Ranking Indirect Connections in Literature-Based Discovery: The Role of Medical Subject Headings (MeSH)

*Don R. Swanson, Division of the Humanities
The University of Chicago
1010 E. 59th Street, Chicago, IL 60637
Phone: 773 702 8267 Fax: 773 834 5867
email:d-swanson@uchicago.edu

Neil R. Smalheiser, Department of Psychiatry, MC912
University of Illinois at Chicago
1601 W. Taylor St., Chicago, IL 60612
Phone: 312 413 4581 Fax: 312 413 4569

Vetle I. Torvik, Department of Psychiatry, MC912
University of Illinois at Chicago
1601 W. Taylor St., Chicago, IL 60612
Phone: 312 355 1707 Fax: 312 413 4569

* Corresponding author.

ABSTRACT

Arrowsmith, a computer-assisted process for literature-based discovery, takes as input two disjoint sets of records (A,C) from the Medline database. It produces a list of title words and phrases, B, that are common to A and C, and displays the title context in which each B-term occurs within A and within C. Subject experts then can try to find A-B and B-C title-pairs that together may suggest novel and plausible indirect A-C relationships (via B-terms) that are of particular interest in the absence of any known direct A-C relationship.

The list of B-terms typically is so large that it is difficult to find the relatively few that contribute to scientifically interesting connections. The purpose of the present paper is to propose and test several techniques for improving the quality of the B-list. These techniques exploit the Medical Subject Headings that are assigned to each input record. A MesH-based concept of literature cohesiveness is defined and plays a key role. The proposed techniques are tested on an earlier published example of indirect connections between migraine and magnesium deficiency. The tests demonstrate how the earlier results can be replicated with a more efficient and more systematic computer-aided process.

1. Introduction

1.1 Direct vs indirect connections between two literatures

To find scientific literature on the relationship between any two searchable terms -- such as a chemical substance (A) and a disease (C) -- one could search Medline for all records on A, all records on C, then form the intersection of set A with set C. This straightforward procedure is called here a search for "direct" A-C connections. (A or C can refer either to a search term or to the set of records created by the search.)

If there are no articles within the AC intersection, an intriguing question arises as to whether

there might exist implicit or indirect connections between A and C, based, for example, on an intermediate literature (B) for which both an AB relationship and a BC relationship have already been separately reported but perhaps not considered together.

It is usually pointless or at least inefficient to pursue indirect connections before conducting a broadly-based literature search for direct connections and analyzing the relevant findings. Any direct connections that exist may eliminate or greatly change the problem and significance of indirect connections.

1.2 Prior and related work on literature-based discovery (LBD).

The problem of finding one or more B that meet the condition stated above has been addressed in a series of papers (Swanson 1986,1987,1988,1989a,1989b,1990a,1991,1993; Swanson & Smalheiser 1997); software designed to aid the discovery process has been developed and made available on a web-based system called Arrowsmith at two sites (http://kiwi.uchicago.edu, http://arrowsmith.psych.uic.edu) (Smalheiser & Swanson 1998b; Swanson & Smalheiser 1996, 1997, 1999). Other researchers, employing statistical methods for the most part, have replicated, evaluated, and/or extended some of the implicit term-associations in this work using a variety of techniques, including lexical statistics (Lindsay & Gordon 1999), latent semantic indexing (Gordon & Dumais 1998), association rule mining (Hristovski et al. 2001), co-word clustering (Stegmann & Grohmann 2003), shared relationship analysis (Wren et al. 2004; Wren 2004), and semantic filtering using the UMLS (Srinivasan 2004, Weeber et al. 2001, 2003).

The possibility of starting with a "problem literature", C (e.g. a disease), then proceeding to an unknown A rather than a given A (Swanson 1991, 1993; Swanson & Smalheiser 1997, 1999) has also been investigated by most of the researchers cited above. Weeber (2001) called this an "open" discovery process, or "hypothesis generation", in contrast to specifying both A and C in a "closed" process. The present paper continues our major focus on the closed process, the improvement of which will have important implications for our open process.

1.3 Implicit connections: An overview of the goal and the problem

The goal of the work reported here is to evaluate and improve Arrowsmith, a computer-assisted process for literature-based discovery in biomedicine. The computer is used to search for, organize, and display information for users who then look for implicit connections that may suggest novel plausible scientific hypotheses.

The "ABC model" described in section 1.1 is implemented by finding all key B-terms (words and phrases) in titles that are common to two disjoint sets of articles, A and C, and then displaying each B-term in the context of its use within A-titles and within C-titles. The user can then assess whether the titles displayed suggest a possibly interesting A-C connection that is worth pursuing further.

If the two input sets each consist of thousands of records, the resulting number of key B-terms they have in common may be so great as to impede a careful search for the few that are novel and of scientific interest. We address this problem on two fronts -- first, by trying to improve the search strategies used in creating files A,C, and, second, by filtering and organizing the B-list itself. Medical Subject Headings (MeSH) play a key role on both fronts.

Any B-term that is judged by the user to be of scientific interest because of its relationship to the A and C literatures is called a "target". It is the target terms that potentially may lead to literature-based discovery. Typically, only a few targets are found among hundreds of B-terms examined.

We tested our proposed solutions to the above problem on a previously analyzed literature-pair (migraine and magnesium) for which the target terms are known (Swanson 1988).

2. Two interpretations of the ABC model.

The focus of interest as discussed in section 1 was on finding an implicit relationship between A and C, but the mechanistic process described lends itself to other objectives as well. For

example, the B-terms themselves could become the focus by virtue of having been investigated in the context of A and separately in the context of C. We reported one such study of viruses as potential biological weapons. Our aim was to find all viruses (B) for which both virulence (A) and stability (C) had been investigated separately, these being properties of particular significance for weaponization and hence for biological defense (Swanson, Smalheiser, & Bookstein 2001; Smalheiser 2001; Kostoff 2001). The nature of the A-C connection per se was not problematic. It follows that there are at least two different ways of using B-terms: i) - to discover novel A-C links, and ii) - to discover B-terms that have a novel conjunction of properties (A and C).

3. Stoplists

Stoplists -- lists of words to be excluded because they are predictably of no interest -- are often used in searching text. Most stoplists consist of up to a few hundred function words that are not subject-oriented. A much longer stoplist (9500 words) has been found valuable in creating title-based B-lists. That stoplist was compiled by selecting words from a composite, frequency-ranked, B-list automatically created by the Arrowsmith system based on a few hundred searches over a multi-year period. Such selection depended on subjective judgments that cannot easily be replicated and the rationale for which is difficult to supply. Evaluation of past searches that influenced the stoplist may not reliably predict the outcome of future searches. Accordingly, for purposes of evaluating specific searches reported in this paper, title-word B-lists are filtered using a short stoplist of only 365 words compiled independently of this project and available at: http://arrowsmith.psych.uic.edu/arrowsmith_uic/data/stopwords_pubmed http://kiwi.uchicago.edu/stopwords_pubmed.

Medical Subject Headings that are used to index Medline records are also filtered using a MeSH-stoplist of 4900 terms that we compiled as part of this project. The criteria for inclusion are for the most part well-defined and replicable, but the need to apply human judgment to some

individual terms has not been eliminated. It is often easier to make lists of stop-terms than it is to find a good rule for doing so. MeSH terms within the following top-level or second-level MeSH categories form the main 4000-term core of the stoplist.

E Analytical, Diagnostic & Therapeutic Techniques and Equipment
F4 Behavioral Disciplines and Activities
G1 Biological Sciences (All G1 terms are names of disciplines.)
G2 Health Occupations
I1 Social Sciences
I2 Education
K Humanities
L Information Science
M Persons (categories of persons -- Age groups, etc.)
N Health Care
Z Geographic Locations

The top or broadest levels just below the single-letter designations of all categories were also added to the stoplist because they were considered too broad to be useful (A1-A16, B1-B7, C1-C23, D1-D27, F4, G3-G12, H, I3 J1-J2). Moreover, the size of the literatures and the numbers of subcategories they subsume provide further criteria. Terms with over 100,000 postings and/or more than 14 subcategories were added to the MeSH stoplist.

A disadvantage of a very long stoplist is that, if it does cause useful terms to be omitted from the B-list, the user is unaware of the loss. We continue to study the consequences of different types and lengths of stoplists.

Any stoplist, once compiled, becomes part of the machine procedure; it is always to some extent fallible, but open to inspection, criticism, and improvement.

4. Target terms from a completed study with known outcome

In 1988, an analysis of the magnesium and migraine literatures led to identifying eleven complementary arguments that connect magnesium deficiency with migraine (Table 1) (Swanson 1988).

TABLE 1 HERE

Each of the statements in Table 1 is taken from or supported by a number of published articles. Altogether 65 migraine articles and 63 magnesium articles were analyzed, after extensive searching of the literature. No single article contains both an "a" statement and a "b" statement, as marked in Table 1, nor do any of the "a" articles cite any of the "b" articles, or vice versa (Swanson 1988).

Prior to 1988, only a few articles had been published concerning a direct magnesium-migraine (A-C) relationship. Subsequently, numerous laboratory and clinical investigations provided supporting evidence for this literature-based hypothesis. Citations to this evidence have been provided elsewhere (Smalheiser & Swanson 1994; Swanson 1993).

Arrowsmith implements the ABC model outlined in Section 1.3. Taking, as input, the results of two Medline searches -- a pre-1988 File A (on magnesium) and File C (on migraine), -- a B-list and its associated titles were produced as output. The search strategy to create file A was based on the occurrence of "magnesium" in both the title and subject heading fields, and similarly for the occurrence of "migraine" in File C. No attempt is made to re-analyze the literature in the light of possibly new valid connections that might be found. Our purpose instead is to replicate the results as reported by Swanson (1988), but now using a more systematic computer-aided process.

Scientific arguments in general cannot be extracted automatically from titles, abstracts, or the full text of articles, but titles often can serve as pointers or clues that guide the viewer to

arguments presented in the text. For that reason Arrowsmith provides a link from each B-term to the A and C titles from which it was extracted, and so helps the user assess whether it might qualify as a target. Examples of target terms and the titles in which they occur have been given elsewhere (Swanson, 1990,1991; Swanson & Smalheiser, 1997).

The eleven argument-pairs connecting magnesium deficiency with migraine can be associated with 43 title-words or phrases selected from a much longer B-list. These 43 terms are "targets" as defined in Section 1.3. Although some are synonyms or variants of others, in general two variants of the same term lead to different title contexts in which they occur, and hence are not redundant. True redundancies, in which similar B-term candidates lead to identical title contexts, are removed.

Table 2 lists the 43 target terms, coded 1 through 11 to reflect the 11 argument pairs, or indirect connections, to which they correspond. (There is one occurrence of code "0", which denotes a direct connection).

TABLE 2 HERE

4.1 Defining the quality of a B-list: Precision and Recall.

The usefulness of any one B-term depends ultimately on the contents of the articles within which that term co-occurs with A and with C. Interpreting that context and its usefulness in suggesting new relationships requires, in general, expert knowledge and human judgment. For any particular example, such as the magnesium-migraine case in which knowledge and judgment have already been applied in order to identify target terms, the effectiveness of the B-list can be defined and measured. We define the following variables:

T = total number of known target terms

(T=43 for the magnesium-migraine case) (Table 2).

S = total number of argument-pairs, or indirect connections
(S=11 for the magnesium-migraine case) (Table 1).
targB = number of target terms on a particular B-list.
argB = number of arguments in Table 1 supported by targB
nB = size of B-list (number of terms on a particular B-list)
P = targB/nB = Precision, or target density.
RT = targB/T = "Recall-T" (term recall)
RS = argB/S = "Recall-S" (connection or argument recall)
RR = RT x RS (combining the two measures of recall)
Ri = recall for argument-pair-i (i = 1,2,3 --- S.)
Rav = {SUM Ri}/S i=1,2,3 ---- S.

In the magnesium-migraine case, we essentially want to know how many out of the 11 argument-pairs were found (R11), what proportion of targets for each one were found (Ri), and what portion of all targets were found (R43). The total information about recall is contained in the eleven values, Ri. How we sum it up and plot it is inevitably somewhat arbitrary. Rav provides one approach to integrating connection-recall, R11, with overall term-recall R43. RR provides a second approach and can be thought of as if it were the product of two probabilities -- the probability of retrieving the 11 connections and that of retrieving the 43 terms.

Other measures can be invented, but none can capture the complexity of the real situation, in which the number and quality of the corresponding articles are crucial factors. The main purpose of any measure is to determine its sensitivity to variation of different system parameters such as weights and rankings of B-terms. We assume that the variables named offer a reasonably good numeric surrogate for B-list quality, and we apply them retrospectively to the magnesium-migraine case. Our purpose next is to gain insight into how to influence precision and recall by assigning weights and rankings to B-terms.

5. Title B-list Ranking using MeSH terms

The problem addressed here is how to identify, automatically, subsets of B-terms that are likely to have higher target density, and hence can be given a higher rank, than other subsets.

Each title B-word or phrase corresponds directly to a small set of, typically, one to a dozen or so records from the A-file and from the C-file. The MeSH terms in these records provide context that might make it easier for the viewer to assess whether the titles are suggestive of an A-C relationship. For example, if an AB title is about magnesium and ischemia, and a corresponding BC title is on ischemia and migraine, then the possibility of a magnesium-migraine connection via the B-term "ischemia" is likely to be greater if the two uses of "ischemia" could be placed in the same context (e.g. both cerebrovascular) rather than in different contexts (such as AB-cardiovascular and BC-cerebrovascular). The corresponding MeSH terms, if displayed to the searcher, might help resolve this point.

For any given title-based B-term, it is plausible that the greater the density of MeSH terms that the corresponding AB and BC records have in common, the more likely these records are to be meaningfully related and to provide a context that might be helpful. We define a subject-heading weight (sh-wt) as follows:

For any given title B-term:

{AB} = subset of records in A containing that title-term.

{BC} = subset of records in C containing that title-term.

nAB = number of records in {AB}

nBC = number of records in {BC}

ncom = the number of unique subject headings that {AB} and {BC} have in common.

The subject-heading weight for a given title B-term is: sh-wt=100*ncom/(nAB*nBC).

This expression represents the density of ncom among all possible pairs of titles displayed (AB with BC), hence the multiplicative denominator AB*BC. Pairs are the most cogent units to

count because the purpose of the display is to facilitate the recognition of potential complementary relationships between A and C titles.

This weight is used to rank title B-terms, placing the higher weights at the top. Note that we use MeSH terms to rank title-derived B-terms. Other researchers who have focused on MeSH terms have used them more directly as a B-list per se rather than as auxiliary to title terms (Hristovski 2001,2003; Srinivasan 2004).

As a test of how well the weighting scheme and stoplists work for a case with known targets, Recall-RR and B-list length as a function of sh-wt cutoff were determined for the magnesium-migraine example under consideration.

A search for migraine as a title-word and as a subject heading (forming an intersection) and a similar search for magnesium as a title-word and a subject heading were used to create files A,C. The resulting B-list consisted of 848 terms with weight 0 and 562 terms with weight >0. The 43 target terms consisted of 7 with weight 0 and 36 with weight >0. The target term distribution between the two weight groups is significantly better than random.

Table 3, and the corresponding solid line curve of Figure 1, show the effect of the B-term weight cutoff on Recall-RR and on the B-list length. Precision and Recall are expressed as percentages.

TABLE 3 HERE

FIGURE 1 HERE

Notable in Table 3, as one progresses from bottom to top, is the two-thirds reduction in the length of the B-list (from 1410 to 466) as a result of keeping only sh-wt>=1. RR is reduced only to 81% and Rav to 88%, so both remain high.

6. Influence of "deficiency" on magnesium-migraine B-list;

implications for subheadings.

Because subheadings qualify and provide context for main subject headings, they are important tools of search strategy development. There are 76 subheadings [MeSH 2004]: each can be attached (by indexers) to MeSH main headings in certain categories. Searchers can similarly attach subheadings to main headings that are specified in a search.

The word "deficiency" is unusual in that it is used either as an ordinary subheading (df), or as part of certain specific main headings. (df) can be attached to any substance, except for certain recognized deficiency states and diseases, for which a main heading that includes the word "deficiency" is used instead -- e.g. "magnesium deficiency" (rather than magnesium/df).

Table 4 shows the results of two Arrowsmith searches, one using "magnesium deficiency" as the A-literature, and the other using "magnesium". Although the first has lower recall, it has disproportionately higher precision. A chi-squared test (using a probabilistic model of distributing target terms randomly between the two B-lists) shows that the improvement in precision is significant: $chi2 = 4.40$ $p<.05$. Thus, qualifying magnesium with deficiency has a significant effect in shortening the B-list. (To avoid what may be an unacceptable loss of target terms, it is preferable to use such a result as a means of ranking rather than filtering.)

TABLE 4 HERE

These results for "deficiency" encourage further investigation of subheadings, and in particular of the potential for identifying other subheadings or qualifiers that play a dominant role, as does "deficiency" in this particular example.

Unlike the weighting scheme described in Section 5, wherein the B-list was organized or filtered after being formed, the B-list in this case is influenced by the search strategy that creates files A and C, and so is more explicitly controlled by the user of Arrowsmith.

7. The concept of literature cohesiveness.

Defining "literature cohesiveness", loosely, as "relatedness of articles to one another", one would expect randomly selected articles to have low cohesiveness, and articles that are all on the same specific subject to have relatively high cohesiveness.

We investigated the effect of subject-relatedness on the length of the B-list. We found, somewhat surprisingly, that several pairs of unrelated disjoint literatures (formed, for example, by searches using subject-neutral words) led to B-lists that were about three times as long as the magnesium-migraine B-list mentioned in Section 4, even though the sizes of the non-subject literature pairs were matched to the sizes of the migraine and magnesium literatures. Moreover the magnesium-migraine B-list contained many terms that correspond to biologically meaningful connections, even though the complete B-list was only about 1/3 as long as that for subject-neutral sets of articles. Why unrelated pairs of literatures tend to have much longer B-lists than a pair of literatures having an extensive pattern of useful connections calls for further explanation.

Searching for a specific subject or topic results in a literature that tends to have a limited or constrained lexicon that can be characterized as a sublanguage, a topic that has been extensively investigated by Zelig Harris, Naomi Sager and other linguists (Harris & Mattick, 1988; Kittredge & Lehrberger, 1982; Grishman & Kittredge, 1986; Sager, 1975). A randomly selected set of articles of the same size, on the other hand, tends to cover a far broader range of topics. Rather than characterizing a random literature as not being subject-oriented, it is more accurate to say that it covers a plethora of subjects so great as to defeat any attempt to characterize its subject matter as a whole. It has therefore a more extensive lexicon. The same would be true of searches based on subject-neutral words.

Two different subject-based literatures in general will differ in their respective sublanguages, and so would be expected to have a much lower overlap of lexicons -- hence a shorter B-list -- than would two randomly selected literatures. Nonetheless, articles based on terms that two subject-literatures have in common (i.e., their B-terms) would tend to be limited and focal in

subject matter, and so would be likely to show a more consistent and meaningful pattern of linkages than in either the random or subject-neutral case.

8. Defining a MeSH-based measure of literature cohesiveness

Successful literature-based discovery depends more on finding a consistent pattern of complementary connections within a literature than on isolated instances of connections, and so requires relatively cohesive literatures at the outset. A literature on magnesium, for example, should not include articles that mention magnesium only incidentally, or only as an inactive component of a compound. To include only articles that have been assigned "magnesium" as a medical subject heading makes use of an indexer's judgment that the topic is covered substantively rather than incidentally. For this reason we base the definition of a cohesiveness measure on MeSH terms.

The limited lexicon in a specific biomedical literature, L, selected by a Medline search will be described in terms of the record-frequency distribution of the MeSH terms, ranked from most frequent to least frequent.

For purposes of computing a cohesiveness measure, the MeSH stoplist is first applied, then all subheadings are stripped off and the resulting number of main headings are counted. (That is, if the same main heading, MHx, occurs four times with four different subheadings in the same record, that record contributes a count of 4 to the total for MHx.)

Define u = total number of unique MeSH terms in L; let k=int(1.7ln(u)+0.5), where "int" (with +0.5) rounds off k to the nearest integer. The multiplier 1.7 is somewhat arbitrary and subject to adjustment; it is used throughout the calculations presented here. (1.7 and the logarithm choice were selected to minimize the variability of coh at low values of size of the record set -- to be discussed in connection with Fig. 3). In effect, k identifies the top most frequent dozen or so MeSH terms in L. Defining "top" as the sum of all frequencies in the range 2 to k+1, and "rem" as the total of all remaining frequencies, then the cohesiveness of L is:

coh=top/(top+rem), a measure of the prominence of the high-frequency peak in the non-stoplist MeSH distribution for L.

The highest frequency term itself is excluded from "top" because it is usually a Medline search term, and so is applied to every record in the resulting set. At issue here is whether the search term is accompanied by a relatively small number (k) of other terms, called the "core", that are heavily used, and so can become the basis for identifying a limited lexicon.

A plausible rationale for the choice of core size is to argue that a typical individual Medline record is assigned anywhere from a few up to 20 or so MeSH terms, and that we can think of the individual article as a model for cohesiveness, or at least an exemplar that anchors our intuitive notion of what cohesiveness ought to mean. Obviously a set of 1000 articles cannot attain the cohesiveness of a single article, but it is reasonable to define the cohesiveness of a set of many articles in terms of the dozen or so most frequently occurring non-stoplist MeSH terms.

We can better appreciate how coh is defined by looking at few histograms of MeSH frequency distributions (Figures 2a,2b,2c).

INSERT FIGURES 2a, 2b, and 2c HERE

Especially worth noting is, first, that the non-subject (essentially random) literature (bottom line in Fig 2a) has no well-defined peak at all, and second that the peak of the distribution for each of the three literatures shown (hypertension, migraine, and magnesium) is confined to nearly the same range of top frequencies in each case -- namely the first 10-20 or so most frequent terms -- normalized to the total number of occurrences.

The purpose of Figures 2a,b,c is to provide a visual image, using a few specific examples, that serve to help the viewer understand the concept of cohesiveness as defined in this paper. One should notice that even though the peaks involve only a small proportion of the total number of terms used to index the literature, they account for a substantial portion of the Medline records. A Medline search on the union of the top 10 terms in the Migraine peak, for example, yields

about 40% of the Migraine records; thus it is plausible that an appreciable sublanguage effect could occur.

8.1 Dependence of coh on size of literature.

We next examine the variance of coh that can arise for different sizes of the same literature. It makes sense to think that two literatures different in size could have more or less the same degree of cohesiveness. Given the definition of coh presented above, its possible dependence on literature-size can be tested. A convenient way to vary size while holding subject matter constant is to take random subsets of each literature generated. The error bars in Figures 3a,3b, show the standard deviation for each measured point where a measurement consists of the mean value for 8 random samples taken from the entire distribution. (This was done only for the smallest literatures, wherein variances are appreciable.) At a single point of size (500 records), the variance of 50 random samples was also measured -- and found to be almost the same as for 8 random samples.

These sets were formed from Medline searches covering the period 1966-1987, corresponding to the time frame of the original study (Swanson, 1988). The magnesium set was formed from a Medline search using magnesium as both a title word and as a subject heading; similarly for migraine. Coh is shown also for two sets of records (10k,40k) using "hypertension" as a title word, selected (consecutively) from a larger set of 75000 records searched on the truncated title-word "hypertens" for the period 1966 to 2004.

One of the literatures generated was itself a pseudorandom or non-subject set of 23,665 records ("nonsubj24k") formed from Medline searches based on a number of title-word searches that appeared to be subject-neutral (standard, balance, including, many, combinations, consideration, cluster, comparing, condition, applied, aproaches, background, bearing, better, evaluating), also called "non-subject".

Results for non-subject, hypertension, magnesium, and migraine are shown in Figures 3a and

3b. The results are notable in three respects:

INSERT FIGURES 3a, 3b HERE

First, variability is quite low -- for most purposes negligible for literatures of 1000 or more records -- but possibly appreciable for small literatures (fewer than 300 articles).

Second, coh is essentially independent of size, at least for literatures of more than 500 articles.

Third, each of the curves for hypertension, magnesium, and migraine are separated by more than 10 standard deviations from the non-subject curve, and the separation remains clear-cut for all literature sizes shown, notwithstanding higher fluctuations for fewer than 300 or so articles.

Nonetheless, we must ask whether the "subject-oriented" literature as a whole, taking into account inter-subject variability, is indisputably distinct from the non-subject literature. The standard deviations of coh shown later answer this question affirmatively, and permit us to draw some useful generalizations about literature cohesiveness.

The average coh for the particular four subject-based literatures given as examples was about 5 times greater than that for the non-subject literature. Because coh applies to any single literature (unlike the B-list, which is defined only in terms of pairs of literatures), it can be applied to files A and C in advance of creating a B-list to assess the prospects that the list will be useful, with greater values of coh being preferable. The possible use of coh in ranking B-lists will be examined in Section 10.

9. Low coh values of literatures formed by searching abstracts

It is plausible that cohesive literatures can be selected by searching titles and/or MeSH terms, but it is not clear that the same is true of searching just abstracts. Even subject-based words or phrases, such as those representing diseases or substances, are often mentioned only incidentally within an abstract, and so are not indexed. It seems unlikely therefore that such occurrences

would contribute to cohesiveness.

Table 5 lists 16 subject terms randomly selected from the 43 target terms in order to show whether coh based on abstract searching is significantly different from coh based on title searching. (Initially the random process selected 18 terms. Several sets of synonymous terms were then combined when conducting the title and abstract searching; the term "magnesium" was added).

TABLE 5,6 HERE

Table 6 shows the search strategy used to compare Medline title searching with abstract searching for the time period 1966-1987. An ovid search of the text field, designated by ".tw." (Set 8), searches both titles and abstracts. By excluding all records found in the title search (Set 1), the result consists only of records in which the search term was found only in the abstract. The cohesiveness of Set 8 compared to that of Set 1 most clearly reveals the coh difference between title-searching and abstract-searching, with title-searching shown to be significantly more cohesive. Set 10 is formed by excluding set 2, the MeSH terms used in conjunction with titles to define the most specific search. It is inevitable that set 10 would therefore have lower cohesiveness because cohesiveness is defined by the most frequent subject headings. However the highest-frequency term (presumed to be a search term) is automatically excluded in any event. The data show that there is a strong additional effect.

Cohesiveness was computed for the literature corresponding to each term for all three downloaded sets (plus set 1, separately downloaded) shown above, using the definition in section 8. The mean values and standard deviations are shown in the last two lines of Table 6.

Three significance tests for the following pairs of sets were calculated by testing the null hypothesis that the two means in a given pair are equal -- i.e. come from the same population.

Test 1: set 1 v set 8: z=3.1

Test 2: set 1 v set 10: z=5.4

Test 3: set 8 v set 10: z=2.2

The null hypothesis is rejected for all three tests.

(z-critical value for $p<.05 = 1.96$). (Repeating with a t-test (t-crit=2.10) led to the same results.). Sign tests also show strong patterns of significance.

We conclude that title-searching leads to significantly more cohesive literatures than does the searching of abstracts alone (i.e. with title results excluded).

These results suggest that, in general, searches for files (A,C) that are to be used as input to develop a B-list should be controlled with respect to field, and based primarily on title words or phrases and/or on MeSH terms.

This conclusion does not imply that the information content of abstracts in general is inferior to that of titles or MeSH terms, for it singles out a limited class of both abstracts and term-occurrences. A recent text mining study led to an estimate that far more information could be retrieved by searching Medline abstracts (on Raynaud's Phenomenon) than could be retrieved by searching titles or MeSH terms (Kostoff et al 2004). Our results do not conflict with Kostoff's because the two studies were conducted under quite different conditions and with different goals. The Arrowsmith goal in revealing implicit connections is related more to precision than to recall, for example. More work is necessary, and underway, to determine what role abstracts should play in the Arrowsmith system.

10. Cohesiveness as a basis for B-list weighting.

Any B-term can be used to form a literature by conducting a Medline title-word search for that term. The coh value of the literature can then be computed and interpreted as being associated with the B-term that was used to create the literature. We have already seen that coh values are higher for subject-oriented literatures than for non-subject literatures, and it follows that coh merits consideration as a means of ranking B-terms.

We report here the coh values for all 43 target terms from the migraine-magnesium

Arrowsmith search, and coh values for two random samples from the remaining (non-target) terms, the first sample consisting only of B-terms with sh-wt=0 and the second with sh-wt>0. It is plausible that the two weights, sh-wt and coh, both measure aspects of cohesiveness, and so are probably correlated; one would expect therefore that coh for sh-wt=0 will be smaller than that for sh-wt>0. (In both cases, the 365-word stoplist was used as a filter for the B-list.) Sample sizes, mean values, and standard deviations are given in Table 7.

TABLE 7 HERE

The null hypothesis that the two mean values of coh (first two lines in Table 7) are equal (i.e. came from the same population) was tested and rejected at the level of $p<<.05$ ($z=11$). zcrit=1.96. The other two pairs of lines (first and third; second and third) were similarly tested and the null hypothesis again rejected in both cases ($z=5.7$; $z=5.6$; $p<.05$).

All three sets of results are thus significantly different from each other and fulfill our expectation that coh for targets should be greater than for non-target terms, and that, among the latter, coh should be smaller if sh-wt=0. These results encourage further testing of coh and sh-wt as a basis for ranking B-terms.

Table 8 shows B-list size as a function of coh cutoff, and so further illustrates the possible use of coh as a means of ranking the B-list. The B-list lengths shown are estimates based on scaling up non-target terms from the sample sizes of 59 and 100 to the populations from which they were selected (873 and 519, resp.), with the number of target terms then added.

TABLE 8 HERE

It is also possible to combine the two types of cutoff criteria. Table 9 shows the result of first applying the sh-wt cutoff at 1 (that is, retain only sh-wt>1 for B-terms), and then removing additional B-terms using the coh cutoff at various values.

TABLE 9 HERE

Figure 1 shows a plot of Recall as a function of B-list size, one curve for various values of the sh-wt cutoffs (Table 3) and a second curve for various values of the coh cutoff (Table 8). A third curve shows the combination of sh-wt >=1 with various coh cutoffs (Table 9). Ideally one would like short B-lists and high recall, so, for any two non-intersecting operating curves, the one that is above and to the left is better.

We can interpret Figure 1 as showing diminishing returns somewhere around the point (x,y) with x=400 B-terms and y=RR%=80%. Above and to the right of this point it is well worth using a weight cutoff to bring an initial length of 1400 terms down to 400 or so, for the price paid in lost recall is relatively small. However, below and to the left of that (x,y) point any further length reduction results in a sharp decline in recall along a curve that is not much better than random deletion of B-list length. (Random deletion from any point (x,y) is represented by a straight line from the origin (0,0) to (x,y)).

Because only the magnesium-migraine case was tested, the results should be interpreted as suggestive rather than definitive.

11. Summary of Results and Implications

The Medline search strategy for creating Arrowsmith's input Files A and C is the first step toward creating a high-coh target-rich B-list that can stimulate literature-based discovery. This first step is under control of the user. Two guidelines to search-strategy development may be helpful:

1. In creating files A and C, search fields should be limited to titles and to Medical Subject Headings (MeSH), and so aim for high precision rather than high recall -- (an important exception being the direct search of the A AND C intersection). Our results imply that the set of all records in which search terms appear only in the abstract has lower cohesiveness, defined by

coh, than records found by title and MeSH term searching.

2. Appropriate subheadings (or other qualifiers) attached to MeSH terms may be of value. In the magnesium-migraine example, we found that the B-list from A-C input that included the A-qualifier "deficiency" led to significantly higher precision than a second B-list from A-C input without the qualifier.

We have defined and developed a concept of literature cohesiveness, and have proposed and tested a MeSH-based measure of it, called coh. We found that coh was much higher for subject-specific literatures than for literatures with no, or very diffuse, subject orientation. Moreover, literature pairs with low coh led to much longer B-lists. It is reasonable therefore to test input files A and C for cohesiveness and consider changing to a more specific search strategy if coh falls below about 0.15.

We further investigated two methods of automatically filtering or ranking the B-list after it has been formed.

1. Assign a rank-weight (sh-wt) to each B-term according to the number of subject headings in common between the AB and BC records for the given B-term, divided by the product of the total number of AB and BC records for the given B-term.

2. Assign a rank-weight to each B-term using coh.

The results for both ranking methods (1. sh-wt; 2. coh) as shown in the upper portion of Figure 1 are remarkable in the very large reduction in the length of the B-list that can be attained with relatively small loss of Recall. By either measure, the first one or two steps of the cutoff are very rewarding, but beyond that point are of questionable value. Thus the curve of Figure 1 displays an "elbow" forming a more or less optimal region of operation. These results were based on a small title-word stoplist of 365 terms; a longer stoplist will lead to shorter B-lists.

Although these techniques are tested using the single example of the magnesium-migaine study, it is plausible that they can profitably be applied to other cases in which novel implicit connections between a substance and a disease are sought.

12. Significance of the study

The significance of this study for literature-based discovery lies in exploiting a pre-existing body of expert human knowledge embedded in the design of Medical Subject Headings and subheadings and their application to the Medline database. The exploitation takes place at two levels within Arrowsmith -- preparation of the input files, (A,C) and in ranking or filtering the output (B).

Although only a closed process was analyzed here, the techniques of B-list ranking are important for reducing the search space of a quasi-open process (Swanson 1993; Swanson & Smalheiser 1999).

13. Acknowledgment

This work has been conducted under an agreement between The University of Chicago and the University of Illinois - Chicago and has been supported through grant 1 R01 LM07292-01, Arrowsmith data mining techniques in neuro-informatics, co-sponsored by NLM and NIMH, 6/15/01- 5/31/06; PI: Neil R. Smalheiser, M.D., Ph.D

We are grateful to Abe Bookstein for helpful discussions.

References

Gordon, MD and Dumais, S. (1998). Using latent semantic indexing for literature-based discovery. Journal of the American Society for Information Science 49(8), 674-685.

Hristkovsky, D., Stare, J.,Peterlin, B., & Dzeroski, S. (2001). Supporting Discovery in Medicine by Association Rule Mining in Medline and UMLS. Medinfo 10(Pt 2), 1344-1348.

Hristovsky, D., Peterlin, B., Mitchell, J.A., & Humphrey, S.M. (2003) Improving literature based discovery support by genetic knowledge integration. Studies in Health Technology & Informatics 95, 68-73.

Kostoff, R.N. (2001) Predicting biowarfare agents takes on priority. The Scientist 15(23), 6.

Kostoff, R.N., Block J.A., Stump, J.A. & Pfeil, K.M. (2004). Information content in Medline record fields. International Journal of Medical Informatics 73(6), 515-527.

Lindsay RK, Gordon MD. (1999). Literature-based discovery by lexical statistics. Journal of the American Society for Information Science 50(7), 574-587.

Narayanasamy V., Mukhopadhyay S., Palakal, M. & Potter, D.A. Transminer: Mining transitive associations among biological objects from text. Journal of Biomedical Science 11(6), 864-873.

Smalheiser, N.R. (2001). Predicting emerging technologies with the aid of text-based data mining: the micro approach. Technovation 21, 689-693.

Smalheiser, NR & Swanson, DR. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. Neuroscience Research Communications, 15(1), 1-9.

Smalheiser, NR & Swanson, DR. (1996a). Indomethacin and Alzheimer's Disease. Neurology 46(2), 583.

Smalheiser, NR & Swanson, DR. (1996b). Linking estrogen to Alzheimer's Disease. An informatics approach. Neurology, 47(3), 809-810.

Smalheiser, NR & Swanson, DR. (1998a). Calcium-independent phospholipase A2 and schizophrenia. Archives of General Psychiatry 55(8), 752-753.

Smalheiser, NR & Swanson, DR. (1998b). Using Arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57, 149-153.

Srinivasan, P. (2004). Text mining: Generating hypotheses from Medline. Journal of the American Society for Information Science and Technology 55(5), 396-413.

Stegmann, J. & Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. Scientometrics 56(1), 111-135.

Swanson, DR. (1986). Fish-oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 30(1), 7-18.

Swanson, DR. (1988). Migraine and magnesium: eleven neglected connections. Perspectives in Biology and Medicine, 31(4), 526-557.

Swanson, DR.(1989). A second example of mutually isolated medical literatures related by implicit unnoticed connections. Journal of the American Society for Information Science 40(6), 432-435.

Swanson, DR.(1990). Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association 78(1), 29-37.

Swanson, DR.(1991). Complementary structures in disjoint science literatures. In A. Bookstein, Y. Chiaramella, G. Salton, & VV Raghavan, eds. SIGIR'91, NY:ACM Press. 280-289.

Swanson, DR. (1993). Intervening in the life cycles of scientific knowledge,

Library Trends 41(4), 606-631.

Swanson, DR. & Smalheiser, NR. (1996). Undiscovered public knowledge: a ten-year update. In: Simoudis, E., Han, J., & Fayyad, U. KDD-96 Proceedings. Second International Conference on Knowledge Discovery & Data Mining. Menlo Park: AAAI Press.

Swanson, DR. & Smalheiser, NR. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence 91(2), 183-203.

Swanson, DR. & Smalheiser, NR. (1999). Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. Library Trends 48(1), 48-59.

Swanson, DR, Smalheiser NR, & Bookstein A. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. Journal of the American Society for Information Science and Technology 52(10), 797-812. August 2001

Weeber, M. (2001). Literature-based Discovery in Biomedicine. Dissertation. Groningen: Rijksuniversiteit

Weeber, M., Vos, R., Klein, H., & de Jong-van den Berg, L.T.W. (2001). Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish-oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology 52(7), 548-557.

Weeber, M., Vos, R., Klein, H., de Jong-Van Den Berg, L.T.W., Aronson A.R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. Journal of the American Medical Informatics Association. 10(3):252-259.

Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., & Garner, H.R.

(2004). Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics 20(3), 389-398.

Wren, J.D. (2004) Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics 5(1), 145.

Table 1. Eleven indirect arguments suggesting that a magnesium deficiency may be implicated in migraine.

a) Statements supported by migraine literature	b) Statements supported by magnesium literature
1a Stress and Type A behavior are associated with migraine.	1b Stress and Type A behavior lead to body loss of magnesium.
2a Excessive vascular tone and reactivity can aggravate or predispose to migraine.	2b Magnesium can reduce vascular tone and reactivity.
3a Calcium channel blockers have been used to prevent migraine.	3b Magnesium is a natural calcium channel blocker.
4a Spreading cortical depression may be implicated in the early phase of a migraine attack.	4b High levels of magnesium in the extracellular cerebral fluid can inhibit spreading cortical depression in animals.
5a There is evidence for a connection between epilepsy and migraine.	5b Magnesium deficiency can increase susceptibility to epilepsy in animals.
6a Migraine patients have abnormally high platelet aggregability.	6b Magnesium can inhibit platelet aggregation.
7a Platelets from migraine patients are abnormally sensitive to serotonin release.	7b Magnesium deficits can lead to high levels of serotonin release.
8a Substance P may be a cause of head pain in migraine.	8b Magnesium deficits can increase Substance P activity.
9a Low levels of prostacyclin or high prostaglandin e1 release can aggravate vasoactivity and Substance P activity in migraine.	9b Magnesium deficits can lead to low levels of prostacyclin release.
10a Migraine may involve sterile inflammation of cerebral blood vessels.	10b Magnesium has anti-inflammatory properties.
11a Cerebral hypoxia may play a key role in migraine.	11b Magnesium can protect against brain damage from hypoxia.

Table 2. Target terms associated with statements 1-11 of Table 1.

0 antimigraine
1 stress
2 arterial spasm
2 cerebral vasospasm
2 coronary spasm
2 paroxysmal
2 reactivity
2 spasm
2 vasospasm
3 calcium antagonist
3 calcium blocker
3 calcium channel
3 calcium channel blocker
3 diltiazem
3 nifedipine
3 verapamil
4 cortical spreading depression
4 spreading
4 spreading depression
5 anticonvulsant
5 convulsion
5 convulsive
5 epilepsy
5 epileptic
5 epileptiform
5 seizure
6 anti aggregation
6 platelet aggregation
6 platelet function
7 5 ht
7 5 hydroxytryptamine
7 5 hydroxytryptamine receptor
7 brain serotonin
7 serotonin
8 substance p
9 prostacyclin
9 prostaglandin
9 prostaglandin e1
9 prostaglandin synthesis
10 anti inflammatory
10 indomethacin
10 inflammatory
11 hypoxia

Table 3. Cumulative Recall and B-list length as a function of weight cutoff (sh-wt) for magnesium-migraine (T=43; S=11) (stoplists applied)

sh-wt>=	#targets	#B-terms	R43(%)	Precis(%)	R11(%)	RR(%)	Rav(%)
100	9	59	21	15	64	13	24
50	13	114	30	11	64	19	32
15	23	204	53	11	91	48	49
10	25	260	58	10	91	53	62
5	29	329	67	9	91	61	70
1	35	466	81	8	100	81	88
0	43	1410	100	3	100	100	100

Table 4. Recall and Precision for A=magnesium deficiency v A=magnesium

A	C	sh-wt	#targets	#Blist	Precis%	R43%
mag-defic	migraine	>0	19	84	23	44
magnesium	migraine	>0	35	279	13	81

Table 5. coh values for literatures formed by searching abstracts only (bab or sab), compared to searching titles (ti) for a random subset of target terms. bab=abstracts excluding title sets. sab=bab excluding MeSH search-term sets.

Search Terms	ti	bab	sab
1. (calcium AND (block$ OR antag$))	.37	.19	.13
2. diltiazem	.39	.33	.26
3. nifedipine	.35	.30	.23
4. verapamil	.27	.27	.22
5. antiaggegation	.31	.23	.22
6. platelet aggregation	.31	.27	.14
7. prostacyclin	.32	.29	.26
8. prostaglandin e1	.25	.22	.20
9. brain serotonin	.28	.26	.16
10. 5-hydroytryptamine receptors	.34	.19	.23
11. spreading AND depression	.26	.26	.22
12. convulsion$	.22	.14	.11
13. spasm OR spasms	.39	.17	.13
14. inflammatory	.17	.07	.07
15. migraine OR antimigraine	.21	.21	.11
16. magnesium	.26	.16	.09
MEAN	.293	.222	.174
SD	.064	.064	.060

Table 6. Medline search strategy for Table 5.
(Ovid notation -- x = search term)

set# search statement	Notes:
1 x.ti. + download	title search
2 exp x	exploded MeSH search
3 1 and 2	intersection of sets
4 limit 3 to yr=1966-1987	date restriction
5 keep (download)	download for title search
6 x.tw.	text-word search (ti + ab)
7 limit 6 to yr=1966-1987	date restriction
8 7 not 1	omit set 1, leaving abstr bab
9 keep (download)	download for abstract search
10 8 not 2	omit set 2, leaving abstr sab
11 keep (download)	download

Table 7. coh values for migraine-magnesium B-list
using 365-title-word stoplist and 4900-term MeSH stoplist.

	coh: mean	std dev
All 43 target terms:	0.29	0.08
59-nontarget sample; sh-wt=0	0.11	0.07
100-nontarget sample; sh-wt>0	0.19	0.11

Table 8. Cumulative Recall and B-list size as a function of coh cutoff in a ranked B-list for magnesium-migraine (T=43; S=11) (stoplists applied)

Bterm-coh>=	#targets	#B-terms	R43(%)	R11(%)	Precis(%)	RR(%)
0.300	18	118	42	45	15	19
0.260	24	208	56	64	12	36
0.230	28	283	65	73	10	48
0.200	33	349	77	91	9	70
0.160	39	518	91	100	8	91
0.000	40	1392	93	100	3	93

Note: recall for coh is based on 43 terms, but 3 were omitted because of having too few records to give reliable results. Thus, maximum possible R43(%) and RR(%) = 93%.

Table 9. Combining coh-cutoff with sh-wt >=1
for magnesium-migraine case (stoplists applied)

coh cutoff>=	#targets	#B-terms	R43(%)	R11(%)	Precis(%)	RR(%)
0.260	21	146	49	64	14	31
0.230	24	190	56	73	13	41
0.200	29	242	67	91	12	61
0.190	31	270	72	100	11	72
0.000	35	466	81	100	8	81

See Note, Table 8

Figure 1

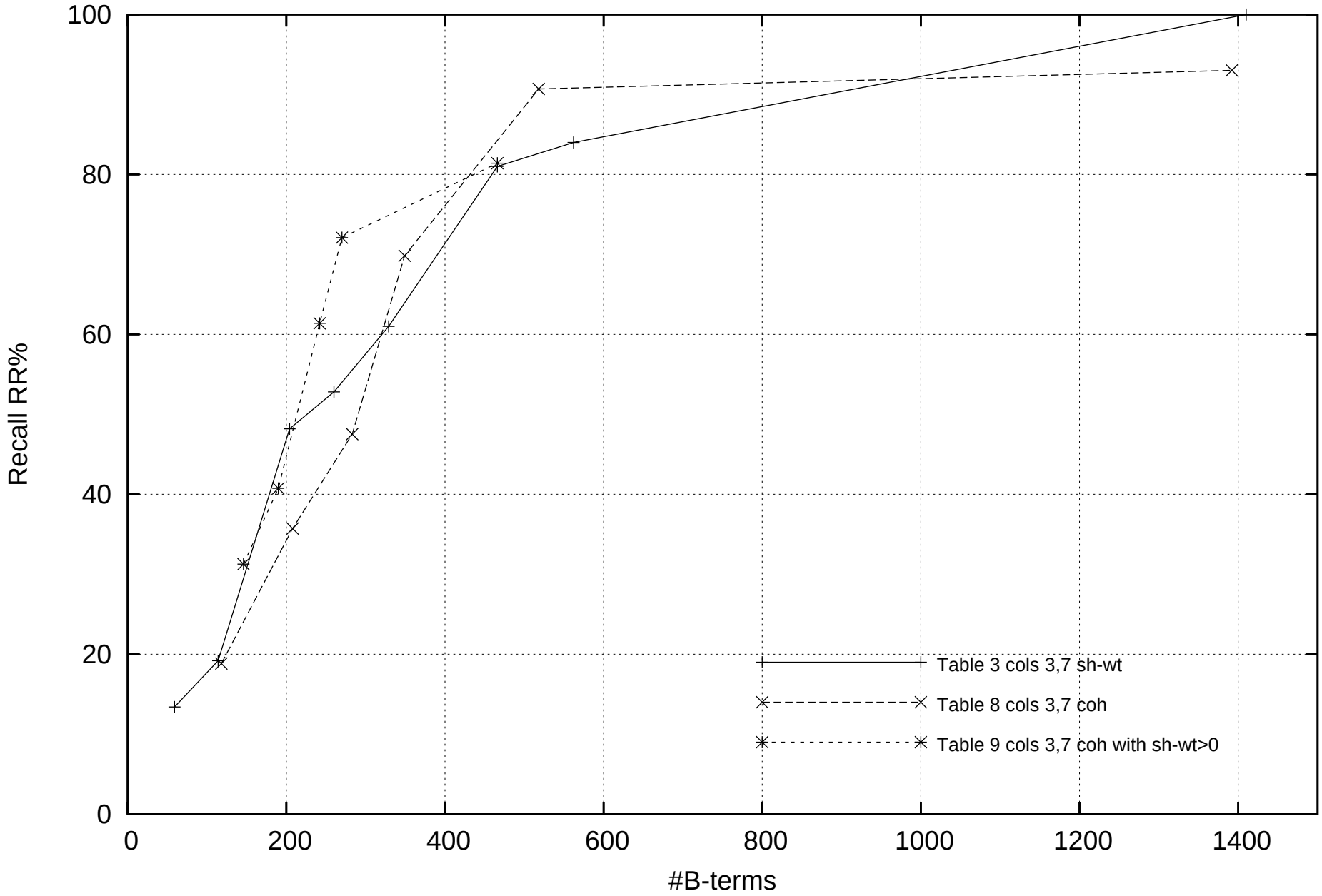

100
80
60
40
20
0
Recall RR%
0
200
400
600
800
1000
1200
1400
#B-terms
Table 3 cols 3,7 sh-wt
Table 8 cols 3,7 coh
Table 9 cols 3,7 coh with sh-wt>0

Figure 2a: MeSH term frequency distributions

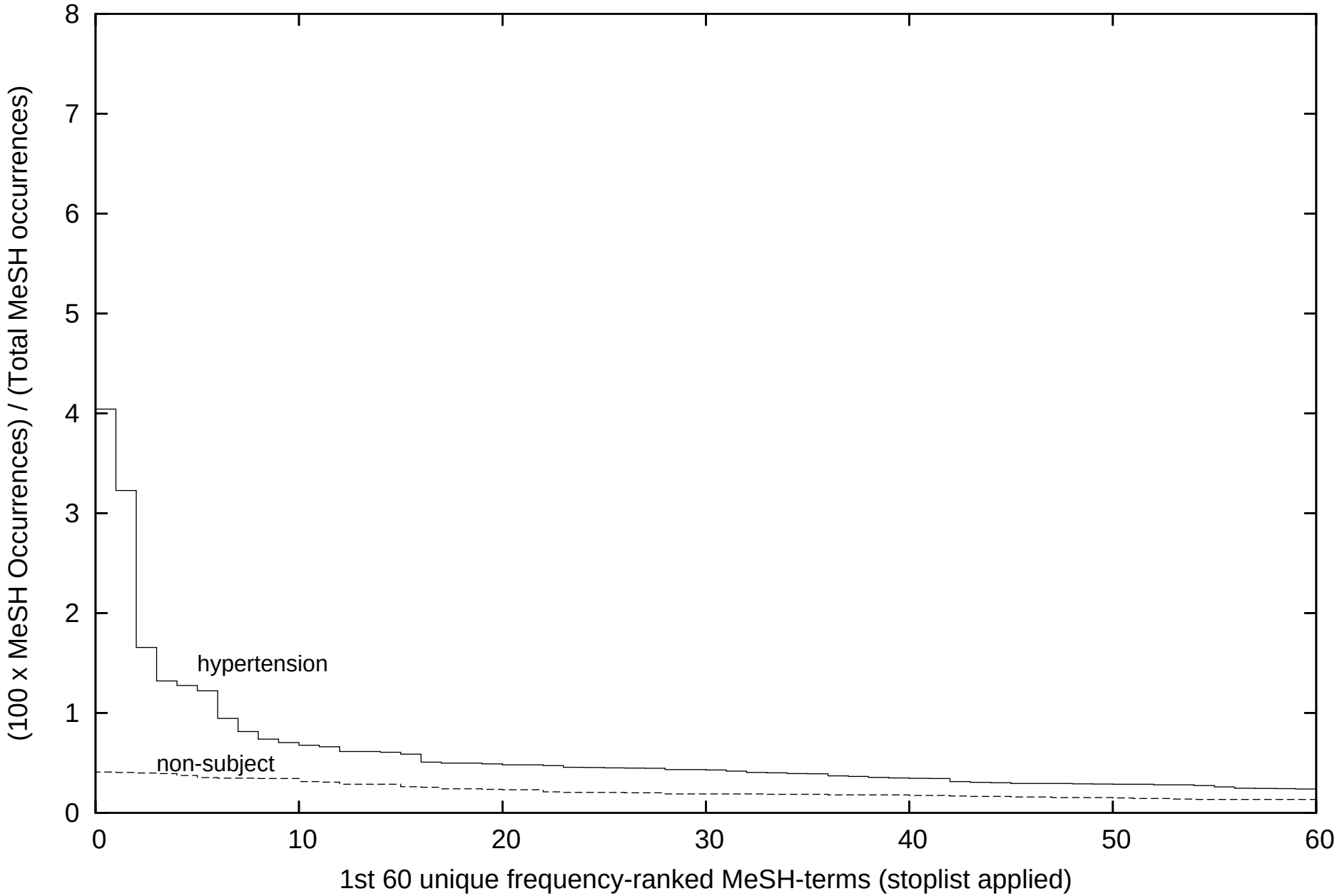
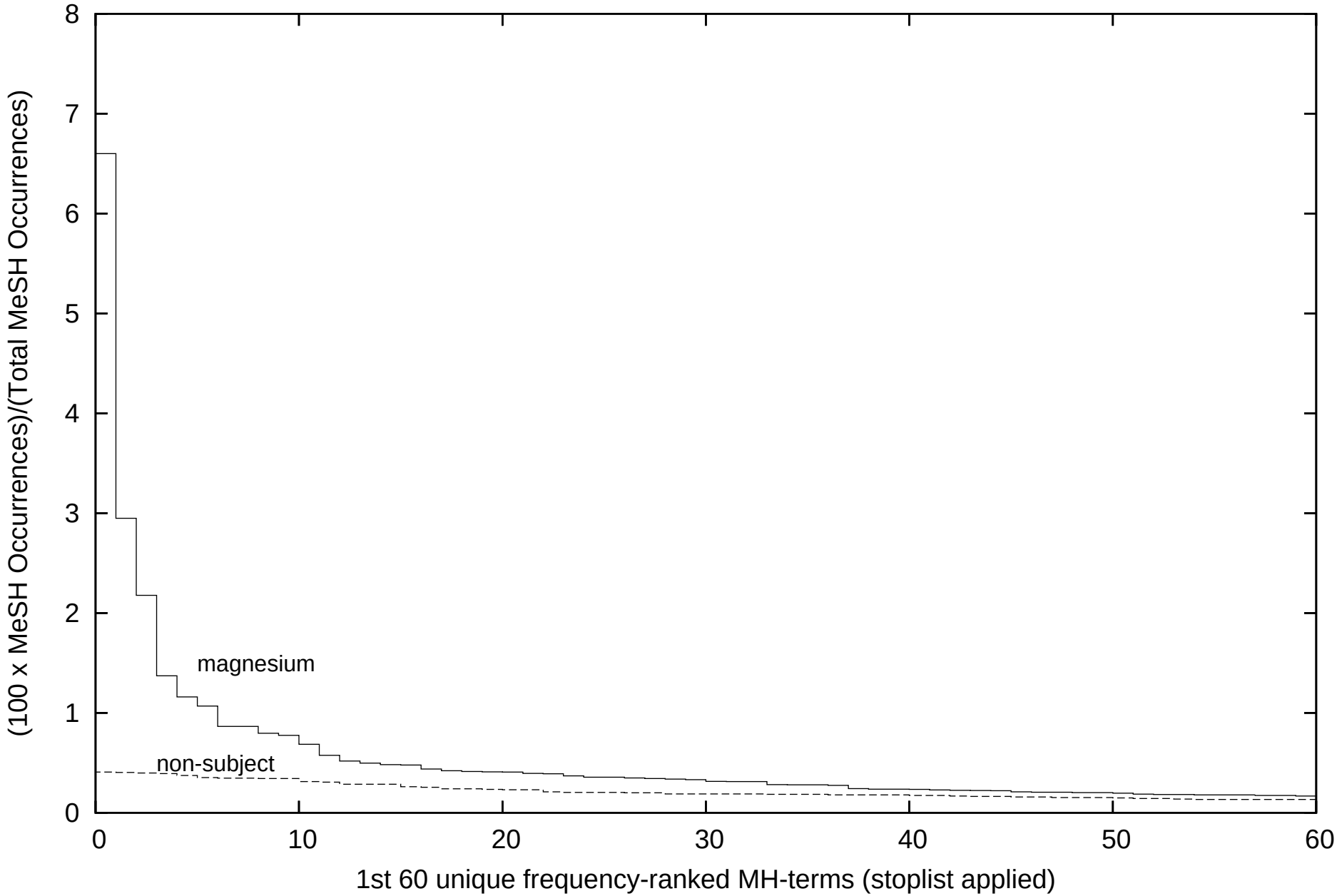


Fig 2b: MeSH term frequency distributions



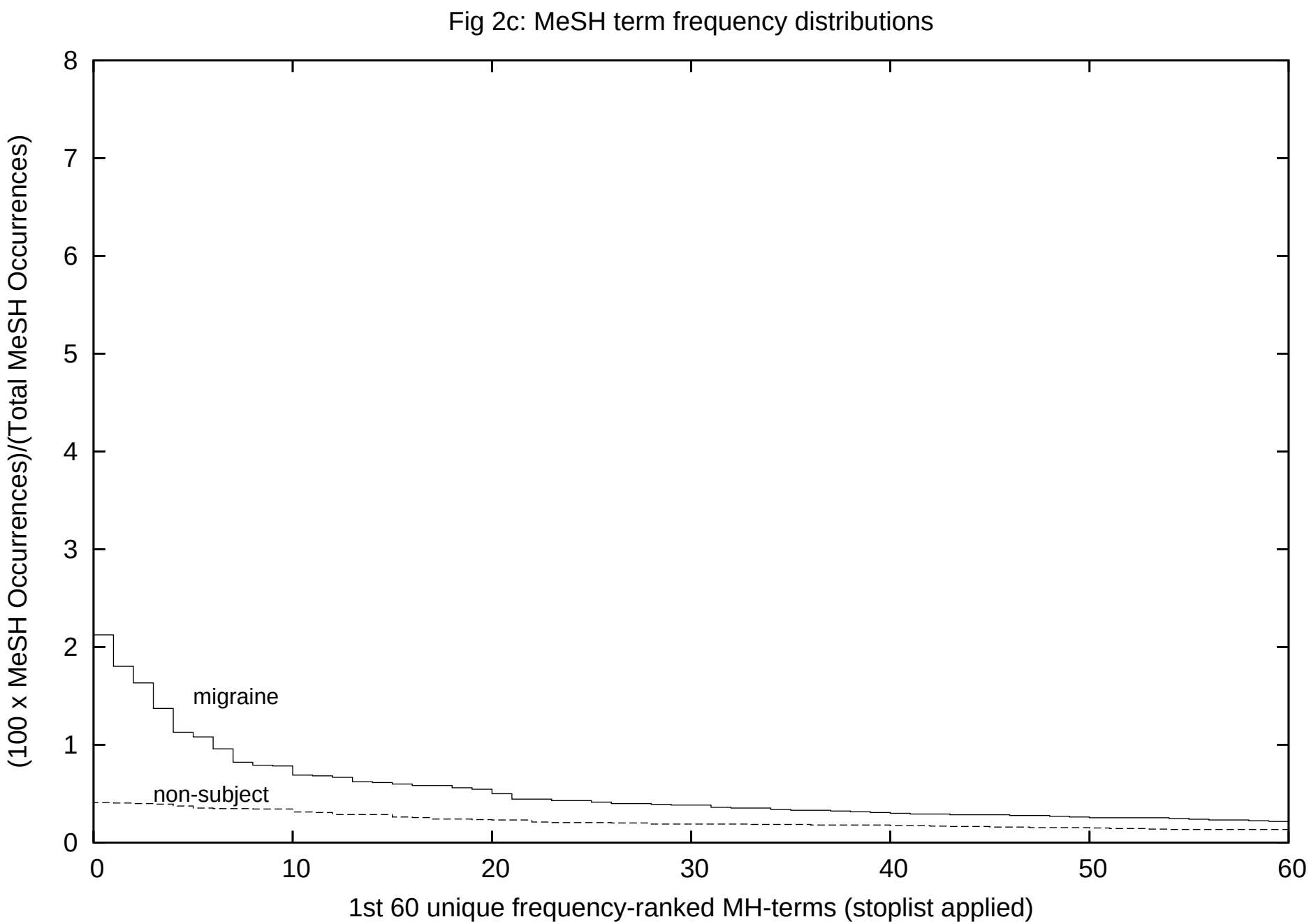


Fig 2c: MeSH term frequency distributions
(100 x MeSH Occurrences)/(Total MeSH Occurrences)
1st 60 unique frequency-ranked MH-terms (stoplist applied)
migraine
non-subject
0
1
2
3
4
5
6
7
8
0
10
20
30
40
50
60

Figure 3a: Cohesiveness (coh) as function of literature size.

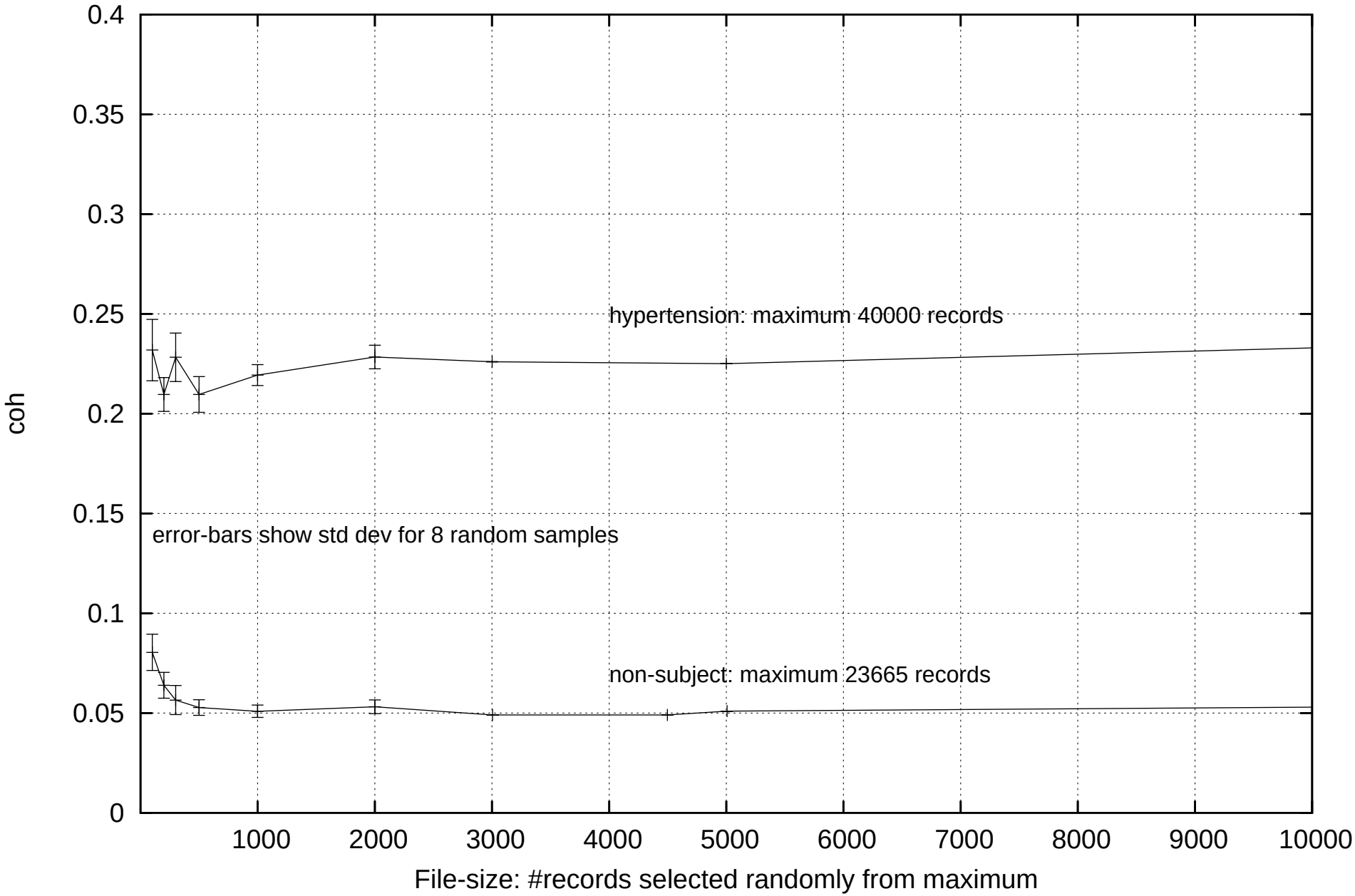


Figure 3b: Cohesiveness (coh) as function of literature size