Literature-Based Discovery by Lexical Statistics

Robert K. Lindsay
Mental Health Research Institute, University of Michigan Medical School, 205 Zina Pitcher Place, Ann Arbor, MI 48109. E-mail: lindsay@umich.edu

Michael D. Gordon
University of Michigan Business School, 701 Tappan Street, Ann Arbor, MI 48109. E-mail: mdgordon@umich.edu

We report experiments that use lexical statistics, such as word frequency counts, to discover hidden connections in the medical literature. Hidden connections are those that are unlikely to be found by examination of bibliographic citations or the use of standard indexing methods and yet establish a relationship between topics that might profitably be explored by scientific research. Our experiments were conducted with the MEDLINE medical literature database and follow and extend the work of Swanson.

Introduction

An enormous and rapidly growing collection of information is available in print and electronic forms. The number of possible connections among elements of this collection is far greater than the number of documents itself. Some organization of this information is provided by indices, summaries, hyperlinks, structured databases, and many other methods of annotation. However, the very magnitude of this information store assures that it must contain ideas that are related in ways that have not been explicitly noted and thus remain difficult to find. Some of these relationships are undoubtedly extremely important.

This situation suggests that there may be ways to use modern computational techniques to aid the discovery of some important implicit connections. Information that is stored in structured databases offers a variety of mechanisms for discovering implicit connections. This has been the impetus for research and practical efforts in database mining in which work has been done to make better use of information contained within and spread among huge, isolated databases (Piatetsky-Shapiro & Frawley, 1991; Fayyad & Uthurusamy, 1995). But, with very few exceptions, this work has been restricted to examining either numeric data or structured (nominal) data such as those managed by traditional database management systems, not text.

However, a far greater amount of information is contained in written reports, e-mail messages, journal articles and other textual forms than in structured databases. To date, efforts to develop techniques to comb through and make discoveries within such textual stores have been very few. However, the gains to be made from literature-based discovery can be enormous, for instance if it leads to a breakthrough in the treatment for a dreaded disease like AIDS or cancer. Certainly, less dramatic discoveries are extremely valuable too, such as a discovery that might lead a firm to make use of engineering research already performed when developing a seemingly unrelated project.

Text, of course, is highly structured by the syntax and semantics of natural language. However, this structure has defied explication to date. There is a large research effort in the field of artificial intelligence called natural language processing (Allen, 1987). Some progress has been made on the very difficult problem of developing computational methods of understanding ordinary language, but this work has not progressed anywhere near the point where scientific articles can be routinely processed and "understood" in a human sense (Lehnert & Sundheim, 1991; Lehnert, Cardie, Fisher, McCarthy, Riloff, & Soderland, 1994).

It is still an open and heatedly argued question whether a computational system will ever be able to fully master the use and comprehension of a natural language. Most would at least agree that any such system will be quite unlike what is available today because far richer forms of experience are necessary to grasp the structure and nuance of natural language. Whatever the merit of this belief, it is clear that procedures for understanding natural language in a manner even roughly approximating that of a human will not be

Received February 16, 1998; revised September 21, 1998; accepted October 13, 1998

available soon. The computational burden of full-text syntactic analysis, the difficulty of creating a grammar that covers a stylistically broad literature such as scientific articles, and the lack of established means for representing complex medical and scientific concepts make this approach impractical at the moment. Consequently, we have avoided any attempt to perform a full syntactic and semantic analysis of text in our work.

By ignoring the complex syntax and conceptual structure of language, our literature-based discovery work must basically take place at the level of words and short phrases, perhaps grouped into categories by stemming rules that combine variants such as singular and plural forms. Such a conception of language need not ignore all syntactic and semantic structure. For example, words can be indexed by part-of-speech information and semantic category information. It is a syntactic fact that *erythrocyte* is a noun, and a semantic fact that it is synonymous with the phrase *red blood cell*, and both of these facts could be available in a lexical database lacking knowledge of a full grammar or semantic model of English.

More tractable still are methods that are based solely on lexical statistics, that is, methods that treat a particular string of characters solely as an instance of a word or phrase, without reference to its deeper linguistic significance. In this article we continue to explore this limited but computationally manageable analysis to see whether it alone can provide leverage for the discovery of implicit connections in literature. The methods we offer appear to have some value by themselves, and may be extendible to deeper word and phrase analyses in the future.

The focus of this work is the scientific literature. Scientific literature is broken up into numerous specializations and, within each, scientists write principally for other specialists. Accordingly, the resulting published literature may contain important information about relations among different specialties that are unnoticed by any of the contributors, each of whom has seen only part of the picture. Insofar as these important relationships are not indexed or cross-cited, they may remain undiscovered if literature searches are conducted solely by the customary methods of keyword and citation searching.

It is important to remember that literature-based discovery cannot replace traditional empirical scientific research or even literature search, but rather supports them by providing the scientist with a means to organize more easily a potentially overwhelming amount of information.

Background

The idea that the medical literature contains unnoticed but discoverable connections has been investigated by Don Swanson of the University of Chicago. In a series of papers (see references) Swanson has described previously unnoticed connections that he discovered using the MEDLINE medical literature database. One of these is that fish oil leads to certain changes in blood viscosity and red blood cell rigidity that may help alleviate the disease known as Raynaud's syndrome (Swanson, 1986). Another is that magnesium may have effects in alleviating migraine headache by, among other things, reducing vascular reactivity, inhibiting cortical depression, and suppressing Substance P (Swanson, 1988, 1989). In a third example, various implicit connections were uncovered relating arginine and blood levels of somatomedin C, which itself is related to growth, nutrition, protein synthesis, and thymic function (Swanson, 1990). Other examples include linkages between magnesium and neurologic health (Smalheiser & Swanson, 1994), indomethacin and Alzheimer's disease (Smalheiser & Swanson, 1996b), estrogen and Alzheimer's (Smalheiser & Swanson, 1996a) and phospholipases and schizophrenia (Smalheiser & Swanson, in press). These examples illustrate different kinds of connections implicit in the literature, including clinical suggestions of potential therapies and basic physiological linkages. In each example, the literatures of the topic pairs were essentially unrelated bibliographically.

Some textual information stores can be conceived as a graph whose nodes are individual documents and whose links represent explicit connections among documents, say by citations. There are a variety of ways to impose additional structure on this graph. For example we can consider, as a unit, all documents that contain a particular word or phrase (e.g., *migraine*). Having defined a set of such (probably overlapping) literatures, one may then define a new set of links among them. For example, one such link might be defined between two literatures that share a common word or phrase (or words and phrases within a common semantic category) at a frequency above a specified threshold. Alternatively, or additionally, one could define a link between two literatures if they both cite each other with sufficient frequency. Many other possible relations could be conceived.

The notion underlying Swanson's work is that one can establish connections between a literature of interest, *A*, and other literatures, *Bi*. This could be done by human effort, including reading extensively, or by using existing indexing schemes, either explicit links, cross-citations, or implicit connections of the sort just mentioned, some of which can be revealed by keyword searches. Then one may use similar techniques to find connections between each *Bi* and other literatures *Cij*. This procedure could perhaps be continued until an interesting target *T* is discovered. At that point an important connection may have been discovered that merits further understanding and investigation if the final target literature, *T*, is bibliographically unrelated to the original literature *A*. That is, though each discovery step may be well-studied and well-described, the source and target literatures may be completely bibliographically isolated from each other: No author writes articles describing both *A* and *T*; there are no articles that an *A* article and a *T* article both cite; nor are there any articles that cite both an *A* and a *T* article. In Swanson's discoveries, the elemental chain of reasoning leading to the previously unnoticed connections

had been only two-steps — *A* to *B*; and *B* to *C*— and yet this yielded the interesting connections cited above. Clearly, the branching factor at each stage will cause the search to be far more difficult as each step is added unless pruning heuristics are sufficient to eliminate most barren paths while retaining the fertile ones.

Swanson's work has demonstrated that diligent exploration of the scientific literature can lead to the discovery of unnoticed connections with the basic two-step sequence. The problem is to find an intermediate literature, *B,* (such as *blood viscosity* or *red cell rigidity*) that can link an *A* (*Raynaud's*) to a *C* (*fish oil*). However, the hypothesis-discovery process, inherently one of trial and error, has demanded an extraordinary amount of time to read and study the literature, generate hypotheses about potential *B*s, conduct MEDLINE searches to test them, and iterate this cycle until a plausible link between *A* and *C* has been established. In Swanson's efforts to find a buried Raynaud's connection, many false leads were considered and rejected leading to the (simplified) line of reasoning: Raynaud's, which affects the circulation, might have a blood-related treatment, especially if blood viscosity can be reduced; and fish oil seems to do that, though no Raynaud's researchers have considered it.

Though no automated process can be a full substitute for reading and understanding portions of the medical literature, computational methods that support such an effort make the quest for literature-based discoveries far more efficient—and even more effective if they allow the discoverer to focus more widely or more systematically, or to perform more purely intellectual tasks, confident that the software will appropriately deal with various bookkeeping tasks that may arise in dealing with and analyzing large amounts of text.

In this spirit, Swanson and Smalheiser (1997) have developed an experimental software system called ARROWSMITH to help detect unnoticed or rarely noticed literature-based linkages. ARROWSMITH performs two discovery tasks: First, it generates "potential discovery terms" that might warrant further investigation. Second, it tries to display the possible relationships that link such potential discovery terms to the initial discovery source. Using the linkage between migraine and dietary magnesium as an example (since it is the example that we will use to describe our own efforts), the essential operation of ARROWSMITH in the generation of potential discovery terms is as follows: Terms that co-occur in journal titles with the word migraine are enumerated. "Stop words" and general terms (like *drug*) are discarded, along with any term that seems to have only a random occurrence frequency among migraine article titles. The terms remaining suggest intermediate literatures (or topics) that might form a bridge between migraine and a new discovery. A new list of "potential discovery" terms or topics is then generated, by searching for any MEDLINE titles containing both one of the intermediate terms and any other term that can be considered a dietary, deficiency, or drug factor involved in migraine, as determined by context and MEDLINE subheading codes. Possible factors that are highly ranked by their relative frequency of occurrence within the intermediate literatures' titles are then ordered by their absolute frequency.

TABLE 1. Comparison of the methods of Swanson/Smalheiser and Gordon/Lindsay for literature-based discovery.

	Swanson/Smalheiser	Gordon/Lindsay
MEDLINE text analyzed	Titles, main headings, MeSH subheadings	Complete record
Unit of lexical analysis	Words	Words and phrases
Selection of intermediate concepts	Relative frequency rules out concepts	Token counts, record counts, and tf * idf suggest concepts
Identification of terminals	No. of links to intermediate topics	Relative frequency (also: token counts, record counts, tf * idf)
Topical restrictions	Dietary, dietary deficiency, toxicity, and poison terms	None

Note: This table highlights the main differences between these methods. Both methods are subject to variation; typical uses of the method are identified here.

Previously (Gordon & Lindsay, 1996), we have developed software that successfully replicated Swanson's Raynaud's/fish-oil connection (Swanson, 1986) in a systematic fashion without relying on an understanding of the medical literature beyond our general knowledge of English and technical discourse. Here we report experiments that apply and extend that approach in studying the migraine/magnesium connection reported by Swanson.

First, a source literature (migraine) is identified, and the complete MEDLINE record of any document mentioning the term *migraine* is downloaded. A complete MEDLINE record includes a bibliographic citation, as well as an abstract and various subject descriptors. Next, all one-, two-, and three-word phrases within this corpus are found, excluding those containing any item on a list of approximately 1000 stop words (*the, of, to,* etc.) and general medical words (*medicine, patient,* etc.) that are so general as to be uninformative. These phrases are combined according to stemming rules that equate singular and plural forms. For each of these words and phrases, we then calculate four statistics: (1) its token frequency (tf) within the downloaded corpus; (2) its document frequency (df), i.e., the number of downloaded records that contain the term or phrase; (3) its relative frequency (rf) in the corpus vs. MEDLINE as a whole; and (4) its value for the statistic tf * idf—where the inverse document frequency (idf) is the log of the ratio of the number of records in MEDLINE to the number of records in MEDLINE using that word or phrase. Relative frequencies and tf * idf statistics rely on MEDLINE base rates, which we obtain by use of an automated telecommunications script.

Based on these statistics one or more intermediate literatures is identified. Each is treated similarly to the migraine

literature; i.e., it is downloaded and statistically analyzed in order to identify potential discovery topics.

Table 1 summarizes the differences between Swanson and Smalheiser's approach for partially automating literature-based discovery and our own. First, our approach is based on complete MEDLINE records, theirs is based on MEDLINE titles. Second, we attempt to identify intermediate literatures or topics based on a term having a very high occurrence frequency; their approach rules out terms with low occurrence frequency. Third, we base our analyses on four statistics that form the basis of much work in information retrieval: token counts, document counts, relative frequencies, and the tf * idf statistic. Their approach seeks terms whose df is significantly greater than would be predicted by a Poisson distribution. Fourth, we employ two- and three-word phrases throughout our work, whereas they use only single words for generating potential discovery terms. Fifth, we seek any kind of novel connection, while their approach restricts connections to certain "search targets," such as dietary factors, dietary deficiencies, toxicities, and poisons. In our work we cull, by hand, the lists of discovered terms to eliminate those that are "nonterminal," that is, those that seem not to provide a mechanism of therapy, prophylaxis, or physiological relevance. Such culling is based on our knowledge and judgment rather than automated classification. Swanson and Smalheiser's selection of search targets is performed using MEDLINE's MeSH subheadings, main headings, and title words.

Our previous work (Gordon & Lindsay, 1996) succeeded in replicating Swanson's generation, prior to the ARROWSMITH system, of the hypothesis that fish oil is a potential treatment for Raynaud's disease. Focusing on the same subset of MEDLINE used by Swanson— the literature that was current at the time of his work— we were able to show that, by the use of lexical statistics a medical scientist could make the following connections between literatures describing the various subsets of MEDLINE:

1. Raynaud's—blood;
2. Raynaud's ∩ blood—blood viscosity;
3. Blood viscosity—fish oil.

These connections require some human judgment but are mainly based upon the descriptive statistics described above. For instance, by examining token counts, the research scientist can learn that *blood* is the second most prominent substantive term in the Raynaud's literature (excluding the term *Raynaud's*, itself). Thus, the decision to pursue the link from *Raynaud's* to *blood* is defensible on lexical and statistical grounds. But so would be the decision to pursue the link from *Raynaud's* to *scleroderma*—which also has a very high token count in relationship to *Raynaud's*.

The results of our work in replicating the Raynaud's/fish-oil connection led us to form the hypothesis that: (a) intermediate literatures are best identified by absolute lexical frequencies; and that (b) candidate discoveries are best generated from intermediate literatures by using relative frequencies.

The purpose of the experiments we report on in this article is to test further the applicability of the methods of literature-based discovery support that we have investigated. It is our belief that literature-based discovery holds great promise for many tasks where taking maximum advantage of information is of paramount importance. We also believe that no single method will be successful in all circumstances. We view our method as a complementary approach to that of Swanson and Smalheiser. It operates differently and should perform differently, sometimes better, sometimes worse. Each will often point to different topics possibly worth further exploration by traditional empirical scientific investigation. Ultimately, it would be useful to have a catalog of various literature-based discovery techniques and some guidelines for choosing among them. We view this work as part of that effort.[1]

Experiments

The basis for the work reported here is Swanson's research suggesting that the scientific literature contained a scarcely noticed hypothesis linking migraine headaches and dietary magnesium. This connection has subsequently been confirmed by empirical research. Swanson (1988) specifically identifies 11 intermediate literatures linking the two topics. We considered these 11 topics, as well as vasospasm, which is related to vascular tone and reactivity.

We ran two sets of experiments. The first was aimed at determining if our methods would lead from the literature on migraine to any or all of the 12 intermediates linking migraine and magnesium. The second set of experiments began by analyzing each of these 12 topics to determine what discovery targets they might lead to. This latter step was performed by considering each of the 12 analyses separately and also by pooling their results.

Discovery of Intermediate Literatures

We began by downloading the 1081 MEDLINE records from 1986 to 1988 (inclusive) that contained the word *migraine*. This approximately represents the state of knowledge that was current at the time of Swanson's investigation. (Swanson examined MEDLINE prior to mid-1987 back to 1966 and by reading the articles discovered the 11 connections mentioned earlier. He then selected 65 of these articles and 63 magnesium articles as being particularly important.) Using software, we selected from these records all one-, two-, and three-word phrases, excluding stop or noise words and phrases. Word boundaries were signaled by spaces and other punctuation; in addition to all single words,

[1] In this spirit, we have compared our techniques to the statistical technique known in the IR literature as latent semantic indexing (Gordon & Dumais, 1998).

TABLE 2. Salience of intermediate term in migraine literature.

Intermediate literature	Token analysis Count	Token analysis Rank	Record analysis Count	Record analysis Rank	Rel freq Rank	tf * idf Rank
Calcium channel blocker	95	1	35	1	49	2
Spreading cortical depression	16	9	15	6	25	8
Cortical spreading depression	5	117*	3	97*	28	20
Either variant	21	(7)	15	(5)	(26)	(4)
Vascular resistance	11	56	9	36	152	86
Vascular responses	3	696*	3	520*	104	167
Vascular reactivity	3	696*	2	139	90	163
All variants	17	(31)	11	(22)	(148)	(57)
Prostacyclin = prostaglandin I_2	5	1810*	4	1492*	71	1808*
Prostaglandin(s)	36	28	12	572*	96	27
Ketoprostaglandin	6	1582*	3	1688*	60	108
All variants	47	(22)	19	(23)	(103)	(22)
Inflammation	11	909*	7	900*	99	83
Inflammatory	13	760*	12	586*	111	82
Either variant	24	(57)	19	(23)	–	(50)
Hypoxia	6	1582*	6	1045*	77	122
Vasospasm	36	28	26	17	16	16
Vasospastic	6	1582*	6	1045*	8	85
Either variant	42	(27)	28	(15)	(21)	(14)
Platelet aggregation	47	13	32	8	99	17
Platelet aggregability	5	495*	4	77	31	111
Either variant	52	(10)	32	(7)	(100)	(15)
Substance P	25	23	9	35	67	32
Stress	101	12	46	12	58	10
Epilepsy	143	7	56	8	26	3
Epileptic(s)	40	25	22	360*	37	15
Either variant	183	(7)	61	(7)	(38)	(1)
Serotonin	169	5	53	11	33	1
Serotoninergic	5	1810*	4	1492*	27	114
Serotonergic	17	579*	7	900*	39	37
All variants	191	(3)	54	(10)	(45)	(1)

* Original rank; no culling; other entries: culling of non-relevant intermediate concepts; ranks are computed separately for 1-, 2-, and 3-word phrases. ( ) = hypothetical rank; – indeterminable rank.

software was used to identify all 2- and 3-word adjacency phrases that were not split by periods, parentheses, or other punctuation marks that identified clause boundaries. We then calculated the four lexical statistics for all the selected items, along with their ranks as determined by these statistics. Table 2 presents the results for the four statistics.

Several notes help explain the table. First, one-, two-, and three-word rankings were computed separately. As shown in Table 2, for example, *calcium channel blocker* was the most frequent three-word term in the migraine literature (by both token and record counts). But other one- and two-word phrases (such as *serotonin*) appeared more frequently. The reason for separately considering phrases of different lengths is that, while a longer term (like *calcium channel blocker*) is precise and descriptive, it necessarily occurs less frequently than shorter phrases like *calcium*, *channel*, *blocker*, or *calcium channel*. Thus, one-word phrases were all analyzed in the context of other one-word phrases, and similarly for two- and three-word phrases. Second, the variants that we considered for a term (given in Table 2) were generated by manually inspecting the list of phrases that the computer analysis produced, not systematically or automatically. Third, rankings consider only topics that are conceivably of interest, based on our non-expert judgment. Since the task at hand was to identify useful intermediate concepts, certain items could be safely excluded. For instance, in the two-word term analysis, *cerebral ischemia* (ischemia means shutting off of blood flow) occurred less frequently than *cluster headache*, *double blind*, and *classic migraine*. However, the latter phrases were ruled out of further consideration (and so were not given a better rank than *cerebral ischemia*) on the grounds that they would not be productive to pursue as an intermediate concept. Third, for each literature we analyzed, we

TABLE 3. Top intermediate terms generated by the four statistics.

Allergy or allergies	Adrenergic beta	Adenosine cyclic monophosphate
Amnesia	Angina pectoris	Adrenal cortex hormones
Aneurysm or aneurysms	Basilar artery	Adrenergic alpha receptor
Angina	Beta blockade	Adrenergic beta receptor
Blood	Beta blockers	Alpha receptor agonists
Calcium	Beta receptor	Alpha receptor blockaders
Cerebrovascular	Blood coagulation	Anti-inflammatory agents
Contraction or contractions	Blood flow	Auditory evoked potentials
csd, or cortical spreading depression	Blood platelets	Benign paroxysmal vertigo
Depression or depressions	Calcium antagonists	Beta receptor blockaders
Endorphin or endorphins	Calcium channel	Blood brain barrier
Epilepsy	Cerebral artery or arteries	Blood coagulation factors
Epileptic or epileptics	Cerebral infarction	Calcium channel blockers
Hemorrhage	Cerebral ischemia	Calcium channel blocking
Hypertension	Cerebrovascular circulation	Cerebral arteriovenous malformations
Ischemia	Cerebrovascular disorders	Cerebral blood flow
Lupus	Channel blockers	Chronic paroxysmal hemicrania
Muscle	Chronic paroxysmal	Cortical spreading depression
Neuralgia	Cortical depression	Lupus coagulation inhibitor
Ophthalmoplegia	Ergotamine derivatives	15 hydroxytryptophan
Ophthalmoplegic	Ischemic attacks	Mitral valve prolapse
Pregnancy or pregnancies	Lupus erythematosus	Monoamine oxidase inhibitors
Prostaglandin or prostaglandins	Myocardial infarction	Monocular visual loss
Raynaud's	Paroxysmal hemicrania	Muscle contraction headache
rcbf	Platelet aggregation	Nerve compression syndromes
Serotonin	Raynaud phenomenon	Plasma beta endorphin
Somnambulism	Receptor blockaders	Posterior cerebral artery
Stress	Spreading cortical	Recurrent abdominal pain
Stroke	Spreading depression	Regional blood flow
Tourette or Tourette's	Subarachnoid hemorrhage	Regional cerebral blood
Vasoconstriction	Temporomandibular joint	Retinal vein occlusion
Vasospasm	Tolfenamic acid	Selective beta blockers
Vertigo	Trigeminal neuralgia	Somatosensory evoked potentials
	Vascular resistance	Spreading cortical depression
	Vertebrobasilar insufficiency	Systemic lupus erythematosus
		Temporomandibular Joint Syndrome
		Tolosa hunt syndrome
		Transient global amnesia
		Transient ischemic attacks
		Visual evoked potential

This table is the set of all items that were among the top 35 intermediate terms or phrases that were identified by two or more of the four lexical statistics. Items are shown alphabetically along with other items of the same phrase length.

established a token frequency and document frequency threshold for each phrase length; phrases falling below these thresholds were automatically eliminated. The thresholds were chosen largely on the basis of what size the result would be. To keep each analysis manageable, size was limited to between two and three thousand phrases. Finally, we wished to explore in a preliminary way the effects that considering concepts rather than words and phrases might have on our results.[2] In this regard, we identified "by hand" lexical and slight semantic variants of the items we were focusing on and calculated the relevant statistics for them. For instance, we identified *cortical spreading depression* as a synonym for *spreading cortical depression* and calculated the four lexical statistics for it. In addition, we calculated statistics for the disjunction of these two phrases. In determining ranks, we acted as if a variant (or a disjunction) was being considered instead of the main variant. So, although *serotonin* was the highest ranked single term by the tf * idf statistic, the disjunction *serotonin* or *serotoninergic* or *serotonergic* was also better than all other single terms by this same statistic, and so was the disjunction *epilepsy* or *epileptics*. Thus, each of these items received a rank of 1. Since the detection and grouping together of these variants was not done in an automatic or rigorous way, we deem the ranks for these lexical variants and disjunctions hypothetical ranks, and we denote "hypothetical ranks" with parentheses in Table 2.

The token and record count data in Table 2 suggest that most of the intermediate concepts that Swanson identified as being important bridges from migraine to magnesium are

[2] See the Discussion section of the paper for a brief discussion about the National Library of Medicine's MetaThesaurus for doing such "collapsing" automatically.

TABLE 4. All-statistics prominence analysis: Distribution of prominence in intermediate literatures.

Number of intermediate literatures = *n*	Number of terms prominent in exactly *n* literatures	Number of terms prominent in *n* or more literatures
1	1582	3159
2	507	1577
3	343	1070
4	172	727
5	116	555
6	96	439
7	83	343
8	72	260
9	51	188
10	69	137
11	55	68
12	13	13

3,159 terms were identified according to a lexical statistic analysis of twelve intermediate literatures. Shown here is the distribution of the number of intermediate analyses by which these terms were detected as prominent. For example, *serotonin* was one of 13 terms that was identified by all 12 intermediate literature analyses, and *cyclooxygenase* was one of 260 terms identified by eight or more of the 12 analyses.

readily identified by lexical statistics. Except for *hypoxia* and *inflammation*, each of these is among the top 31 concepts by at least one statistic. *Calcium channel blocker* is the most prominent three-word phrase in the migraine literature analyzed, and *spreading cortical depression, platelet aggregation, epilepsy,* and *serotonin* are all in the top 10.

In summary, 10 of the 12 intermediate concepts linking migraine and magnesium were among the first few dozen items suggested by either token or record count analyses of the migraine literature.

Table 3 lists the union of all phrases that were among the top 35 for two or more of our four statistics, after obvious artifacts were removed by the authors (i.e., without employing any specialized knowledge of medicine or physiology).

The prominence of 10 of the 12 concepts identified by Swanson confirms the usefulness of identifying fertile intermediate terms by means of lexical statistics. Of course, we had the advantage at the outset of aiming for targets Swanson had previously identified, and hitting them was our purpose. In a different sort of experiment, where we were not trying to replicate a previous discovery but, rather, create a new one, we would have no basis for focusing further study on just these 12, but would consider as many of the highest ranked terms as would be manageable. We did not pursue that strategy here, but instead turned to investigate the value of lexical statistical methods in dealing with the second stage of the process.

Discovery of Target Terms

An argument can be made that if some item turns up prominently among the 12 intermediate topics, and if that item is previously unrecognized as having a relationship to migraine, then it deserves further attention by medical researchers. In this spirit we conducted the following analysis. First, for a given intermediate literature, we determined the top 500 items for each of the four lexical statistics we employed. Second, we created the union of these lists (thus creating a list of potential size between 500 and 2000 items). Third, we performed the same two steps (separately) for the other 11 intermediate literatures, yielding 12 discovery candidate lists. Fourth, we formed the union of these 12 different lists. The resulting list contained 3159 terms, each of which occurred on from 1 to 12 of the discovery candidate lists. The distribution of the number of discovery candidate lists containing a given item is shown in Table 4.

We next selected the items that occurred on eight or more of the 12 discovery candidate lists, reasoning that each of these was possibly interesting to explore further since it was lexically prominent according to at least two thirds of our intermediate literature analyses. There turned out to be 260 such terms. Since we were only interested in items that were either rarely or never before used in conjunction with *migraine* (prior to 1989), we next consulted MEDLINE to determine the frequency with which each of these items appeared along with *migraine* (through 1988). However, even *magnesium* had not gone completely unmentioned, but in fact seven documents mentioned both it and *migraine*. Intersection frequencies are shown in Table 5.

The 57 terms that co-occurred with *migraine* seven or fewer times were shown to a medical student who had far

TABLE 5. All-statistics novelty analysis: MEDLINE co-occurrence frequencies with *migraine* of the most prominent 260 terms identified by all-statistics intermediate literature analyses.

Migraine co-occurrence frequency = *f*	Number of terms with migraine co-occurrence frequency *f*	Number of terms with migraine co-occurrence frequency *f* or less
0	12	12
1	5	17
2	8	25
3	11	36
4	6	42
5	6	48
6	5	53
7	4	57
8	4	61
9	6	67
10	5	72
11	6	78
12	7	85
13	2	87
14	6	93
15	2	95
>15	165	260

The 260 terms that were in the top 500 (on at least one of the four statistics) of at least 8 intermediate literature analyses were checked to determine how often they appeared in a MEDLINE record along with the term *migraine* prior to 1989. The 57 terms with *migraine* intersection less than or equal to seven were examined more closely for their discovery potential.

better understanding than we of medical terminology and some knowledge of hypothesized mechanisms for migraine. He was able to combine some of these terms because they were synonyms or subclasses of other terms, and to eliminate others as being completely unrelated to the physiology of migraine. The following are the remaining terms with his brief annotations. Some are related to migraine in a way that suggests increasing their presence may be beneficial, others more likely have a deleterious effect. Either category merits investigation, since identifying a substance that is deleterious could lead to ways to reduce it or block its production physiologically.

a23187 = calcimycin—a calcium ionophore (i.e., it transports calcium ions across membranes) which can stimulate platelet aggregation and secretion of 5HT (see ketanserin), a potentially positive effect.

cyclooxygenase—an enzyme inhibited by migraine medications such as aspirin and nonsteroidal anti-inflammatory drugs such as ibuprofen. Cyclooxygenase leads to the creation of prostaglandin and is a key enzyme in the metabolic pathway from arachadonic acid to prostaglandin and leukotriene. Thus this substance has a potentially deleterious effect.

glutathione—an antioxidant within cells that keeps intracellular enzymes from being oxidized and becoming potentially damaged. Platelet superoxide dismutase (SOD) is an enzyme that destroys oxygen free radicals. In patients with migraine with aura, low concentrations of SOD have been found. This indicates susceptibility to oxidative stress which could possibly be ameliorated with glutathione. Thus this substance has a potentially positive effect.

kcl—can initiate spreading cortical depression, probably by increasing the excitability of neurons. Used as an inducer of spreading depression in experiments. Thus this substance has a potentially deleterious effect, or its prominence may be an artifact.

ketanserin—a 5-hydroxytryptamine (5HT2) receptor antagonist. For example, ergotamine, a well-known migraine prophylactic, can induce venoconstriction in hand veins. This effect is blocked by ketanserin, therefore implicating the 5HT2 receptors in the process. Thus this substance has a potentially deleterious effect or might be indicative of one.

protein kinase C—some effects of migraine therapies have been reported to be linked to protein kinase C, which activates calcium stores within cells. Certain forms of PKC are altered during spreading cortical depression, but its relation to migraine pathophysiology is not known. Thus this substance may have either a positive or a deleterious effect.

leukotriene—these metabolites of arachodonic acid are a target of some migraine therapies. Thus these substances have a potentially deleterious effect.

ltd4—a subclass of leukotrienes. It is a bronchoconstrictor and inflammatory mediator in asthma. Thus these substances have a potentially deleterious effect.

monophosphate—cyclic AMP is a monophosphate taking part in the protein kinase C signal transduction pathway. Suggests a reduced beta-adrenergic receptor sensitivity that could to be an indicator of possible central nervous system reduced sensitivity.

oxidation deficiency (rather than just oxidation)—a percentage of the migraine population has oxidation deficiency problems, but studies so far have not shown any real relation to migraine. A similar percentage of people have been reported to be deficient in cytochrome P450, an oxidative enzyme, and to suffer from migraine. Thus this condition has a potentially deleterious effect.

pge2—a subtype of prostaglandin. It was seen to be the strongest in a series of prostaglandins used to relax constricted arteries. It is supposed to play the largest role in producing pain during the migraine attack; perhaps the extended relaxation and stretching of blood vessels causes pain. Thus this condition has a potentially deleterious effect.

pgf1—6-keto-pgf1-alpha is a stable prostacyclin metabolite and is found to be decreased in migraineurs. Probably indicates a decrease in the amount of prostacyclin during migraine, but most likely only an indicator and not a therapeutic agent. Thus this substance may have a positive effect or merely be indicative of one.

pgf2—pgf2-alpha contracts tissue and reduces intraocular pressure. Thus this substance may have a positive effect or merely be indicative of one.

phenylephrine—postsynaptic alpha-1 agonist. This induces sympathetic responses, but is not a direct therapy. Some studies have shown that there is a hyper-responsiveness to phenylephrine in migraineurs, indicating a chronic deficiency of sympathetic response. Thus this substance may have a positive effect or merely be indicative of one.

prazosin—a dilator of arteriovenous shunts. Arteriovenous shunts are supposed to play a role in the pathophysiology of migraine; prazosin is an antihypertensive agent. A dilatory agent might not be beneficial since most therapies seem to promote venoconstriction. Thus these substances have a potentially deleterious effect.

thrombin—induces platelet aggregation. In all studies, used strictly as an inducer of platelet aggregation to simulate the events of migraine. Thus this substance has a potentially deleterious effect.

Revised Analysis Ignoring Nondiscovered Intermediate Literatures

As seen in Table 2, two of the 12 intermediate literatures identified by Swanson did not yield a rank that could justify the claim that they had been discovered. We asked how our final results would differ if these two, *hypoxia* and *inflammation*, were not included in the preceding analysis. Repeating the above procedure without those data and selecting items that were prominent on at least seven of the 10 intermediate lists yielded a few differences. The 12 literature analysis yielded the following final target terms that were not found by the 10 literature analysis: *Cyclooxygenase, glutathion, ltd4,* and *pgf1*.

In addition, the 10 literature analysis yielded the following terms that the 12 literature analysis did not by promoting terms mentioned in neither the *hypoxia* or *inflammation*

TABLE 6. Salience of magnesium in intermediate literatures.

Intermediate literature		Token analysis Count	Token analysis Rank	Record analysis Count	Record analysis Rank	Rel freq Rank	tf * idf Rank
	Mg **	129		60			
Calcium	Magnesium	86		36			
channel blocker	Union	215	87	80	110	292	87
	Mg **	22		7			
Spreading	Magnesium	19		9			
cortical depression	Union	41	39	12	59	139	40
	Mg **	116		17			
Vascular resistance	Magnesium	106		26			
	Union	222	56	27	188	582	49
	Mg **	17		5			
Prostacyclin	Magnesium	35		10			
	Union	52	162	12	245	415	156
	Mg **	19		9			
Inflammation	Magnesium	22		9			
	Union	41	868	16	961	1422	874
	Mg **	45		16			
Hypoxia	Magnesium	32		11			
	Union	77	229	20	365	468	224
	Mg **	32		5			
Vasospasm	Magnesium	54		9			
	Union	86	75	9	270	538	57
	Mg **	39		23			
Platelet aggregation	Magnesium	105		34			
	Union	144	188	47	260	281	181
	Mg **	16		8			
Substance P	Magnesium	13		6			
	Union	29	539	12	584	647	542
	Mg **	36		12			
Stress	Magnesium	50		14			
	Union	86	214	21	373	495	194
	Mg **	73		24			
Epilepsy	Magnesium	123		34			
	Union	196	71	43	156	316	64
	Mg **	87		39			
Serotonin	Magnesium	90		30			
	Union	177	299	55	446	900	291

Note: The 12 intermediate literatures were analyzed via lexical statistics to determine the prominence of *magnesium* (and *Mg*). Ranks and counts include all items possibly causing or capable of treating any disease. Ranks shown are the lowest possible, in the case where several items have the same statistical value (tied ranks). Ranks may drop markedly when items already linked to Migraine bibliographically are eliminated from consideration. **Mg compounds counted, incl: Mg, $MgCl_2$, MgATPase, Mg_2.

literatures: *npy*, *phenoxybenzamine*, and *yohimbine*. Of these, all had frequencies of intersection with *migraine* of five or less, and thus could be considered potential discoveries by our criteria.

Testing Our Discovery Hypothesis

Our earlier work led us to suggest that relative frequencies are the appropriate lexical statistic to use to analyze intermediate literatures in order to generate possible discovery targets. Our reasoning was that an item that has a high relative frequency within an intermediate literature but not an especially high absolute frequency of occurrence—that is, an unexpectedly high percentage of its occurrences are in the intermediate literature—will be strongly related to the intermediate literature but will not be so commonplace there as to be known by all researchers familiar with the initial seed topic (in this case migraine). In other words, a term uncovered by a relative frequency analysis will introduce an element of novelty, and thus is an item that may be a

concept with an unnoticed connection to the discovery seed.

Table 6 shows the lexical statistics supporting the connection between each of the 12 intermediate literatures and magnesium. In general, the relative frequency statistic was the worst of the four statistics in suggesting magnesium. This is contrary to our hypothesis and is consistent with Swanson's (1991) observation. One explanation for the failure of our hypothesis is that *magnesium* has a much higher frequency in the medical literature as a whole than does the rather obscure term *fish-oil*. This could mask any slight increased relative frequency it had in the intermediate literatures. Alternatively, an article about a comparatively widely studied substance like magnesium may be able safely to omit central characteristics about this substance in a way that articles on less fully investigated topics could not. If correct, these explanations suggest that the relative frequency hypothesis must be modified by considering the magnitude of the base-rate frequencies. Only if these are small will relative frequencies be efficacious.

However, a novel concept may be detected by consistently being somewhat prominent in many different intermediate literatures, rather than by its prominence in a single intermediate literature. In other words, the "clues" a concept leaves in any one literature may be weak, but together these clues may be convincing. We sought to test the usefulness of relatively frequencies considered in this way. Specifically, the process that we outlined above (in Discovery of Target Terms) for determining the top items according to any lexical statistic was modified to produce a list of top terms according to their relative frequencies alone. Specifically, we obtained 12 separate lists—one for each intermediate concept—each containing the top 500 terms according to a relative frequency analysis of that concept. The union of these 12 lists produced 2,956 unique items, identified according to the 12 separate relative frequency analyses, as shown in Table 7. For example, there are 608 terms that occurred prominently in exactly two of the 12 intermediate literatures; the two hits were of course not the same for all 608.

TABLE 7. Relative frequency prominence analysis: Distribution of prominence by relative frequency.

No. of intermediate literatures *n*	No. of terms prominent by relative frequency in exactly *n* literatures	No. of terms prominent in *n* or more literatures
1	1688	2956
2	608	1268
3	289	660
4	141	371
5	95	230
6	54	135
7	36	81
8	27	45
9	8	18
10	7	10
11	3	3
12	0	0

2,956 terms were identified according to a relative frequency analysis of twelve intermediate literatures. Shown here is the distribution of the number of intermediate analyses by which these terms were detected. For example, 230 terms each had a high relative frequency in one or another subset of size five or more of the 12 intermediate literatures.

TABLE 8. Relative frequency novelty analysis: MEDLINE co-occurrence frequencies with *migraine* of the most prominent 230 terms identified by relative frequency analyses.

Co-occurrence frequency with *migraine* = *f*	No. of terms with co-occurrence frequency *f*	Cumulative no. of terms with frequency *f* or less
0	33	33
1	9	42
2	17	59
3	14	73
4	7	80
5	8	88
6	5	93
7	7	100
8	4	104
9	2	106
10	5	111
11	5	116
12	7	123
13	1	124
14	6	130
15	4	134
16 or more	96	230

The 230 terms that were prominent in five or more intermediate literatures by a relative frequency analysis were checked to determine how often they appeared in a MEDLINE record along with the term *migraine* prior to 1989. The 100 terms with migraine intersection less than or equal to seven were examined more closely for their discovery potential.

We selected those terms identified by at least five relative frequency analyses of intermediate terms; there were 230 of these. We then queried MEDLINE to determine how frequently each of these terms occurred within MEDLINE along with the term *migraine* prior to 1989. The distribution of these intersections is shown in Table 8.

Again we reasoned that any term discovered by our procedure that had already been frequently mentioned in the migraine literature would not likely comprise a new discovery. We looked at those terms from the 230 that had an intersection frequency with migraine of seven or fewer. This analysis revealed 33 terms not found by the 12 way all-statistics analysis (Table 9). Among them was *magnesium*, the predefined target we were hoping to hit. In addition, this relative frequency analysis revealed a few items also uncovered by the 12-way all-statistics analysis: *a23187* (and its synonym *calcimycin*), *cyclooxygenase,* and *ltd4*. But, for the most part, this relative frequency analysis yielded items not identified by the all-terms analysis.

In summary, a relative frequency analysis that sought to combine evidence of novel association that was spread among different intermediate literatures successfully identified the previously known target, *magnesium*, along with other possibly interesting items related to migraine. Thus,

TABLE 9. Terms that were prominent by relative frequency analyses that were not prominent by the other lexical statistics.

Aldosterone
Apomorphine
Arachidonate
Argipressin
avp (antiviral protein)
Bicuculline
cgrp (calcitonin gene-related peptide)
CO_2
Creatine
Decarboxylase
Dinoprost
Dipeptidase
edrf (endothelium-derived relaxing factor)
Ether
Hydralazine
Hydroxydopamine
Inositol
Magnesium
Nitrendipine
npy (neuropeptide Y)
O_2
Ouabain
Oxygenase
paf (platelet aggregating factor)
Pentobarbital
Phenoxybenzamine
Pyrilamine
Tetrodotoxin
Theophylline
txa2
txb2
Xanthine
Yohimbine

Items on this list a) were prominent (top 500) on at least five of the relative frequency analyses performed on one of the twelve intermediate literatures; b) were relatively unexplored in relationship to migraine (intersection frequency of 7 or less); and c) were not uncovered by the twelve-way analysis focusing on token frequency, document frequency, and tf * idf statistics.

these results suggest another approach that can be profitably employed in literature-based discovery.

Additional Experiments

Another possible approach is to understand more completely the connection between lexically prominent phrases and a particular intermediate concept. For example, one theory of migraine is that it is caused by vasoconstriction followed by vasodilation, the latter causing the onset of pain. Prostacyclin and thromboxane act together to maintain the proper level of blood coagulation and, indirectly, dilation; following constriction; however, prostacyclin tends to bring on migraine by dilating blood vessels. Thus, using *prostacyclin* as a focus, we determined its co-occurrence frequency with other potential intermediate phrases. Items with high co-occurrence frequency with *prostacylin* presumably are conceptually and biologically related to it.

Further, in some cases it is known whether the referents of items co-occurring with *prostacyclin* tend to raise or reduce levels of prostacyclin. Since prostacyclin is implicated in migraine headaches, any of these frequent phrases that are previously bibliographically disjoint from migraine may be a new discovery.

In Table 10 we indicate 74 items uncovered by such an analysis that were either bibliographically disjoint from migraine or co-occurred with it in a single article. Items in the table whose effect is *positive* are those that increase levels of prostacyclin, while those whose effect is negative reduce it. The size of each term's intersection with the prostacylin literature is also indicated.

Discussion

When a plausible link between literatures has been detected, medical research must validate or invalidate it. If a hypothesized link turns out to be fruitful, it leads to new medical knowledge such as a treatment for a currently untreatable disease, a treatment that is more cost efficient or has fewer side effects, or an advance in basic science. Subsequent to Swanson's work on Raynaud's, DiGiacomo, Kremer, & Shah (1989) conducted clinical tests and determined that secondary, but not primary, Raynaud's could be effectively treated by fish oil. Migraine/magnesium connections have also been published after the dates used to establish the literature-based connection (Castelli, Meossi, Domenici, Fontana, & Stefani, 1993; Gallai, Sarchielli, Morucci, & Abbritti, 1993). Thus, in spite of the effort in finding these links and the chance that the quest will not be successful, searching the medical literature to suggest new discoveries may still be justified.

Davies (1989, 1990) has delineated some of the ways new literature-based, knowledge-producing techniques can serve science: By uncovering hidden refutations or qualifications; by accumulating the evidence from several independent studies that are inconclusive or of borderline significance individually (meta-analysis); by providing solutions to analogous problems in unrelated scientific disciplines; by uncovering hidden correlations; and, as in the work of Swanson and Smalheiser and our own, by drawing conclusions from two or more premises. But, as an examination of *Science Citation Index* and *Social Science Citation Index* shows, other than ourselves (Gordon & Lindsay, 1996; Gordon & Dumais, 1998) there have been few attempts (Kostoff, 1997, 1998) to replicate, extend, or modify Swanson's hypothesis discovery-seeking methods.

As we have noted, our analyses do not make use of even the limited semantic information provided by synonym classification. If we were to collapse all uses of synonyms into single categories and apply our lexical statistical methods to these categories rather than word (stem) tokens, some important phrases—those that admit of several common and distinct lexical variants (such as *erythrocyte* and *red blood cell*)—might become much more prominent. There is available a substantial thesaurus of medical terms in machine-

TABLE 10. Terms influencing prostacyclin with limited bibliographic relation to migraine.

Co-occurrence frequency with prostacyclin	Phrase	Comment	Effect
3	Acether	PAF-acether induces aggregation; antagonizes effect of PGI-2	Negative
16	Actinomycin	RNA synthesis inhibitor-actinomycin D inhibits induced PGI-2 synthesis	Negative
6	Bisphosphate	PGI-2 suppresses effects mediated by bisphosphate hydrolysis in platelets; antagonist	Negative
10	BW755C	Dual cyclooxygenase/lipoxygenase inhibitor; decreases synthesis of PGI-2	Negative
1	Calcimycin	Stimulates arachadonic acid metabolism via CO-increase in PGI-2 synthesis	Positive
14	Calmodulin	Calcium/calmodulin-dependent protein Kinase II-activates phospholipase-A-2 to release AA for prostacyclin synthesis	Positive
10	Carbachol	Little effect on PGI-2 release; however, as a parasympathomimetic, can stimulate PGI-2 synthesis with NaF, a protein activator	Positive
2	Carrageenan	Induces macrophages to produce eicosanoids-an inflammatory agent	Positive
6	Cathepsin	A protease-prevents cytosolic calcium utilization and PGI-2 synthesis induced by thrombin	Negative
5	Chorionic	Supernatant of placental villi has action which protects activity of PGI-2; human chorionic gonadotropin perhaps-villi and PGI-2 play role in maintaining placental microcirculation	Positive
13	Cicletanine	Antihypertensive drug-increases PGI-2 synthesis	Positive
3	Cotinine	Levels weakly inversely correlated to PGI-2 activity	Negative
18	CsA	Cyclosporine A-immunosuppressor-stimulates platelet aggregation, increases TxA2 synthesis and decreases PGI-2 synthesis.	Negative
36	Cycloheximide	Inhibits eicosanoid production	Negative
9	D4	Leukotriene subclass-induces vasoconstriction but increases synthesis of PGI-2	Positive
10	Dazmegrel	TxA2 synthetase inhibitor-increases PGI-2 synthesis	Positive
13	Dazoxiben	Inhibitor of thromboxane synthase-increases PGI-2	Positive
26	Defibrotide	Synthetic analogue of heparin-enhances PGI-2 production	Positive
15	EGTA	Calcium chelator-decreases PGI-2 synthesis	Negative
35	Eicosapentaenoic	Eicosapentaenoic acid: the main fatty acid of fish oil. Inhibits PGI-2 production	Negative
7	Eicosatetraenoic	15-HPETE-product of lipoxygenase other AA metabolism pathway--inhibits prostacyclin	Negative
5	Eicosatetraynoic	A xenobiotic-inhibits both CO and LO pathways	Negative
3	Eicosatrienoic	Decreases PGI-2 synthesis	Negative
18	Endotoxemia	Presence of bacterial membrane products in blood can lead to inflammation and production of PGI-2	Positive
4	Flunixin	Flunixin meglumine decreases PGI-2 synthesis	Negative
18	Fluoride	Decreases antiaggregatory activity of PGI-2	Negative
15	FMLP	Stimulates neutrophil aggregation-known stimulus for leukotriene synthesis and coronary vasospasm-significant increase in PGI-2	Positive
7	FPL	Leukotriene receptor blocker stimulates PGI-2	Positive
37	Glutathione	An essential component of glutathione peroxidase, an enzyme that protects cells by reducing intracellular peroxides-increases glutathione activity-production of PGI-2 goes up-under oxidant stress depletion of glutathione leads to PGI-2 increase	Unsure
20	H_2O_2	Hydrogen peroxide decreases PGI-2 production	Negative
13	Hydroperoxide	t-butyl hydroperoxide causes vasoconstriction, facilitates lipid peroxidation and inhibits PGI-2	Negative
34	Hydroxyeicosatetraen	12-HETE is a lipoxygenase product inhibits PGI-2 synthesis	Negative
53	Lipopolysaccharide	Bacterial cell wall constituent stimulates inflammation and PGI-2	Positive
49	Lipoprotein	Decreases PGI-2	Negative
10	LTC4	Leukotriene C4-stimulates synthesis of PGI-2	Positive
10	LTD4	Leukotriene D4-stimulates synthesis of PGI-2	Positive
5	LTE4	Leukotriene E4-stimulates synthesis of PGI-2	Positive
25	Meclofenamate	Decreases eicosanoid production	Negative
2	Meclofenamic	Cyclooxygenase inhibitor	Negative
3	Meglumine	See flunixin	Negative
12	Melittin	Stimulates PGI-2 production	Positive
8	Mepacrine	Phospholipase A-2 inhibitor	Negative
9	Monocrotaline	Monocrotaline pyrrole (MCTP)-vascular injury agent enhances release of PGI-2	
2	Mononitrate	IS-5-MN increases PGI-2 synthesis	Positive

TABLE 10. (continued)

Co-occurrence frequency with prostacyclin	Phrase	Comment	Effect
17	NaF	NaF, sodium fluoride activator of G proteins increases PGI-2	Positive
4	Nafazatrom	Cyclooxygenase inhibitor	Negative
14	NDGA	Nordihydroguaiaretic acid inhibits CO/LO pathway	Negative
16	Nordihydroguaiaretic	See ndga	Negative
14	Normoxia	During normal oxygen concentrations increased PGI-2 synthesis vs. hypoxia	Positive
7	NSAID	Non-steroidal inflammatory agent CO/LO inhibitor	Negative
37	O_2	Increases PGI-2 in pulmonary vasomotor areas	Positive
10	Ouabain	Increases PGI-2 in response to ATP	Positive
7	Oxytocin	Can stimulate PGI-2 release	Positive
7	P2Y	Purinoreceptors that mediate PGI-2 increase	Positive
8	Pentoxifylline	Antiaggregator increases PGI-2	Positive
2	Phenylbutazone	NSAID	Negative
5	Polytetrafluoroethyle	Synthesize more PGI-2 as a result of polymer on vascular grafts	Positive
9	Prothrombin	Thrombin stimulates PGI-2	Positive
12	PUFA	Polyunsaturated fatty acids-i.e. HPETE decrease PGI-2 synthesis	Negative
6	Quin	Quin-2 stimulates PGI-2 increase	Positive
14	Selenium	Decreases synthesis	Negative
4	Septicemia	Endotoxin release increases PGI-2 synthesis	Positive
10	Soybean	Low fat dietary constituent decreases synthesis	Negative
4	Streptococcal	Bacteria that induces PGI-2 production	Positive
6	Sulindac	NSAID	Negative
2	Terephthalate	Dacron sleeve constituent increases PGI-2 synthesis	Positive
11	Tetradecanoylphorbol	Phorbol ester increases PGI-2 synthesis	Positive
2	Thimerosal	Inhibitor	Negative
18	Tranylcypromine	Inhibitor of PGI-2 synthase	Negative
18	Trifluoperazine	Calmodulin antagonist-decreases release of PGI-2	Negative
8	Venom	Can cause PLA-2 activity increase	Positive
4	VLDL	Decreases PGI-2 synthesis	Negative
17	Zymosan	Activates edema stimulates endogenous AA metabolism and PGI-2	Positive

All terms above had an intersection of 0 or 1 with *migraine.* All were judged by a medical student to have an effect on prostacyclin. Terms with *positive* effects increase prostacyclin (which is implicated in migraine) and those with *negative* effects decrease it.

readable form, the National Library of Medicine's *Meta-Thesaurus*, that we are now using for this purpose.

In addition to making the MEDLINE literature a source of new hypotheses, our techniques might yield useful results when applied to nonbibliographic databases that are indexed, for example in molecular biology, genomics, and biotechnology.

For instance, the GENBANK is a database of DNA sequence data and related bibliographic and biological annotations. GENBANK is indexed by the National Library of Medicine, using sequence entry fields that are quite similar to those used in MEDLINE. Our techniques might be effective in identifying unrecognized relationships within these biotechnology databases as well as between the contents of these databases and the published literature described in MEDLINE.

There is an even larger potential gain with other medical databases, particularly in genomics. A very large percentage of the information that will find its way into various genome databases will not be published or peer reviewed. Most frequently, it will be found in working databases of Genome Centers or consortia of Genome sites. It is important that this information be shared. Our technique has the potential to make it easier and more effective for investigators to identify areas of these genomic databases that have useful information. This would require a somewhat greater degree of annotation and/or indexing than is now found in these databases. However, this degree of annotation and indexing may be necessary for the effective use of the databases—even by their creators.

Further, it should be possible for the methods we are developing to provide hypothesis discovery support for searchers on the World Wide Web. This large body of largely textual, semistructured information is growing daily, providing a new means for people both to publish their ideas and learn about the ideas of others. The breadth and range of contributions to the Web argue that many implicit, but unnoticed, connections are contained within it. Our methods may help forge these connections and give greater utility to this self-perpetuating knowledge base.

Conclusions

We have conducted experiments in which candidate leads (intermediates and targets) are generated by lexical statistics alone, and then these automatically produced lists are culled with two human filters. The first human filter (the authors) eliminates candidates on the basis of general

knowledge. The second human filter (a medical student) eliminates and combines the remaining candidates on the basis of nonspecialist medical knowledge.

The first result of this work is that our purely automatic methods placed 10 of 12 previously known intermediate topics relating migraine and magnesium into a reasonably sized list of candidates. The methods did not single out these 10 from a number of others, and so an unprimed study would necessarily need to examine many more intermediate candidates. It is, of course, possible that some of these candidates, in addition to the 10 known ones, could lead to interesting connections as well. However, we have not pursued that possibility in this article.

The second result of this work is that the previously known target term, *magnesium*, could be identified by a relative frequency analysis that considered all intermediate literatures at once, but not by a relative frequency analysis that focused on a single intermediate literature.

In addition, a number of possibly promising additional candidates were discovered by both manner of relative frequency statistics. Whether any of them is a valuable discovery is not known. One test that could be performed systematically to tell would be to find out if a term has been linked to migraine in the literature published subsequently to the literature used to generate it. At least one, *cyclooxygenase*, has recently shown an important connection in treating many pain-related afflictions. COX-2 inhibitors selectively inhibit one type of cyclooxygenase, a variety that has deleterious effects, while sparing other types that have valuable effects (protection of the kidneys and stomach). Subsequently to our uncovering of the term *cyclooxygenase* via lexical statistics, COX-2 inhibitors have become a topic of considerable interest as pain relievers in pharmacological research.

Our previous work proposed the hypothesis that relative frequency (frequency within a topic literature divided by frequency in the medical literature as a whole) would be valuable in selecting target discovery terms latent in intermediate literatures. In light of our present results, we suggest that the hypothesis be refined to consider MEDLINE base rates: High base-rate terms (such as *magnesium*) appear not to be unearthed in this way, whereas uncommon terms (such as *fish-oil*) do. In addition, we found in this work that relative frequency statistics were successful in uncovering the high base-rate term *magnesium* when these statistics pooled the results of many intermediate literatures at once. Further exploration needs to be done on these matters.

Acknowledgments

The authors would like to thank Peter Lin for his considerable help in categorizing medical terms and describing for us the physiological mechanisms of migraine, as they are currently known. We also wish to thank Praveen Pathak and David Haines for their discussions about these ideas and their assistance with programming.

References

Allen, J. (1987). Natural language understanding. Menlo Park, CA: Benjamin/Cummings.

Castelli, S., Meossi, G., Domenici, R., Fontana, F., & Stefani, G. (1993). Magnesium in the prophylazis of primary headache and other periodic disorders in children. Pediatric Medical Chirurgia, 15, 481–488.

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. The Journal of Documentation, 45, 273–301.

Davies, R. (1990). Generating new knowledge by retrieving information. The Journal of Documentation, 46, 368–372.

DiGiacomo, R.A., Kremer, J.M., & Shah, D.M. (1989) Fish oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. American Journal of Medicine, 8, 158–164.

Fayyad, U., & Uthurusamy, R. (1995) Proceedings of the First International Conference of Knowledge Discovery and Data Mining (KDD-95), Menlo Park, CA: AAAI Press.

Gallai, V., Sarchielli, P., Morucci, P., & Abbritti, G. (1993). Serum and salivary magnesium levels in migraine: Results in a group of juvenile patients. Cephalalgia, 13, 94–98.

Gordon, M.D., & Dumais, S. (1998). Using latent semantic indexing for literature based discovery. Journal of the American Society of Information Sceince, 49, 674–685.

Gordon, M.D., & Lindsay, R.K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science, 47, 116–128.

Kostoff, R.N. (1998). Database tomography for technical intelligence: A roadmap of the near-earth space science and technology literature. Information Processing and Management, 34(1), 69–85.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., & Soderland, S. (1994). Evaluating an information extraction system. Journal of Integrated Computer-Aided Engineering, 1(6), 453–472.

Lehnert, W., & Sundheim, B. (1991). A performance evaluation of text analysis technologies. AI Magazine, 12, 81–94.

Piatetsky-Shapiro, G., & Frawley, W.J.E. (Eds.). (1991). Knowledge discovery in databases. Cambridge, MA: MIT Press.

Smalheiser, N.R., & Swanson, D.R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. Neuroscience Research Communications, 15, 1–9.

Smalheiser, N.R., & Swanson, D.R. (1996a). Indomethacin and Alzheimer's disease. Neurology, 46, 583.

Smalheiser, N.R., & Swanson, D.R. (1996b). Linking estrogen to Alzheimer's disease: An informatics approach. Neurology, 47, 809–810.

Smalheiser, N.R., & Swanson, D.R. (in press). Calcium-independent phospholipase A2 and schizophrenia. Archives of General Psychiatry.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30, 7–18.

Swanson, D.R. (1988). Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine, 31, 526–557.

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

Swanson, D.R. (1990). Somatomedin C and arginine: Implicit connections between mutually isolated literatures. Perspectives in Biology and Medicine, 33, 157–186.

Swanson, D.R. (1991). Complementary structures in disjoint scientific literatures. In International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 280–289), New York, NY: ACM Press.

Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence, 91, 183–203.