Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud–Fish Oil and Migraine–Magnesium Discoveries

Marc Weeber,* Henny Klein, and Lolkje T.W. de Jong-van den Berg
Social Pharmacy and Pharmacoepidemiology, Groningen University Institute for Drug Exploration, A. Deusinglaan 1, 9713 AV Groningen, the Netherlands. E-mail: lolkje@farm.rug.nl

Rein Vos
Health Ethics and Philosophy, Faculty of Health Sciences, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, the Netherlands. E-mail: rein.vos@zw.unimaas.nl

Literature-based discovery has resulted in new knowledge. In the biomedical context, Don R. Swanson has generated several literature-based hypotheses that have been corroborated experimentally and clinically. In this paper, we propose a two-step model of the discovery process in which hypotheses are generated and subsequently tested. We have implemented this model in a Natural Language Processing system that uses biomedical Unified Medical Language System (UMLS) concepts as its unit of analysis. We use the semantic information that is provided with these concepts as a powerful filter to successfully simulate Swanson's discoveries of connecting Raynaud's disease with fish oil and migraine with a magnesium deficiency.

1. Introduction

Scientific knowledge has grown immensely in the past century. For the individual scientist, this means that he has to focus on or specialize in only a few scientific subdisciplines. To advance science, the researcher has to know and understand the current state of the art in his field(s). It is common practice to keep up to date with research reported in journals that are related to his areas of expertise. Regularly, a scientist searches in bibliographic databases for other and new developments directly related to his work. Current bibliographic databases are huge and expand at a fast rate. Information Sciences and Information Retrieval (IR) research investigate both the scientist's information needs and the bibliographic data in order to develop systems that retrieve the most relevant information. Combining the retrieved information with his own experiments and observations, the scientist creates new scientific knowledge.

During the past two decades, Don R. Swanson (University of Chicago) has advanced a different view of creating new scientific knowledge. He proposes that combining existing, though not connected, bibliographic information results in new knowledge. One publication may state the relationship between the two phenomena *A* and *B* while another reports on the relationship between the phenomena *B* and *C*. If no one has reported on the association between *A* and *C*, this association can considered to be new and may be of scientific interest. The crucial notion in this view is that two pieces of information are not related directly: there is only a *hidden* connection. One or more common aspects of the two pieces provides indirect linking. Once these links have been found, a connection can be made and new knowledge has been created.

Since 1986, Swanson has made literature-based discoveries on a regular basis in the scientific field of biomedicine. The first discoveries (Swanson, 1986, 1987, 1988, 1989) have been corroborated experimentally and clinically (Smalheiser & Swanson, 1998). The potential of Swanson's research has been widely acknowledged, but likewise its complexity (Hearst, 1999). This complexity concerns the vast information space and the possible number of connections. Swanson and Smalheiser use MEDLINE (NLM, 2000a) as their bibliographic database, with over 10 million

Received August 10, 2000; accepted December 15, 2000; Revised December 15, 2000.

* To whom all correspondence should be addressed, at the National Library of Medicine, National Institutes of Health, Bethesda, MD 20894. E-mail: weeber@nlm.nih.gov

Henny Klein is also affiliated with the Department of Humanities Computing, Center for Language and Cognition Groningen, P.O. Box 716, 9700 AS Groningen, the Netherlands. E-mail: hklein@let.rug.nl

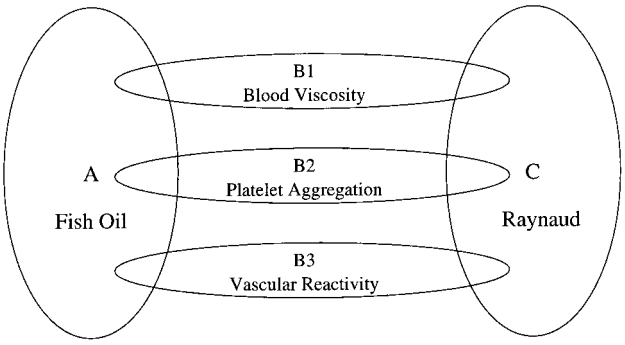

FIG. 1. Venn diagram representing Swanson's first discovery. Fish oil is beneficial for patients with Raynaud's disease through three pathways. Modified from Swanson and Smalheiser (1997) with permission.

citations of publications covering almost every scientific discipline in biomedicine. An addition to this complexity is that most information has been represented by natural language, therefore Natural Language Processing (NLP) techniques are needed to tackle the variation and intricacies of natural language.

The complexity may be the reason that only few scientists actually contribute to literature-based discovery. Michael D. Gordon and Robert K. Lindsay (University of Michigan, Ann Arbor) explore IR techniques for building discovery tools (Gordon & Lindsay, 1996; Gordon & Dumais, 1998; Lindsay & Gordon, 1999). Kenneth A. Cory (Wayne State University, Detroit) showed that Swanson's ideas are valid outside the biomedical field. He found an indirect link between a 20th century poet and an ancient philosopher in a humanities bibliographic database (Cory, 1997).

In this article, we contribute to literature-based discovery by proposing a two-step model of discovery in which new scientific hypotheses can be generated and subsequently tested. Applying advanced NLP techniques to find biomedical concepts in text, we have implemented the model in a versatile interactive discovery support tool. The semantic information available with the concepts assists the user in reducing the huge search space. The user's activities focus on the use of his expert knowledge to assess generated hypotheses. Using this tool, we are able to successfully simulate Swanson's first two discoveries, linking Raynaud's disease to fish oil and migraine to magnesium.

2. A Two-Step Model of Discovery

Swanson's first discovery was a coincidence. By reading two different literatures (for two different purposes), he made a connection between these literatures and formulated the hypothesis that fish oil may be used for treating Raynaud's disease (Swanson, 1986). Patients with Raynaud's disease suffer from intermittent blood flow in the extremities (fingers, toes, and ears). There is neither a general treatment nor cure for Raynaud's disease ("Raynaud" for short).

Swanson was able to validate his hypothesis by exploring MEDLINE extensively and by reading many scientific articles. Figure 1 shows a Venn diagram of his argument. Studying the literature on Raynaud's disease (*C*), Swanson observed that many blood- and blood vessel-related characteristics are typical for Raynaud patients: Blood viscosity and platelet aggregability are high. Also, there are vascular reactions such as vasoconstriction. These characteristics are the *B*-terms, and together they form the known *BC*-knowledge. Swanson also found that fish oil (*A*) and its active ingredient eicosapentaenoic acid (EPA) lowered blood viscosity and platelet aggregation. Additionally, fish oil may cause vasodilation. The *AB*-associations have thus been reported in the literature; however, Swanson was the first one to make the hypothetical *AC*-connection. His second discovery concerned the previously unknown association of magnesium deficiency and migraine (Swanson, 1988).

Already, in Swanson's first discovery, we may observe a two-step approach in the actual discovery process. A hypothesis has to be formulated or generated that may subsequently be validated or tested by extensive bibliographical analysis. The hypothesis can be generated in many not necessarily text-based ways. Testing a hypothesis means assessing its plausibility.[1] In his early discoveries, Swanson had formed, in one way or another, a hypothesis that he subsequently tested by extensive literature search and analysis. In later research, Swanson developed a method to also generate the hypothesis by bibliographic analysis (Swanson, 1991). We define the generating approach as an *open* and the testing approach as a *closed* discovery process.

2.1. Open and Closed Discovery Processes

An open discovery process is characterized by the *generation* of a hypothesis. Initially, there is only the scientific problem or research question and there is no idea as to where the discovery will end. In terms of Swanson, the question is, can we find a new treatment for Raynaud's disease. Figure 2 depicts the open approach beginning with disease *C*. The discoverer will try to find interesting clues (*B*), typically physiological processes that play a role in the disease under scrutiny. Next, he tries to identify *A*-terms, typically substances, that act on the selected *B*s. In the discovery process, it is likely that many *B*s and *A*s will be found. In fact, the challenge of discovery support tools is to contain the vast amount of possibilities. As a result of the process, the discoverer may form the hypothesis that substance A_n can be used for the treatment of disease *C* via pathway B_n.

A closed discovery process is the *testing* of a hypothesis. If the researcher has already formed a hypothesis about a treatment, possibly by the open discovery route described

[1] Note that our definition of testing does not include the aspect of *proof*; only experimental research in a laboratory or clinical setting can ultimately prove the formulated hypothesis.

above, he can elaborate and test it on the basis of the literature. Figure 3 depicts this approach. Starting from both disease *C* and substance *A*, the researcher tries to find common intermediate *B*-terms. The more pathways between *A* and *C* he finds, the more likely his hypothesis will be. The main difference between the two approaches concerns the literatures that are studied. In an open search, the literatures on *C* and *B* are studied; in the closed search, the literatures on *C* and *A* are studied.

Our open and closed definitions resemble respectively *procedure I* and *procedure II* in Swanson and Smalheiser (1997). Procedure I (Swanson, 1991) is complicated and difficult to automate. The web version of the ARROWSMITH discovery tool (Swanson & Smalheiser, 1997; Smalheiser & Swanson, 1998) uses a closed approach only. The user should already have formulated a hypothesis in order to upload two literature files: a disease file, *C* (Raynaud), and a dietary file, *A* (fish oil). However, opting for more general selection criteria for the *A*-file, a semi-open approach can be pursued. Swanson and Smalheiser (1999) mention general selection criteria such as dietary factors, toxins, or pharmaceutical agents. This approach is successful in finding both the *B*- and *A*-terms when simulating the Raynaud–fish oil discovery (Swanson, personal communication). Note that the *B*-literature on the intermediate physiological aspects will not be analyzed.

Gordon and Lindsay (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999) are mainly concerned with an open discovery process. In both papers, they use lexical statistics to simulate Swanson's discoveries on Raynaud–fish oil (Swanson, 1986) and migraine–magnesium (Swanson, 1988). Although they have to use different statistics in these two cases, they manage to find most of the *B*-terms found by Swanson, and they succeed in placing fish oil and magnesium at a high rank in the final list of *A*-terms.

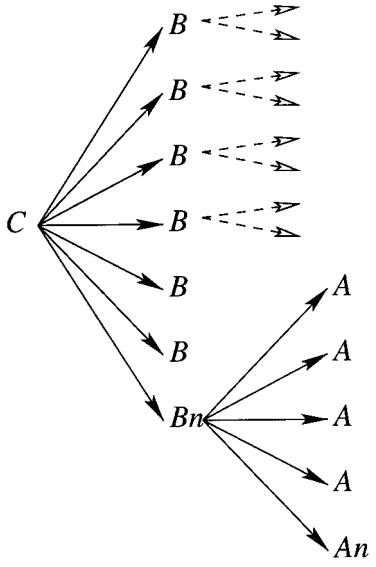

FIG. 2. Open discovery process: a one direction search process which starts at *C* and results in *A*. The solid arrows indicate potentially interesting pathways of discovery, the dashed ones unsuccessful pathways.

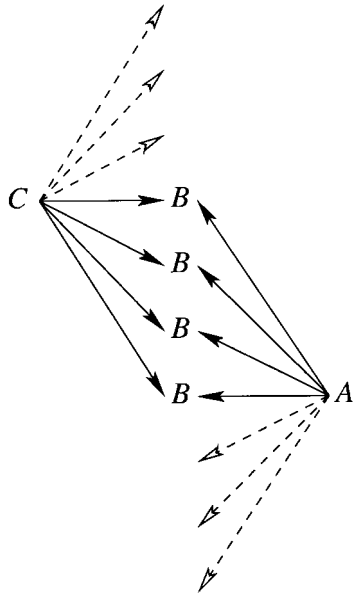

FIG. 3. Closed discovery process. The search process starts simultaneously from *C* and *A* resulting in overlapping *B*s. The solid arrows indicate potentially interesting pathways of discovery, the dashed ones unsuccessful pathways.

2.2. Open Versus Closed: Consequences for Search Spaces

In generating a hypothesis, a user knows only the starting point of his discovery. Using PubMed (NLM, 2000a), which is the free web-based interface to MEDLINE, we start an open discovery process with the initial query[2] "Raynaud's disease AND 1960:1986[dp]". We restrict the Raynaud query with "dp", date of publication, to simulate the bibliographic situation in which Swanson made is first discovery (1986). 1960 is the starting date of MEDLINE. The query results in 2,375 hits. Studying these citations, one may observe, or already know, that Raynaud affects blood and blood circulation in the extremities. A query on blood-related issues with the goal of treating the non-normal or pathologic blood characteristics will result in many hits. The query "(platelet adhesiveness OR platelet aggregation OR plasma viscosity OR blood viscosity OR blood coagulation OR erythrocyte deformability) AND 1960:1986[dp]" results in 49,332 hits. In Swanson's model, these are the references of the potential blood-related *B*-terms. Each one of these citations may lead to one or more interesting *A*-terms. Note that when using more general blood terms, we obtain many more hits. The challenge to single out fish oil as an end point is obvious. The expert's knowledge together with advanced analysis techniques such as described in the next section are needed to generate only few likely end points.

Selecting those few end points, the user can assess them in a more restricted search space (closed discovery). For

[2] Concepts are in *italics* and semantic types are in **bold**.

instance, the query on both the *A*- and *C*-terms, "(Raynaud's disease OR fish oil OR salmon oil OR cod liver oil OR eicosapentaenoic acid) AND 1960:1986[dp]", results in 3,219 PubMed citations. However, even in this restricted search space there are too many citations to assess manually. Again, NLP techniques are necessary.

The presented examples show that the information space is vast and finding an interesting endpoint is non-trivial. Both the discoverer's expert knowledge and sophisticated analytic tools should reduce the space to a workable size. In our approach, we have opted for semantic text analysis.

TABLE 1. Differences in tokens, types, and concepts for 1,246 MEDLINE citations.[a]

	Tokens	Types	Filtered types	Concept types
Single words	111,521	8,627	4,969	2,871
Bigrams	105,730	48,255	2,625	1,953
Trigrams	99,963	76,576	768	894
$n > 3$	—	—	—	280
Total	317,205	133,428	8,362	5,998

[a] Types and tokens for four types of *n*-grams. Also, results when applying Swanson's stop list and the number of UMLS concepts are tabulated.

3. Semantic Text Analysis

Literature-based discovery is about connecting disconnected entities of scientific *knowledge*. For practical reasons, the citations in MEDLINE constitute the knowledge base in which the discovery can be made. In the textual analysis of the citations, single or compound terms represent the knowledge items. Originally, Swanson used a single word as unit of his analysis; the current ARROWSMITH system, however, uses both single and multiple word terms (*n*-grams). Gordon and Lindsay use single words, bigrams (two-word terms), and trigrams (three-word terms). Using single words is an oversimplification, whereas multiple word terms are often more meaningful. However, many sequences of two or three words are available in MEDLINE citations with only a limited amount of them being meaningful, both linguistically and biomedically.

To distinguish the meaningful *n*-grams from the non-interesting ones, collocations can be computed (Church & Hanks, 1990; Weeber, Vos, & Baayen, 2000); however, Swanson and Smalheiser as well as Gordon and Lindsay use a simpler but sufficiently adequate approach. They use a *stop list*, a list of words that are not interesting for their purposes and that has been compiled gradually during the course of their research activities. General categories such as numbers, determiners, prepositions, and adverbs are excluded as well as too general biomedical terms, e.g., *aetiology*, *enzyme*, *irradiation*, or *overdose*. If one of the stop words occurs in an *n*-gram, the complete *n*-gram is discarded. In fact, the power of the ARROWSMITH system lies in its extensive stop list of about 8,000 words.

In our discovery system, we proceed beyond the lexical analysis based on the (multiword) terms in a more systematic manner. We have followed the suggestions of Swanson and Smalheiser (1997) and Lindsay and Gordon (1999) to use the Unified Medical Language System (UMLS) Metathesaurus (Lindberg, Humphreys, & McCray, 1993; McCray, Razi, Bangalore, Browne, & Stravi, 1996; NLM, 2000b). This most comprehensive biomedical thesaurus contains about 730,000 concepts in the 2000 edition that originate from over 90 source vocabularies and thesauri. Our system has to translate, or map, the raw text from the titles and abstracts to UMLS concepts. Once we have identified these concepts, we can proceed to the discovery process. The remainder of this section exemplifies the main advantages of using UMLS concepts over *n*-grams.

3.1. *Comparing* n-*Grams and Concepts*

To extract all UMLS concepts from the titles and abstracts, we use the MetaMap program (Rindflesch & Aronson, 1994; Aronson, 1996; Aronson & Rindflesch, 1997), which maps raw text to UMLS concepts. We have downloaded 1,246 MEDLINE citations in which the title or abstract contains the text word "raynaud". This collection comprises 5,782 sentences and 111,512 word tokens. The number of different word types is 8,627 (see Table 1). This means that there are 8,627 different words, many occurring only a few times (low number of tokens), others occurring often (high number of tokens). In a similar fashion, there are *n*-gram types that occur a certain number of times in the citations (*n*-gram tokens) and there are concept types that have a number of instances (tokens).

When we also consider all possible bigrams and trigrams, more than 133,000 different types are potentially interesting in our discovery process; filtering is absolutely needed. Applying the current ARROWSMITH stop list filter, which is, among other things, tuned for the Raynaud–fish oil discovery, we obtain the term types in the third column of Table 1. The result is a reduction of the number of interesting terms to 8,362 (6%). This list of single words, bi-, and trigrams consists of relatively interesting terms. For comparison, MetaMap has identified all UMLS concepts in the 1,246 MEDLINE citations. We observe only 5,998 distinct UMLS concepts in these citations. The advantage of using concepts is fourfold: first, we obtain a further reduction in types; second, we know by their origin that they are biomedically relevant; third, we are not dependent on the subject-specific focus of the stop list—this means that a discovery tool based on concepts is less prone to an explosion of different word/concept types and more robust for new applications in different biomedical research disciplines. Finally, by using concepts, we can collapse synonyms and textual variants in text to one concept type.

Interestingly, we observe that only 41% of the filtered word types are compound terms, whereas 52% of the concepts are compound concepts, i.e., consist of more than one

TABLE 2. The semantic types that form the functional semantic filter.

Semantic type
Biologic function
Cell function
Finding
Molecular function
Organism function
Organ or tissue function
Pathologic function
Phenomenon or process
Physiologic function

word. Apparently, biomedical knowledge is better described with more than one word; reducing knowledge to single words is indeed an oversimplification.

3.2. Using Semantic Information for Filtering

One of the knowledge sources of the UMLS is its semantic network (NLM, 2000b). This network consists of 134 semantic categories called *semantic types* and 54 relationships between these types. Because every concept in the Metathesaurus has been assigned one or more semantic type, the network provides a consistent, source vocabulary-independent, categorization. Examples of semantic types are: **Body location or region**, **Vitamin**, and **Physiologic function**.

To further reduce the search space in the discovery process, we filter the results of the text-to-concept mapping process by means of the semantic types. In the different stages of the process, we employ different semantic filters. For example, in the stage of selecting intermediate *B*-concepts we single out those concepts that have a functional semantic type such as **Biologic function**, **Cell function**, **Phenomenon or process**, and **Physiologic function**. Table 2 provides the functional filter that we use in our simulations. Looking for dietary factors as *A*-concepts, a typical semantic filter consists of semantic types such as **Vitamin**, **Lipid**, and **Element, ion, or isotope.**

Semantic filters are not static, but query dependent. Swanson, for instance, is interested in dietary and food deficiency aspects as a probable treatment of diseases. In our own research, we are also interested in pharmaceuticals. In the latter case, the semantic filter for *A*-concepts will at least include the semantic type **Pharmacologic substance**. Additionally, open and closed searches may employ different filters. As an open search will result in a large number of MEDLINE citations and UMLS concepts, a restrictive filter with only few semantic types is needed, whereas a closed search may benefit from a less-restricted filter.

4. The *DAD*-system and MetaMap

We have implemented the open and closed approaches to discovery as well as the semantic analysis in our discovery support system, the *DAD*-system,[3] an NLP tool that guides the user in his discovery process. The key resource the *DAD*-system uses is MetaMap.[4] This program, designed by Aronson and his colleagues at the National Library of Medicine (Rindflesch & Aronson, 1994; Aronson, 1996; Aronson & Rindflesch, 1997), maps natural language text to UMLS concepts. Table 3 provides an example analysis of the sentence "Platelet aggregation is known to be high in patients with Raynaud's disease." MetaMap uses the Xerox tagger to assign syntactic parts of speech. Using these tags, a minimal commitment parser recognizes phrases (Rindflesch, Rajan, & Hunter, 2000). These phrases are the units for which mappings to UMLS concepts are sought. First, variants are generated based on the SPECIALIST lexicon, one of the UMLS knowledge sources. The variant generation process accounts for mapping different textual variants to one UMLS concept. Variants include synonyms, derivations, and inflections. These variants are matched against the UMLS Metathesaurus to retrieve candidate concepts. Using a linguistically rigorous evaluation metric, MetaMap selects the final concepts. Table 3 shows that *Platelet aggregation*, *Platelet aggregation* ⟨2⟩, *Known*, *High*, *Patients*, and *Raynaud's disease* are the mapped concepts.

This example also shows that ambiguity is not resolved. There are two types of platelet aggregation concepts in the UMLS: one has the semantic type of **Cell function**, the other one **Laboratory or test result**. The number in angle brackets indicates that there is more than one meaning for a concept in the UMLS. Currently, MetaMap is not able to disambiguate ambiguous concepts, and we therefore include both concepts in the analysis.

The *DAD*-system is the discovery environment that supports the researcher in generating new hypotheses from the literature. The system handles the queries to PubMed, the text to concept mapping through MetaMap, and local storage and retrieval of the results. A query to PubMed always results in a full citation. Currently, we are only interested in the titles and abstracts. More specifically, we are interested only in sentences in which *A* and *B*, respectively *B*- and *C*-concepts co-occur. This co-occurrence is our definition of an association between concepts. We use sentences as the unit of analysis because a sentence often expresses the actual relationship between the concepts. Using co-occurrence within one abstract will lead to many more (most often incorrect) associations that are difficult to assess by the expert. After MetaMapping the titles and abstracts, the *DAD*-system selects only these sentences and the concepts they comprise for further analysis.

The next step in the discovery process is for the user to experiment with the settings of the semantic filters to reduce

[3] The acronym *DAD* expands to *D*isease-*A*dverse drug reaction-*D*rug, or vice versa. It represents one of the pathways we currently explore in literature-based discovery in biomedicine.

[4] More information on MetaMap and its applications is available at: http://nls9.nlm.nih.gov. Requests for using MetaMap can be directed to Dr. Alan R. Aronson (alan@nlm.nih.gov).

TABLE 3. Example of MetaMap's text to concept mapping process.

Sentence													
Platelet	aggregation	is	known	to	be	high	in	patients	with	Raynaud	'	s	disease
Tags													
Noun	noun	aux/2	verb/2	adv/2	aux/2	adv/3	prep/2	noun	prep/2	noun	ap	noun	noun
Phrases													
Platelet aggregation			known	to		high		patients			raynaud disease		
Variants													
platelet aggregation			known			high		patient		raynaud			disease
pa			know					patients		raynauds			diseases
pa's			knowing										diseased
pas			know										diss
blood platelet	aggregation		knowledge										maladie
blood platelets	aggregations		knowledgeable										mal
platelet	aggregate												mals
platelets	aggregates												morbus
	aggregating												di
	aggregated												di
													di's
													dis
													dis's
Candidate concepts													
Platelet aggregation			Known			High		Patients			Raynaud's disease		
Platelet aggregation ⟨2⟩			Knowledge								Disease		
Pa			Knowledgeable								DIS		
PAS													
% aggregation													
Aggregation, NOS													
Platelet													
Aggregate													
Final concepts													
Platelet aggregation			Known			High		Patients			Raynaud's disease		
Platelet aggregation ⟨2⟩													

the amount of non-interesting concepts. The resulting concepts are ranked. For an open discovery process, rank ordering of concepts is based on concept frequency; for the closed search, the number of links between *A* and *C* is also included. The more pathways there are between *A* and *C*, the more plausible the hypothesis is. Additionally, a higher concept frequency means that there is more contextual information for the user to evaluate the hypothesis.

To study the filtered concepts, the user can retrieve the sentences in which the interesting concepts occur. For a certain *B*, the *AB*-sentences and the *BC*-sentences are presented in a juxtaposed fashion, similar to Swanson's ARROWSMITH system. For a *B*-concept like *Blood viscosity*, the sentences "Blood viscosity and Raynaud's disease" and "Reduction in blood viscosity by eicosapentaenoic acid" are presented positioned next to each other for an easy formulation of the hypothesis that fish oil may treat patients with Raynaud's disease. If the user needs more contextual information, the full citation in PubMed is available through a hyperlink. For a technical description we refer to Weeber et al. (2000).

5. Simulating Two Literature-Based Discoveries

To show the validity of our *DAD*-system, we simulate two of Swanson's early and well-known discoveries of linking Raynaud's disease to fish oil and migraine to magnesium. We try not only to replicate Swanson's results but also to simulate the actual discovery. This means that the interesting concepts should not only be mentioned in the results, but they should also be noticeable for a biomedical expert with no prior knowledge. A high rank is an indication, for instance, many quite similar concepts that occur in the results is another one. One concept in a list of 500 is difficult to uncover, but five closely related ones, with a not-too-low ranking, will likely draw the expert's attention. Additionally, by using the two-step approach as described above, one or two strong pathways should suffice to make the initial link in the open discovery process. The closed process may then uncover additional pathways.

5.1. Raynaud–Fish Oil

Swanson found three general pathways through which fish oil acts on Raynaud's disease: platelet aggregation, blood viscosity, and vascular reactivity. When we start our analysis with Raynaud, we find 1,246 MEDLINE citations that have the word "raynaud" in their title or abstract. After MetaMapping these citations, we apply a filter of functional semantic types (Table 2) to restrict our search to only functional *B*-concepts that co-occur with the Raynaud con-

TABLE 4. The three Raynaud–fish oil pathways and their corresponding *B*-concepts.

Blood viscosity	Platelet aggregation	Vascular reactivity
Blood viscosity	Fibrinolysis	Vasodilatation ⟨1⟩
Erythrocyte deformability	Platelet aggregation ⟨1⟩	Vasodilation ⟨1⟩
Plasma viscosity level	Thrombosis	Vasospasm
Hemorheology	Platelet adhesiveness	Vasospasm mechanisms
Decreased vascular flow	Effects, blood coagulation	Vasomotion
Hyperviscosity		Decreased vascular resistance
		Decreased vascular flow

cept in one sentence. This filtering results in a list of 145 *B*-concepts. Note the reduction from 5,998 (Table 1) to 145.

Simulating a real-life situation with no prior knowledge, we will try to find the Raynaud–fish oil connection through each pathway in the open discovery process in order to find other pathways in the more restricted, closed search. Table 4 shows the concepts we used for the three pathways.

For each pathway, we start a new discovery process to find *A*-concepts. For vascular reactivity, we analyzed 9,129 MEDLINE citations; for platelet aggregation, 14,851; and for the blood viscosity pathway, 1,515 citations. We filtered the resulting concepts through a dietary filter that consists of the semantic types **Vitamin**, **Lipid**, and **Element, ion, or isotope**. For the blood viscosity pathway, this results in 61 concepts; for the platelet aggregation pathway, 215; and for the vascular reactivity pathway, 96 concepts. Table 5 shows the relevant fish oil *A*-concepts together with their ranks.

We observe that the viscosity and platelet pathways both lead to several concepts concerning fish oil and its active ingredients. Because of their number and their rank we think it not unlikely that a biomedical researcher not yet familiar with the Raynaud–fish oil connection will notice these concepts. The vasoreactivity pathways seem to lead to a dead end. We only observe general dietary fat concepts with no reference to fish oil. Without prior knowledge, we doubt that a researcher will mark this as an interesting lead on which to follow up.

When we use the fish oil concepts found via a single pathway to start a closed search by considering the **A**- and *C*-literature jointly, we also find the *B*-concepts that are relevant to the other pathways. We have thus strengthened the single pathway-generated hypothesis with the additional pathways. If we apply a broader semantic filter, i.e., the functional filter with an addition of the semantic types **Body location**, **Body part**, **Body space**, **Sign or symptom**, **Organism attribute**, and **Laboratory or test result**, even more relevant *B*-concepts are found: *Platelet aggregation* ⟨2⟩, *Blood viscosity level*, and *Whole blood viscosity level*. Interestingly, in a closed search with the general dietary fatty acid concepts found via the vasoreactivity pathway, we still can find relevant *B*-concepts: The concepts *Platelet aggregation* ⟨1⟩, *Vasospasm*, *Vasospasm mechanism*, *Plasma viscosity level*, and *Platelet adhesiveness* are among a list of 68 functional *B*-concepts.

Although the fish oil concepts did not have the top ranks in the open process results, there were many of them, and they had a reasonable rank. We think that an expert user will mark them as "interesting," and therefore a hypothesis has been generated successfully. Scrutinizing the single pathway-generated hypothesis in a closed search, we uncover the other pathways enhancing the plausibility of the hypothesis.

5.2. Migraine–Magnesium

Swanson's discovery of the relation between a magnesium deficiency and migraine headaches is a more complicated one. He found eleven pathways through which magnesium affects migraine. Starting with an open discovery process, we use 3,310 MEDLINE citations on migraine in which over 3,000 concepts co-occur in sentences with the migraine concept. After applying our functional filter, 504 concepts remain. Two pathways, vascular reactivity and platelet aggregation, are represented by many concepts, a biomedical expert will likely notice them. The first two columns of Table 6 show the relevant *B*-concepts. The other two functional pathways concern hypoxia (oxygen deficiency) and spreading cortical depression.

TABLE 5. Fish oil *A*-concepts and their ranks for each pathway in the Raynaud experiment (open discovery process).

Blood viscosity ($n = 61$)		Platelet aggregation ($n = 215$)		Vascular reactivity ($n = 96$)	
20	Fish oils	50	Eicosapentaenoic acid	56	Fatty acids, essential
21	Maxepa	54	Cod liver oil	57	Dietary fats
23	Fatty acids, omega-3	77	Fish oils		
26	Omega-3 polyunsaturated fatty acid	94	Maxepa		
32	Eicosapentaenoic acid	135	Fatty acids, omega-3		
54	Epa-e	155	Omega-3 polyunsaturated fatty acid		
		163	Salmon oil		

TABLE 6. Four functional migraine–magnesium pathways and their corresponding *B*-concepts (open discovery process).

Vascular tone and reactivity	Platelet aggregation	Hypoxia	Spreading cortical depression
Vasoconstriction	Thrombocytopenia ⟨1⟩	Oxygen deficiency	Spreading cortical depression
Vasodilitation ⟨1⟩	Platelet aggregation ⟨1⟩		
Vasodilation ⟨1⟩	Thrombosis		
Vasospasm	Platelet activation		
Vasospasm mechanism	Blood coagulation		
Oligemia	Abnormal platelets		
Spasm of cerebral arteries			
Vasomotor functions, NOS			
Peripheral vasoconstriction			

The literature for the vasoreactivity pathway consists of 6,000 MEDLINE citations; the platelet aggregation pathway, 15,099; oxygen deficiency, 11,951; and spreading cortical depression, 167 citations. After MetaMapping the titles and abstracts to UMLS concepts and filtering through the dietary semantic filter, *Magnesium* appears in the list of *A*-concepts for all but the spreading cortical depression pathway with a very high rank (third column in Table 7).

At first sight, this may appear a very promising result. However, upon closer inspection of the sentences in which *Magnesium* occurs according to our system, we observe that MetaMap maps the text strings and abbreviations "mg" (milligram) and "Mg" (magnesium) both to the concept *Magnesium*. Unfortunately, MetaMap cannot resolve ambiguous text-to-concept mappings. To obtain a better estimate of the rank of *Magnesium*, we analyzed by hand the list of sentences in which the concept *Magnesium* occurs and recomputed its rank. The fourth column in Table 7 shows the adjusted ranks. We still find magnesium in the top 15% of the rank ordered list of *A*-concepts for all three pathways.

The spreading cortical depression pathway does not lead to magnesium. Analyzing the 167 citations, we observe that in many cases textual variants of the concept, most notably "spreading depression", are not mapped to the target *B*-concept. Additionally, many citations mention magnesium in a different sentence so that the *DAD*-system does not consider magnesium associated with spreading cortical depression.

Via three different pathways, we can generate the *AC*-hypothesis of magnesium–migraine. In the closed discovery phase, we have to analyze 3,310 migraine and 7,621 magnesium citations jointly. After applying the functional filter, 253 concepts remain that co-occur with both migraine and magnesium. Table 8 on page 29 shows us that 6 of the 11 known pathways can be found. Expanding the filter to include the semantic types **Amino acid, peptide, or protein**, **Neuroreactive substance or biogenic amine**, and **Disease or syndrome**, we find 9 of the 11 pathways. Note that the concept *Spreading cortical depression* is no longer in the list of *B*-terms. In the *C*-literature, we did not find correct mappings for the same reasons why we only found few *B*-literature citations for this concept.

In the open discovery process, magnesium receives top ranks, and the migraine–magnesium hypothesis is therefore very plausible. We found this hypothesis through three different pathways. This is only a limited number compared to the results fouind by Swanson, who discovered 11 pathways. However, in a case where the user has no knowledge regarding the connection beforehand, we think that each pathway is strong enough to be followed. Once the hypothesis is tested in a closed process, six, and with an expanded, less restrictive, filter nine, pathways can be found, which adds to the plausibility of the hypothesis.

TABLE 7. Number of *A*-concepts, rank, and adjusted rank of magnesium (Mg).[a]

Pathway	No. *A*-concepts	Rank magnesium	Adj. rank magnesium
Platelet aggregation	266	3	42
Vascular tone and reactivity	114	5	10
Oxygen deficiency	206	6	29
Spreading cortical depression	8	—	—

[a] Magnesium appears at a relatively high rank for three of the four different functional pathways in the migraine experiment.

6. Discussion

In this article, we have simulated two literature-based discoveries with our discovery support tool, the *DAD*-system, which employs open and closed discovery processes and semantic text processing. We consider the simulations successful because the two hypotheses could reliably be generated via different pathways. A biomedical expert will likely consider at least one pathway interesting, and we have argued that the end points (dietary factors) are sufficiently represented and ranked to be considered an interesting hypothesis. Testing the hypotheses leads to additional pathways, thus strengthening them.

Without prior knowledge, it will be nearly impossible for a domain expert to find all pathways at once. Both Swanson's ARROWSMITH system (Swanson, personal communication) and our *DAD*-system are successful at employing a more practical two-step strategy. In their 1996 paper, Gordon and Lindsay do not pursue a single pathway strat-

TABLE 8. *B*-concepts found in a closed discovery procedure for migraine–magnesium.[a]

Concept	Semantic type(s)
1. SPREADING CORTICAL DEPRESSION	
**NONE	
2. EPILEPSY	
Epilepsy	*Disease or syndrome*
3. SUBSTANCE P	
Substance P	*Amino acid, peptide, or protein*
	Neuroreactive substance or biogenic amine
4. PLATELET AGGREGATION	
Platelet aggregation ⟨1⟩	Cell function
Platelet activation	Organ or tissue function
Thrombocytopenia ⟨1⟩	Finding
Blood coagulation	Organ or tissue function
Blood coagulation factors	*Amino acid, peptide, or protein*
	Biologically active substance
Platelet storage pool deficiency	*Disease or syndrome*
5. SEROTONIN RELEASE	
Serotonin	*Neuroreactive substance*
	Organic chemical
6. CALCIUM CHANNEL BLOCKERS	
Calcium channels	*Amino acid, peptide, or protein*
	Biologically active substance
7. STRESS AND TYPE A PERSONALITY	
Personality sensitivity	Finding
	Mental process
8. VASCULAR TONE AND REACTIVITY	
Vasoconstriction	Finding
	Organ or tissue function
Vasodilation ⟨1⟩	Organ or tissue function
Vasodilatation ⟨1⟩	Pathologic function
Vasospasm	Pathologic function
Vasospasm mechanism	Phenomenon or process
Arteriospasm	Pathologic function
Vascular permeability	Organ or tissue function
Peripheral vasoconstriction	Finding
9. PROSTACYCLIN/PROSTAGLANDIN	
NONE	
10. INFLAMMATION	
Inflammation	Pathologic function
11. HYPOXIA	
Oxygen deficiency	Pathologic function
Ischemia	*Disease or syndrome*
Cerebral ischemia, transient	*Disease or syndrome*

[a] Concepts in normal font have been found with the functional filter, the concepts in *italics* have been found by adding the semantic types **Amino acid, peptide, or protein, Neuroreactive substance or biogenic amine,** and **Disease or syndrome** to the filter.

egy. In their 1999 paper, they observe that they find many relevant intermediate (*B*) literatures, but with their methods, no single literature (or pathway) leads to magnesium. Only the combination of all known pathways places magnesium at a relatively high rank.

Neither Swanson's nor our approach employ statistical measures. The only characteristics we use are concept frequency and the number of *B*-terms to assess the strength of an *AC*-hypothesis. Gordon and Lindsay, on the other hand, use many statistical procedures. This is a very salient aspect of their research because of the domain independence. Once they have found some general statistic, it would be relatively easy for even a non-expert to use them. In our approach we find fish oil and magnesium because we limit ourselves to dietary factors (through the semantic filter) *beforehand*. This involves domain-dependent expert knowledge and interest from the user. However, Gordon and Lindsay have not shown a single statistic that works in all cases; they use a combination of different measures. In Gordon and Lindsay (1996), revisiting the Raynaud–fish oil case, they explain why the different statistics work in the different phases of the discovery process; but when applied to migraine–magnesium, the same statistics fail to work (Lindsay & Gordon, 1999).

Gordon and Lindsay (1996) view literature-based discovery not as an automatic procedure. Instead, their goal is to develop tools that can assist the human expert user in formulating new hypotheses. We agree with this point of view (Vos & Rikken, 1998). Literature-based discovery does not concern the artificial intelligence subfield of *machine discovery*, but concerns *discovery in science*, which is, according to Valdés-Pérez (1999), "the generation of novel, interesting, plausible, and intelligible knowledge about the objects of study." The aspects of novelty, interestingness, plausibility, and intelligibility of a discovery have to be interpreted and judged by human discoverers in order to be added to the scientific knowledge base. In fact, the successful use of discovery tools will depend on the communication between the system and the domain expert (Simon, Valdés-Pérez, & Sleeman, 1997). Our *DAD*-system supports the human researcher in three ways: first by restricting the search space, second by assisting in interpretation through semantic analysis, and third by providing the textual context of the hypotheses. Scientists can scan efficiently through large mounts of literature to look for new ideas or to strengthen their initial hypotheses.

7. Future Perspectives

The semantic filtering as implemented in the *DAD*-system assists in reducing the huge search space to workable dimensions. We are convinced that more and deeper semantic analysis can reduce the human workload in the process even more. In the current system, we have defined the relation between *A*- and *B*-concepts and between *B*- and *C*-concepts by their co-occurrence in one sentence. A more thorough semantic analysis seems to be promising. With the semantic types available, general predicate templates, or semantic rules, can be defined beforehand. This approach has shown to be successful in biomedical text analysis (Rindflesch et al., 2000).

Furthermore, discovery tools will profit from improved text to concept mapping. The identification of chemical names is still elementary in MetaMap, but new and very successful recognition algorithms (Wilbur et al., 1999) will be included. Also, mapping ambiguity is a problem for MetaMap and needs to be addressed, as the magnesium/milligram case has shown.

Swanson and Smalheiser's prime disease interest concerns complex diseases and syndromes. The first reason for this is that there is an actual need in finding new treatments for these diseases. Second, complex diseases will have many pathways involved, and it is the power of literature-based discovery to traverse many different biomedical disciplines. In our own research, we are also interested in finding new applications (*C*) for existing drugs (*A*) through side effects (*B*) of these drugs (Rikken & Vos, 1995).

All literature-based discovery researchers in biomedicine have used MEDLINE as their source of information, but other sources may be interesting as well: bibliographical (EMBASE, Biological Abstracts, Chemical Abstracts) and other types. Annotated genetic databases, for instance, may prove valuable. Starting at disease *C*, try to find pathways *B* which have a genetic origin *A*. With the electronic availability of many and diverse information sources, text-based discovery tools will become invaluable for scientists.

Acknowledgments

We are indebted to Alan R. Aronson and James G. Mork of the National Library of Medicine, Bethesda, Maryland for their generous help and efforts in providing the MetaMap program to us. We would also like to thank Don Swanson for the stimulating discussions and for providing the ARROWSMITH stop list used in Section 3.1.

References

Aronson, A.R. (1996). The effect of textual variation on concept based information retrieval. In J.J. Cimino (Ed.), Proceedings of the 1996 AMIA Annual Fall Symposium (pp. 373–377). Philadelphia, PA: Hanley and Belfus.

Aronson, A.R., & Rindflesch, T.C. (1997). Query expansion using the UMLS Metathesaurus. In D.R. Masys (Ed.), Proceedings of the 1997 AMIA Annual Fall Symposium (pp. 485–489). Philadelphia, PA: Hanley and Belfus.

Church, K.W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16, 22–29.

Cory, K.A. (1997). Discovering hidden analogies in an online humanities database. Computers and the Humanities, 31, 1–12.

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

Hearst, M.A. (1999). Untangling text data mining. In R. Dale (Ed.), Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (pp. 3–10). San Francisco, CA: Morgan Kaufmann Publishers.

Lindberg, D.A.B., Humphreys, B.L., & McCray, A.T. (1993). The unified medical language system. Methods of Information in Medicine, 32, 281–291.

Lindsay, R.K., & Gordon, M.D. (1999). Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50, 574–587.

McCray, A.T., Razi, A.M., Bangalore, A.K., Browne, A.C., & Stravi, P.Z. (1996). The UMLS knowledge source server: A versatile internet-based research tool. In J.J. Cimino (Ed.), Proceedings of the 1996 AMIA Annual Fall Symposium (pp. 164–168). Philadelphia, PA: Hanley and Belfus.

National Library of Medicine (NLM). (2000a). PubMed. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

National Library of Medicine (NLM). (2000b). Unified medical language system knowledge sources. Available at: http://umlsks.nlm.nih.gov/

Rikken, F., & Vos, R. (1995). How adverse drug reactions can play a role in innovative drug research. Pharmacy World & Science, 17, 195–200.

Rindflesch, T.C., & Aronson, A.R. (1994). Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In J.G. Ozbolt (Ed.), Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care (pp. 240–244). Philadelphia, PA: Hanley and Belfus.

Rindflesch, T.C., Rajan, J.V., & Hunter, L. (2000). Extracting molecular binding relationships from biomedical text. In M. Meteer (Ed.), Proceedings of the 6th Applied Natural Language Processing Conference (pp. 188–195). New Brunswick, NJ: Association for Computational Linguistics.

Simon, H.A., Valdés-Pérez, R.E., & Sleeman, D.H. (1997). Scientific discovery and simplicity of method. Artificial Intelligence, 91, 177–181.

Smalheiser, N.R., & Swanson, D.R. (1998). Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine, 57, 149–153.

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

Swanson, D.R. (1987). Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Science, 38, 228–233.

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. (1991). Complementary structures in disjoint science literatures. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghavan (Eds.), Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (pp. 280–289). New York: 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.

Swanson, D.R., & Smalheiser, N.R. (1999). Link analysis of MEDLINE titles as an aid to scientific discovery: Using arrowsmith as an aid to scientific discovery. Library Trends, 48, 48–59.

Valdés-Pérez, R.E. (1999). Principles of human computer collaboration for knowledge discovery in science. Artificial Intelligence, 107, 335–346.

Vos, R., & Rikken, F. (1998). Connecting disconnected structures: The modelling of scientific discovery in medical literature databases. In P. Ahrweiler & N. Gilbert (Eds.), Computer simulations in science and technology studies (pp. 91–101). Heidelberg, Germany: Springer.

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., de Jong-van den Berg, L.T.W., & Vos, R. (2000). Text-based discovery in biomedicine: The architecture of the DAD-system. In J.M. Overhage (Ed.), Proceedings of the 2000 AMIA Annual Fall Symposium (pp. 903–907). Philadelphia, PA: Hanley and Belfus.

Weeber, M., Vos, R., & Baayen, R.H. (2000). Extracting the lowest-frequency words: Pitfalls and possibilities. Computational Linguistics, 26, 301–317.

Wilbur, W.J., Hazard, G.F., Divita, G., Mork, J.G., Aronson, A.R., & Browne, A.C. (1999). Analysis of biomedical text for chemical names: A comparision of three methods. In N.M. Lorenzi (Ed.), Proceedings of the 1999 AMIA Annual Fall Symposium (pp. 176–180). Philadelphia, PA: Hanley and Belfus.