LITERATURE-RELATED DISCOVERY: A REVIEW

Ronald N. Kostoff, Ph.D., Office of Naval Research, 875 N. Randolph St, Arlington, VA 22217
Internet: kostofr@onr.navy.mil

Joel A. Block, M.D., Section of Rheumatology, Rush Medical College
Rush University Medical Center
1725 W. Harrison St., Suite 1017, Chicago, IL 60612

Jeffrey L. Solka, Ph.D., Naval Surface Weapons Center Dahlgren Division
Dahlgren, VA 22448-5100

Mr. Michael B. Briggs,
Arlington, VA 22204

Mr. Robert L. Rushenberg, DDL-OMNI Engineering, LLC
8260 Greensboro Drive, McLean, VA 22102

Mr. Jesse A. Stump
Catonsville, MD 21228

Mr. Dustin Johnson
Arlington, VA 22201

Terence J. Lyons, M.D.
Air Force Office of Scientific Research
Arlington, VA 22203

Jeffrey R. Wyatt, Ph.D., DDL-OMNI Engineering, LLC
8260 Greensboro Drive, Mclean, VA 22102

DISCLAIMER

(The views in this report are solely those of the authors, and do not necessarily represent the views of the Department of the Navy, Department of the Air Force, Rush University Medical Center, or DDL-OMNI Engineering, LLC)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
05 NOV 2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Literature-Related Discovery: A Review	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research,Ronald N. Kostoff, Ph.D,875 N. Randolph St.,Arlington,VA,22217**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Discovery in science is the generation of novel, interesting, plausible, and intelligible knowledge about the objects of study. Literature-related discovery (LRD) is the linking of two or more literature concepts that have heretofore not been linked (i.e., disjoint), in order to produce novel interesting, plausible, and intelligible knowledge (i.e., potential discovery). Two major variants of LRD are: 1) open discovery systems (ODS), where one starts with a problem and generates a potential solution (or vice versa) and closed discovery systems (CDS), where one starts with a problem and a potential solution, and generates linking mechanism(s).

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**58**	

KEYWORDS

Discovery; Innovation; Science and Technology; Text Mining; Literature-Based Discovery; Literature-Assisted Discovery; Information Retrieval; Interdisciplinary; Multidisciplinary; Raynaud's Phenomenon; Cataracts; Parkinson's Disease; Multiple Sclerosis; Solicitations.

ABSTRACT

Discovery in science is the generation of novel, interesting, plausible, and intelligible knowledge about the objects of study. Literature-related discovery (LRD) is the linking of two or more literature concepts that have heretofore not been linked (i.e., disjoint), in order to produce novel, interesting, plausible, and intelligible knowledge (i.e., potential discovery). Two major variants of LRD are: 1) open discovery systems (ODS), where one starts with a problem and generates a potential solution (or vice versa); and closed discovery systems (CDS), where one starts with a problem and a potential solution, and generates linking mechanism(s).

This report reviews the state-of-the-art in ODS LRD only. It examines the major LRD concepts, evaluates each concept in detail from the perspective of discovery capability, and examines the level of potential 'discovery' reported in the literature from each concept's implementation. In the evaluation of potential discovery claimed in the published literature, a 'vetting' process is used that requires both characteristics of ODS LRD are present in order for potential discovery to be affirmed: concepts are linked that have not been linked previously, and novel, interesting, plausible, and intelligible knowledge is produced.

The major conclusions are that, until recently, most of the reported ODS LRD techniques had not generated discovery, and this lack of discovery had hampered the growth of ODS LRD substantially. However, ODS LRD techniques have been developed that allow significantly greater amounts of potential discovery to be generated systematically.

DEFINITIONS

Discovery is ascertaining something previously unknown or unrecognized. More formally, "Discovery in science is the generation of novel, interesting, plausible, and intelligible knowledge about the objects of study" [Valdes-Perez, 1999]. It can result from uncovering previously unknown information, or synthesis of publicly available knowledge whose independent segments have never been combined, and/ or invention. In turn, the discovery could derive from logical exploitation of a knowledge base, and/ or from spontaneous creativity (e.g., Edisonian discoveries from trial and error). [Kostoff, 2003]. Innovation reflects the metamorphosis from present practice to some new, hopefully "better" practice. It can be based on existing non-implemented knowledge. It can follow discovery directly, or resuscitate dormant discovery that has languished for decades.

Literature-related discovery (LRD) is a systematic approach to bridging unconnected disciplines based on text mining procedures. LRD allows potentially ***radical*** discovery to be hypothesized using either the technical literature alone, or the literature and its authors.

In the LRD context, **discovery is linking two or more literature concepts that have heretofore not been linked** (i.e., disjoint), **in order to produce novel, interesting, plausible, and intelligible knowledge**. Thus, simply linking two or more disparate concepts is a necessary, but not sufficient, condition for LRD. In particular, concepts may be disjoint because the value of their integration has not been recognized previously, or they may be disjoint because there appears to be little value in linking them formally. Examples of the latter (which had been claimed as potential discovery) will be shown in this report.

There are two types of discovery approaches commonly used in LRD: open discovery systems (ODS) and closed discovery systems (CDS). Only the **ODS** types of discovery approaches, where one starts with a problem and arrives at a solution, will be considered. These are perceived to be more challenging (because of their open-endedness) than the **CDS** types of discovery approaches, where one starts with a problem literature and a solution literature, and tries to understand the intermediate mechanisms that link the two literatures.

Also, in the LRD context, innovation is the exploitation of a discovery linkage, mainly the identification of a linkage that was not being exploited at a sufficient pace or magnitude (based on subjective evaluations). As will be shown in this report, many of the claimed potential discoveries are at best potential innovations, for the generic concept connections had been identified previously.

Discovery can also be sub-divided into incremental and radical. Incremental discovery reflects small steps into the unknown, with typically commensurate payoffs. Radical discovery depends on the source of the inspiration and/or the magnitude of the impact. Potential discovery becomes more radical when 1) the source of ideas becomes more disparate from the target problem discipline and 2) the magnitude of change/impact resulting from the discovery becomes greater. The emphasis of the present report is on the former, identifying myriad disparate sources of ideas using text mining principles (where text mining is the extraction of useful information from large volumes of text).

There are two main LRD methods for extrapolating knowledge and insights from one discipline/ technology to another: ***literature-based discovery (LBD)*** and ***literature-assisted discovery (LAD)***. The ***LBD*** approach uses technical experts to access and examine the literature from 'external' disciplines to help solve problems in the 'internal' discipline. The main LBD focus is finding potential discovery from literature analysis. The ***LAD*** approach uses technical experts from 'external' disciplines in a variety of interactive and/ or independent creative modes for the same purpose. The main LAD focus is finding potential discovery from the literature's authors.

INTRODUCTION

This report will focus mainly, but not exclusively, on ODS LBD. Both the ODS LBD and ODS LAD concepts have been described in detail in Kostoff [2006]. This report will critically review the ODS LBD literature, and one concept from the almost non-existent ODS LAD literature.

ODS LBD first surfaced in Swanson's 1986 pioneering paper on potential treatments for Raynaud's Phenomenon (RP) [Swanson, 1986]. ODS LBD has powerful capabilities intrinsically; given all the disparate medical/technical disciplines and their literatures, and the number of possible connections among all these disciplines, there is much opportunity

for potential discovery. However, one needs to distinguish between a concept's potential and its implementation. While we believe the ODS LBD concept has much to offer, we believe implementation has not yet begun to exploit the potential. One would have expected that, after two decades, there would be treatments proposed for major chronic diseases, similar implementations for their non-medical equivalents, as well as major sponsored research programs on ODS LBD. As far as we know, no major benefits resulting from these ODS LBD studies have yet to be realized.

The focus of the present review will not be a handbook-style recitation of the mechanics of the burgeoning number of techniques that address variants of ODS LBD. Rather, in order to surface the root causes of the lack of ODS LBD progress, we will focus on some of the major studies that have been reported, and address two aspects of each study; how well the underlying ODS LBD ***concept*** promotes and contributes to potential discovery, and whether the ***claimed*** potential discovery is truly potential discovery. In particular, we will show that a major roadblock to wider-scale acceptance of ODS LBD has been its inability to generate potential discovery, systematically. We will also show emerging ODS LBD methods that can generate potential discovery, systematically, on a wide scale, and have the capability for fulfilling the promise of ODS LBD mentioned previously.

There have been many papers written since 1986 that could be categorized as ODS LBD. This report will focus on the most well known and well regarded papers from the body of literature that originated with Swanson's 1986 paper. Much (not all) of the literature reviewed for this report is drawn from highly cited documents that cite both Swanson's 1986 paper and later related works, and that also cite succeeding generations of citing documents.

The general theory behind Swanson's ODS LBD approach, applied to two separate literatures, is based upon the following considerations [Swanson, 1986].

Assume that two disjoint literatures can be generated, the first literature AB having a central theme "A" and sub-themes "B," and the second literature BC having a central theme(s) "B" and sub-themes "C." Further assume that linkages can be generated through the "B" themes that connect both literatures (e.g., AB-->BC). Those linkages that connect the disjoint components of the two literatures (e.g., the components of AB and BC whose intersection is zero) are candidates for discovery, since the disjoint

themes "C" identified in literature BC could not have been obtained from reading literature AB alone.

One interesting ‘discovery’ from Swanson’s initial paper was that dietary eicosapentaenoic acid (theme "A" from literature AB) can decrease blood viscosity (theme "B" from both literatures AB and literatures BC) and alleviate symptoms of RP (theme "C" from literature BC). There was no mention of eicosapentaenoic acid in the RP literature, but the acid was linked to the disease through the blood viscosity themes in both literatures.

A central issue with all the ODS LBD studies that have been reported in the open literature (including Swanson’s) is the absence of a ‘gold standard’ that can be used as a basis of comparison [Ganiz et al, 2005]. A true ‘gold standard’ would allow comparisons of quality and quantity of potential discoveries. Many of the studies that followed Swanson’s pioneering Fish Oil paper used Swanson’s results (Fish Oil and Eicosapentanoic Acid) as a comparison standard. As we point out later in the discussion of Swanson’s initial results, we have questions as to whether Swanson’s hypotheses are potential discoveries or potential innovations. In other words, was Swanson the first to link fish oil/eicosapentaenoic acid to the treatment of RP, or had the linkage been made previously, with Swanson’s observations serving to accelerate the use of fish oil/eicosapentaenoic acid to treat RP? In any case, his results give no indication of the extent of discoveries possible.

In science, if we want to estimate the quality of a predictive tool, we have two main choices. If we have an exact solution to the problem, we can compare the predictive tool’s solution to the exact solution, and estimate the error as the difference between the exact solution and the predictive tool solution. Alternatively, if we have some way of estimating the error that accompanies a predictive tool solution, we can estimate the accuracy by that approach.

For discovery identification, we don’t know the extent of discovery possible for any problem, and therefore are not able to estimate the comprehensiveness of any approach (recall or sensitivity). Further, we are not able to estimate the quality of any discovery until much testing has been done, which means that a long time will be required before we can state definitively the fraction of estimated potential discoveries that are real potential discoveries (precision or specificity).

Therefore, any 'gold standards' would have to be individual examples of discovery predictions that have been validated. For the ODS LBD approaches reported in the literature, we have serious questions about whether any of them have generated potential 'discovery'. In the ODS LBD Approaches section of this report, we will address some examples of reported potential 'discovery' about which we have some concern.

Most of the techniques to be reviewed in this report contain intrinsic similarities to Swanson's basic approach, with their main differences being in the types of variables employed and in the mechanics of implementation. For variables, some studies use Title words (as did Swanson in the initial Fish Oil paper), some use Abstract terms, some use thesaurus-standardized terms, and some use MeSH terms. Some, like Latent Semantic Indexing, transform from Abstract terms to 'concepts, but the basic variables are still the Abstract terms. The review will address the strengths and weaknesses of using these different variable types.

The major challenges used as part of the review criteria are based on the above description of Swanson's approach, and are summarized below.

1. Definition of Core Literature

If the core problem literature is designated as 'C' (the RP literature in the example above), the solution literature is designated as 'A' (the Fish Oil/ Eicosapentaenoic Acid literature in the example), and the intermediate literature connecting 'A' and 'C' is designated as 'B' (the blood viscosity literature in the example), then how well-defined or bounded is the core literature 'C'? Since one validation check (vetting) for a potential discovery 'A' is its absence in the core problem literature 'C', the boundaries established for 'C' are critical, including the time frames and breadth of databases selected.

2. Identification of Intermediate Literature(s)

The intermediate literature(s) 'B', whether defined explicitly or implicitly, stems from the core problem literature 'C'. Since 'C' can contain tens of thousands of documents and hundreds of thousands of phrases, there are many options for selecting intermediate literatures. There is nothing that excludes multi-step intermediate literatures (e.g., C -→ B1 -→B2 -→ B3 -→ A), further compounding the problem and the options. A major challenge is

identifying the key intermediate literatures ‘B’ that will lead to promising solution literatures ‘A’ in a systematic and efficient manner.

3. Identification of Solution Literatures

Typically, the intermediate literatures ‘B’ are many times larger than the problem literature ‘C’. There can be hundreds of thousands of records that could contain potential discovery, with perhaps one million or more associated phrases. A major challenge is identifying a systematic approach that will sift through this enormous literature and extract nuggets of discovery.

Two central issues associated with this ‘mining’ process are defining the characteristics of discovery, and insuring that the techniques used to search for potential discovery are compatible with the characteristics of discovery. Much of the critique in this report will focus on how the above issues were addressed in the discovery approaches reviewed. Our definitions of discovery and innovation presented initially included their characteristics, and how discovery differs from innovation. We believe a number of reported potential ‘discoveries’ in the ODS LBD literature might be categorized more appropriately as potential innovations. In any case, the strength/ importance of the potential discovery should be addressed.

LBD APPROACHES

I. Title/ Abstract Words/ Phrases/Some MeSH

A. SWANSON’S INITIAL APPROACH - 1986

The concept that spawned ODS LBD was developed by Swanson starting in 1986 and continuing today. The main focus of the concept’s evaluation will be on the initial 1986 paper [Swanson, 1986], since it has served as a model for many future ODS LBD studies, and its principles haven’t changed very much in their incorporation in succeeding ODS LBD studies. His fundamental ODS LBD concept was outlined in the Introduction.

Swanson used two main conditions in the initial ODS LBD paper for ranking phrases. Because these conditions have been used in full or part by other ODS LBD research groups, and because we are concerned that their implementation in many studies has overly constrained the identification of

potential discoveries, these conditions will be presented verbatim from a later more comprehensive paper [Swanson and Smalheiser, 1997]:

1. In Swanson's terminology, C is a "source literature" [migraine, in his example], "B-terms" are "intermediate title words or literatures", and A-words are "title words that can represent promising target literatures". His first main condition states: "Each [A] candidate is then assigned a rank *according to the number of different B-words in the AB-BC co-occurrence linkages in which it participates* ….. This ranking algorithm is based on a presumption that the *greater the number of B-linkages, the greater the chance that some of them will be biologically important*."

2. "Each of the remaining B-word candidates is then searched in MEDLINE to determine the total number of titles in which it occurs. Restriction (ii): these words are further screened automatically to *retain only those that occur with greater relative frequency in migraine titles than in titles from MEDLINE as a whole.* The latter frequency is determined from the information displayed in the online search which shows each search statement and the corresponding number of items found. More specifically, we retain only words for which *the probability is small that a random allocation of words to titles could lead to a number of co-occurrences with "migraine" equal to or greater than the observed number*."

The first condition assumes the expanded literature can be separated into multiple orthogonal literatures, where each literature addresses a major thrust of the problem being addressed. It then assumes that a word or phrase found in multiple literatures (that categorize the problem of interest) is a higher priority discovery candidate than a word or phrase that was found in a single literature. In other words, the more characteristics of a problem that the word or phrase addressed, the higher the probability it could lead to discovery. For the RP example, if three important medical thrusts or characteristics are blood viscosity, platelet aggregation, and vasodilation, then a word or phrase that occurred in all three expanded literatures would rank higher than a word or phrase that occurred in only one literature.

The second condition assumes that the more a word or phrase stands out in an expanded literature relative to background because of its occurrence frequency, the more evidence exists that there is a stronger tie between the word or phrase and the problem, and the more likely the word or phrase is to lead to discovery. Swanson used these two conditions to essentially filter

out most of the phrases in the expanded literature, and the discovery phrases he identified were at or near the top of the rankings when these two conditions were applied.

While application of these two conditions reduces the number of candidates to be considered drastically, it is our opinion that much potential discovery is also being eliminated. Further, it is our opinion that these two conditions are completely arbitrary, and there is no theoretical basis to expect them to preferentially identify discovery in general.

Two simple examples will illustrate why we find these conditions problematical, at least for the RP problem. Suppose there are three main themes for RP: blood viscosity, platelet aggregation, vasoconstriction. Suppose further that by 2025 treatments have been found for RP, and the optimal treatment is found (hypothetically) to be Vitamin L to reduce blood viscosity, Vitamin M to reduce platelet aggregation, and Vitamin N to reduce vasoconstriction. Then, rather than researchers having looked for one substance that would impact all three main themes, or even two themes, the optimal trajectory to discovery would have been researchers looking for three separate substances. Only in the case where it was found (hypothetically) in 2025 that the optimal treatment for RP was one substance (to address all three main themes above) would the optimal trajectory to discovery have been researchers applying Swanson's first condition. Thus, unless the optimal treatment was known beforehand, prioritizing a potential discovery candidate by the number of problem characteristics it impacts serves no useful purpose, and over-constrains the solution.

However, we should not underestimate the importance of substances that impact a number of problem characteristics. For example, dehydration (water deficiency) will result in a number of symptoms/biomarkers, and correction of this deficiency will eliminate the symptoms. For this type of causal situation, substances that impact multiple symptoms/biomarkers should certainly receive high consideration. That is a different statement from requiring that number of problem characteristics impacted be imposed as a generic filtering/ranking condition.

The second condition of high relative frequency appears counter-intuitive to us, and we believe it leads away from discovery, not towards discovery. Because this condition has appeared ubiquitously in almost every ODS LBD

study successive to Swanson's, we will devote some discussion to its consequences at this point.

In our own (and other) studies of the RP problem, one of the highest relative frequency discovery candidates is Fish Oil. It was the central hypothesized discovery of Swanson's initial ODS LBD paper. The high frequency is based on many papers having been written about laboratory experiments and clinical trials that show Fish Oil reduces blood viscosity and platelet aggregation.

Realistically, it is difficult to believe that so many researchers would be involved in these Fish Oil-blood viscosity/platelet aggregation experiments and not one would recognize the link between reducing blood viscosity/platelet aggregation and treating RP.

> ***In fact, we believe the linkage was essentially recognized in print! We assert that Fish Oil is an incremental potential discovery at best, and could have been classified as a potential innovation, even though it has been presented in multiple ODS LBD studies as a potential discovery for the treatment of RP.***
>
> *Starting in the late 1970s, there was an explosion of articles in the medical literature that addressed the anti-thrombotic and circulatory enhancement effects of fish, fish oil, and constitutent acids (e.g., "We suggest that the mechanism behind this reduction was a changed balance between pro- and anti-aggregatory prostaglandins towards the anti-aggregatory side, caused by eicosapentaenoic acid from fish lipids" (Bang and Dyerberg, 1981); "The results suggest that dietary supplementation with fish oil may be beneficial in reducing myocardial damage associated with coronary artery thrombosis"(Culp et al, 1980); "The present findings suggest that moderate dietary supplements of fish oil may be beneficial in the prophylactic treatment of ischemic cerebral vascular disease".(Black et al, 1979); "These findings suggest that a diet rich in omega-3 polyunsaturated fatty acids, such as eicosapentaenoic acid, will reduce platelet/vessel-wall interaction and may reduce the risk of ischaemic heart disease" (Hay et al, 1982); "We conclude that the consumption of as little as one or two fish dishes per week may be of preventive value in relation to coronary heart disease." (Kromhout et al, 1985). Prior to that time, fish oil articles addressed first the*

properties of fish oil, and then the impact of fish oil as a food supplement on livestock.

Thus, it was well known by the late 1970s-early 1980s that fish oil and its constituents had a positive impact on thrombotic, arteriosclerotic, and other circulatory disorders. In this time frame, at least six articles suggested a link between fish oil or its constituents and vascular diseases.

"Eicosapentaenoic acid and prevention of atherosclerosis" (Angelico and Amodeo, 1978); "This finding suggests that, in vivo, high levels of E.P.A. and low levels of A.A. could lead to an antithrombotic state in which an active P.G.I3 and a non-active T.X.A3 are formed. Eskimos have high levels of E.P.A. and low levels of A.A. and they also have a low incidence of myocardial infarction and a tendency to bleed. It is possible that dietary enrichment with E.P.A. will protect against thrombosis" (Dyerberg et al, 1978); "Evidence for the mechanism by which eicosapentaenoic acid inhibits human platelet aggregation and secretion - implications for the prevention of vascular disease." (Jakubowski and Ardlie, 1979); "Dietary use of a fatty acid like eicosapentaenoic acid (which would be the precursor for a delta17-prostacyclin (PGI3) but is transformed by the platelets into nonaggregating thromboxane A3) might have beneficial effects as antithrombotic therapies" (Moncada and Vane, 1979); "Modification of blood rheology by dietary omega-3 fatty acids is of potential value in the treatment of vascular disease" (Cartwright et al, 1985).

Finally, "…. in patients with peripheral arterial disease …. It is concluded that rheological changes that result from a diet rich in eicosapentaenoic acid may contribute to the suggested protective effects of such a diet against arterial disease and that such changes are of potential therapeutic importance in established arterial disease." (Woodcock et al, 1984). ***What more is needed for establishing prior art than the title of Woodcock et al's paper: "Beneficial effect of fish oil on blood viscosity in peripheral vascular disease"***

While none of these papers mentioned RP specifically, how much of a leap is it from peripheral vascular disease to RP? For example, [SIGN, 1998] lists drug therapies for peripheral vascular disease, and presents this

information in two categories: intermittent claudication and Raynaud's Disease. Additionally, most of the hospital Web sites we examined list Raynaud's Disease under 'peripheral vascular diseases'. Thus, depending on how broadly the core RP literature is defined, Fish Oil may or may not have been a potential discovery. The oversight in the literature appears due more to poor indexing (not adding the Mesh term Raynaud's Disease to at least some of these articles) and inadequate retrieval (not using synonyms for RP) than the lack of information availability within the medical community.

Therefore, the 'discovery' of Fish Oil and Eicosapentaenoic Acid by literature-based techniques is an incremental discovery at best, since the linkages between peripheral vascular disease and Fish Oil had already been shown in openly-published literature. The revolutionary discovery was proposing/demonstrating the linkage between the ingestion of fish (and/or its constituents) and its positive effects on the circulatory system. While it is difficult to pinpoint a specific discovery date, certainly articles suggesting such a linkage were appearing in the open literature in the late-1970s or before.

These remarks are not meant to denigrate Swanson's ODS LBD concept. We believe the fundamental concept remains valid, and a major step forward using the literature as a basis for hypothesizing discovery. However, we believe there is an incompatibility between discovery and **extensively reported research** in **directly related** literatures. We believe discovery will have higher probability if **indirectly related** literatures are accessed, or perhaps **isolated low frequency phenomena** are found in **directly related** literatures. Intuitively, one would expect that output from one or two researchers who are at least somewhat (and preferably very much) removed from the core literature research would have a better chance of resulting in discovery through a literature-linking mechanism.

These two conditions constituted the primary numerical filters used by Swanson to narrow the pool of candidate discovery phrases to more manageable numbers. Either the same, or related, numerical conditions have been used by later ODS LBD researchers for similar filtering purposes. We do not believe these are either necessary or sufficient conditions for discovery. They represent attempts to apply information technology principles to automation of literature-based discovery, but there is no evidence that they are associated with discovery. There is nothing to rule

out a discovery word or phrase being associated with a high value of either of these filters, but the same could apply to a non-discovery word or phrase. There is also nothing to rule out a discovery or non-discovery word or phrase being associated with a low value of either of these filters.

Finally, for 'vetting' his potential discovery candidates as being independent from the core RP literature, Swanson used a co-citation analysis approach to insure that RP and the potential discovery concept were not being co-cited in the same paper. His approach was highly intensive manually, which probably motivated the need for developing filtering metrics. His database was mainly MEDLINE records published between 1975 and mid-1985. He did not seem to have a formal approach for generating the intermediate B literatures.

Most of Swanson's later work appears to be focused on CDS LBD problems for the medical literature. He and Smalheiser [Swanson and Smalheiser, 1997, Smalheiser and Swanson, 1998] developed a software system called Arrowsmith to simplify use of his closed system approach. Because Arrowsmith could be used in a quasi-open system mode (quasi-ODS), we summarize its operation.

B. SWANSON-SMALHEISER – 1997, 2005 - ARROWSMITH

Arrowsmith operation is described in detail in two major references [Swanson and Smalheiser, 1997; Smalheiser, 2005], and its improved selection of 'B' literatures (terms) is described in a more recent paper [Torvik and Smalheiser, 2007]; we select the 2005 paper for purposes of this review. Arrowsmith presently operates from the MEDLINE database only, and therefore can generate only medical potential discovery. The user initiates Arrowsmith operation by conducting two PubMed searches for literatures 'C' and 'A'. The Arrowsmith software then stems the titles of the papers in each literature, and makes a list of all single, double, and triple word phrases that are found in common in the titles of both literatures. The resulting raw 'B' list is then filtered and ranked further before being displayed to the user. The user can then examine A -→ B -→ C list linkages to identify credible paths that show how the solution impacts the problem. While Arrowsmith presently uses titles only, it could be upgraded to include Abstracts. It could also be upgraded to include additional literature links (e.g., 'A', 'B', 'C', 'D'). Because of the constraints imposed by the requirement for exact phrase matching between the 'AB' and 'BC'

literatures, adding the capability for using synonyms would seem to offer a major step forward as well. In addition, MeSH terms have been integrated into the matching process and the title display.

There can be many raw phrases on the intermediate 'B' list. Therefore, much effort has been expended with Arrowsmith to reduce the number of phrases on the 'B' list. Seven filters are listed in [Smalheiser, 2005]. The value of these filters is that they allow the user to insert or remove assumptions that may impact the level and quality of potential discovery. In particular, assumptions made by Swanson and his successors about discovery phrase frequency above MEDLINE background and desirability of maximal pathways connecting 'C' and 'A' literatures can be removed, and the full range of potential discovery candidates can be examined.

Filter 1 allows 'B' terms to be limited to desired UMLS-based semantic categories. Filter 2 allows users to examine 'B' terms that only occur more (or less) than a certain number of times in either the 'A' or 'C' literatures. Filter 3 allows the user to examine 'B' terms that appeared for the first time more recently than (or only earlier than) a given year in either literature. Filter 4 removes 'B' terms if the 'A' and 'C' literatures have either zero or less than a threshold number of MeSH terms in common. Filter 5 merges highly related terms within the same semantic category into a single composite 'B' term, using a statistical model of term co-occurrence within title or Abstract fields. Filter 6 removes terms that are not 'characteristic' (characteristic terms in title or Abstract fields occur in the 'A' or 'C' literatures significantly more often than in MEDLINE as a whole). Filter 7 gives higher ranking to the most cohesive 'B' terms (terms that represent the more narrowly focused literature). Torvik and Smalheiser [2007] described a logistic regression model that allowed estimation of the probability of relevance for each 'B' term, that allowed 'B' term rankings according to their likely relevance, and allowed estimation of the overall number of relevant 'B' terms inherent in a given two-node search.

Arrowsmith can operate in a quasi-ODS discovery mode as follows. If the semantic categories of solutions are known beforehand, then a query can be generated to retrieve all records that include members of the desired semantic categories. This would constitute the 'A' potential solution literature. The 'C' problem literature would also be specified. Then, all 'A', 'B', 'C' links would be identified and the most credible paths (and solutions) could be selected. It would allow an alternative mechanistic approach to the

technique described in section 1-D of this review. We are presently using this mode of Arrowsmith operation to supplement our ODS LBD approaches.

C. GORDON AND LINDSAY - 1996

The next article that addressed the Swanson-based ODS LBD problem was [Gordon and Lindsay, 1996]. They used an information technology-based approach (especially information retrieval), with the goals of introducing some automation to the ODS LBD process and replicating the Fish Oil 'discovery' results. In parallel to Swanson, they used the A-B-C literature philosophy for discovery. In contrast to Swanson's use of words from titles as variables, Lindsay and Gordon used words and phrases from full MEDLINE Abstracts as variables. For reasons unexplained, they used the MEDLINE database from 1983-1985, rather than from the 1975-1985 time frame Swanson used.

They used phrase, word, term, and record counts as their key metrics, both absolute frequencies and relative frequencies. After they retrieved the RP core literature, they used metrics to identify the intermediate 'B' literatures. They essentially accomplished this through a phrase frequency analysis of the core RP literature, and examination of the high frequency high medical content phrases. Use of these phrases can be viewed as a less formal type of clustering, as opposed to the more formal clustering approach used by Kostoff et al [2008b] to define the intermediate literatures.

Until this point, the analysis appears objective, with knowledge of the known results not influencing the analysis. After this point, it appears to us that the analysis is influenced by knowledge of the final results. Gordon and Lindsay grouped blood-related terms into essentially a blood-related cluster, although other clusters would have been possible based on the terms generated and their frequencies. They then did a frequency analysis of the blood-related records in the RP core literature, and argued that blood viscosity is an important concept that merits further analysis. However, while blood viscosity ranked high on the frequency lists, it was not first, and many other concepts could have been selected with equal logic. There is no evidence that blood viscosity would have been selected for further examination (at least as the first choice) had the end result not been known a priori.

They then downloaded the blood viscosity literature, and examined high frequency terms as they did twice before. They identified those terms that they viewed as potential solutions (terminals), and found that two of those 'terminals' were Fish Oil and Eicosapentaenoic Acid (Swanson's findings). They then concluded: "We have thus replicated Swanson's finding".

However, 'Swanson's finding' was that the highest recommended potential treatment discovery for RP was Fish Oil/ Eicosapentaenoic Acid. Gordon and Lindsay found that one of many potential treatment discoveries for RP was Fish Oil/ Eicosapentaenoic Acid. That is a different finding from Swanson.

We have no problem with their selecting high frequency phrases for proxy clustering to identify intermediate literatures. However, we question the use of identifying high frequency phrases as potential discoveries (as was done when finally selecting Fish Oil/ Eicosapentaenoic Acid) for the reasons presented in our discussion of Swanson's approach. There is no reason a priori that a high frequency phrase should be related to discovery, and much intuitive reason to believe its chances of being a potential discovery item are reduced. To summarize, we don't believe that Gordon and Lindsay replicated Swanson's result in all its dimensions, and we don't believe their method has the potential for generating the full scope of potential discovery possible with a literature-based approach.

D. KOSTOFF ET AL - 2008

A recent ODS LBD group of five studies [Kostoff et al, 2008a-2008h] followed the classic Swanson approach. In the first study [Kostoff et al, 2008c], the MEDLINE database used by Swanson for the original RP study, 1975-1985, was re-evaluated. In the second [Kostoff, 2008d], third [Kostoff et al, 2008e], and fourth [Kostoff et al, 2008f] studies, potential treatments for cataracts, Parkinson's Disease (PD), and Multiple Sclerosis (MS), respectively, were researched. In the fifth study [Kostoff et al, 2008g], potentially low cost alternatives to present water purification (WP) approaches were identified using the SCI.

In all five cases, the 'A', 'B', 'C', literatures described previously were used as an initial model. The starting problem literature, 'C' was initially retrieved. Then, the retrieved 'C' literature was ***clustered*** to identify systematically the main technical/ medical thrusts that characterized the

problem literature. The subset of clusters deemed the most important by experts constituted the 'B' literatures.

One difference from the other ODS LBD approaches was the method of 'A' topic identification. No filtering assumptions were made related to frequency. Conditions were not required that 1) 'A' term frequencies had to be much higher in 'B' literatures than in MEDLINE background, or 2) the number of different 'B' literatures in which 'A' terms appeared should impact ranking. The only filtering assumptions made were specification of the semantic category of the 'A' terms and use of term combinations in queries to retrieve related literatures. Depending on the specific study, visual inspections were performed of the phrases or records that had been selected from the desired semantic categories. In most of the studies, the 'A' literature was divided into two components, one related more directly to the original core literature than the other.

Citation relationships in the SCI were used to link related literatures. This approach complements the Swanson approach, where the literatures are linked by terminology. In the citation linking mode, after identification of promising 'A' terms and their associated records was made, searches were made for: a) records that cite or are cited by the promising 'A' record; b) other records that are cited by the records that cite the promising 'A' record; and c) other records that cite the records cited by the promising 'A' record. These citation relationship searches were initiated only at the end of the studies, but the results look highly promising. Citation linking allows similar concepts to be accessed without the need for similar terminology, and offers the promise that very disparate literatures can be accessed for potentially radical discovery.

When what appeared to be a potential discovery was surfaced, it was labeled as a 'potential discovery ***candidate***'. Those potential discovery candidates were 'vetted' before they were claimed to be 'potential discoveries' (in ODS LBD, the term 'hypotheses' is used, since lab tests and field/ clinical trials are required to verify that the hypotheses are true discoveries). To 'vet' a potential discovery candidate, at least the following four steps were required: 1) check the core 'problem' literature for co-occurrence of the potential 'discovery' and the 'problem'; 2) check the potential 'discovery' citing papers for mention of the 'problem'; 3) check the patent literature for co-occurrence of the potential 'discovery' and the 'problem'; 4) ask an expert(s) in the core 'problem' literature whether the claimed 'discovery' is an actual

discovery. If a potential 'discovery' candidate passed these four 'wickets', the potential 'discovery' candidate could be published as a potential discovery.

1. Raynaud's Phenomenon – MEDLINE Data (1975-1985; vetted 1965-1985)

For the RP study, approximately 130 potential discoveries were identified [Kostoff et al, 2007b], substantially more potential discovery than all the other reported studies on the RP problem combined. Some of these potential discoveries include:

Fish oil for inhibiting platelet aggregation and reducing blood viscosity; Agar-Agar for peripheral arterial vasodilation; Enkephalins for vasculature casodilation in skeletal muscle; Nitric Oxide for smooth muscle relaxation/ vasodilation; Benzoic Acid for arterial vasodilation, platelet aggregation inhibition, blood viscosity reduction; Reflexotherapy for reduction in total peripheral vascular resistance; Huang Chin extract for peripheral vasodilation; Secretin for vasodilation; Vernolepin/dried fruit for platelet disaggregation; Guar Gum for decrease in plasma fibrinogen and viscosity; Cell hydration to improve cell deformity and increase arm blood flow.

This additional discovery was due mainly to removal of the numerical filters that other researchers added (in order to reduce the number of potential discovery candidates required to examine). The cost of this added discovery for the RP study was visual examination of a large number of phrases for candidate discovery (~270,000 phrases for the RP study, where the semantic filtering was done at the phrase visual inspection), and the subsequent requirement to examine the many articles retrieved that contain these candidate discovery phrases. As will be shown in the (next) cataracts study section, visual inspection could be replaced by semantic category specification (i.e., specifying MeSH categories in MEDLINE), reducing the visual inspection time considerably.

Only the tip of the discovery iceberg was uncovered, due to shortcomings in the RP study. Much more potential discovery for the RP problem should be possible if the following improvements to the study procedure were to be implemented:

- Expanding the core literature comprehensively by using larger queries

- Increasing the number of possible expanded literatures
- Increasing medically-oriented personnel in all phases of the study. Mainly non-biomedical personnel were involved in the discovery identification process, and were not equipped to identify the more subtle relationships
- Examining single frequency phrases. The ~900,000 phrases with a frequency of unity were not examined systematically for potential discovery (only phrases with a frequency of two or greater were examined during the initial visual inspection), although single frequency events were identified during the citation linking process at the end of the study. From the rare event perspective discussed previously, these rare phrases may have the most potential for ***real*** discovery! Coupled with a more complete expanded literature, the unity frequency phrases could total well over a million, and serve as a potential 'gold mine' for real discovery
- Increasing use of citation linkages. The citation-based discovery pathways had only been used for a minute number of cases, and these pathways appeared to offer enormous potential for discovery
- Improving information content of records (not within purview of analysts). Many records contained insufficient information to make a determination of discovery. Almost all records prior to 1975 did not contain Abstracts, and a significant number of records in the decade after 1975 also did not contain Abstracts. Because of the large volumes of papers without Abstracts, it was not feasible to track down the full texts of these papers. Unless a positive determination of potential discovery could be made from the information contained in the MEDLINE record, an item was not included in the discovery list

Is this manually intensive approach to discovery cost-effective? At present, there appears to be no alternative. As this review shows, ***most ODS LBD approaches have not demonstrated discovery.*** The cost of a manually intensive study is miniscule compared to the magnitude of potential benefits, if in fact discovery results.

2. Cataracts

A short study on potential discovery for cataract treatments [Kostoff et al, 2008d] was performed using lessons learned in the RP study. 'Cataract*' was used to define the 'C' literature, and semantic categories were selected

for the ‘A’ literature. In the spirit of Swanson’s original RP paper, only non-drug approaches were examined. In particular, as a first approximation, “Plants, Medicinal” or “Plants, Edible” or "Plant Extracts" or "Plant Oils" or Phytotherapy or Fruit or "Fish Oils" or Flavonoids or Dietary Supplements were selected as the ‘A’ literature semantic categories.

The ‘C’ cataract literature was then clustered using the CLUTO document clustering software [CLUTO, 2006] to identify the main medical thrusts that characterized cataracts, and a subset of the main medical thrusts (the ‘B’ terms) believed to be most promising was selected. Terms in this subset were used as a MEDLINE query, in addition to the ‘A’ term semantic categories listed above, along with exclusion of the ‘C’ cataracts core literature. The retrieved records were examined for potential ‘discovery’ from this ‘AB’ literature.

To generate an indirectly related literature, ‘B’ terms from the directly related literature query above (with the word ‘cataract*’ as a negation term) were used as a MEDLINE query, and all the appropriate MEDLINE records with no semantic category restrictions were retrieved. This retrieval was then clustered, and the main medical thrusts were identified.

In parallel, it was observed that the records from the semantically restricted directly related literature that were potential discovery candidates tended to have two or more of these main medical thrusts as MeSH terms, whereas the non-potential discovery candidates tended to have one, or even none, of these major medical thrusts listed as MeSH terms. The following steps were then performed: 1) a sub-set of these cluster categories that overlapped with the MeSH terms in the directly related literature promising discovery candidates was selected; 2) all intra- and inter-medical thrust combinations of three MeSH terms (although the latest medical study performed by the authors [Kostoff et al, 2008f] showed that only the intra-thrust terms should be combined in a combinatorial manner) were identified and intersected with the ‘A’ term semantic categories; 3) the core and directly related literature queries were subtracted; and 4) the resulting query was used to retrieve these indirectly-related records. Thus, for example, if there were a total of ten important MeSH terms identified from the directly related literature potential discovery candidates examined, the number of combinations of any three MeSH terms would be specified by the binomial coefficient (10!/(7!*3!)=120, in this case). The citation approach described in the

previous RP section was still used for a few cases to access the indirectly-related literatures.

Examples of potential discoveries from literatures related directly and indirectly to the cataracts core literature include:

Isogentisin for activating cellular repair functions through oxidative stress and protein oxidation reduction; cultured Cordyceps militaris and natural Cordyceps sinensis for protection against oxidative damage of biomolecules are a result of their free radical scavenging abilities; three active components of the root of Scutellaria baicalensis Georgi, i.e. baicalin, baicalein and wogonin, for inhibiting oxidation and nitration; Walnut (Juglans regia L.) bark for antioxidant potential; Acorus calamus to reduce oxidative stress; mixtures of phytochemicals to minimize oxidative stress; Ziziphora clinopoides to reduce lipid peroxidation and cellular oxidative stress; Gypenosides for multiple antioxidative actions and for reducing glutamate and oxidative stress; guava leaf extracts for antioxidant activity and free radical scavenging; Biophytum sensitivum (L.) DC (Oxalidaceae) for antioxidant activity; Eruca sativa seeds as a powerful antioxidant and protector against oxidative damage.

The large number of potential discoveries reported here that relate to antioxidant activity reflect the dominant mechanism ('B' literature) of oxidation/degradation of the protein in the lens, and the need for antioxidants to counter the lens oxidation if progress is to be made in treating cataracts non-surgically.

Also as in the RP study, only the tip of the iceberg with respect to quantity or quality of potential discovery was visible. Even though the streamlined study had a number of shortcomings (not all semantic classes used, relatively abbreviated query with limited biomedical phenomena used, limited numbers of MeSH terms used in query, citation approach barely used), hundreds of potential discovery items were generated.

3. Parkinson's Disease (PD)

A short study on potential discovery for PD treatments [Kostoff et al, 2008e] was performed using lessons learned in the cataracts and RP studies. The query 'Parkinson* NOT (Parkinson*[AU] OR Wolff-Parkinson*)' was used to define the 'C' literature, and semantic categories similar to those for the

cataracts study were selected as the 'A' literature. An analysis process similar to that for the cataracts study was used, including the use of combination intra- and inter-medical thrust biomedical phenomena terms in the query to target records that contain similar characteristics to potential discovery items already identified (again, only intra-thrust combinatorials are recommended for future studies). The citation approach described in the previous cataracts and RP sections was still used for a few cases to access the indirectly-related literatures.

While only a fraction of the retrieved records have been examined, substantial potential discovery has been identified, as in the cataracts and RP studies above. Examples from literatures related directly and indirectly to the PD core literature include:

Cordyceps militaris and Cordyceps sinensis for protection against oxidative damage of biomolecules are a result of their free radical scavenging abilities; malanga carotenoids extract, and malanga leaf powder for protection against oxidative damage; kolaviron for protection against oxidative damage to molecular targets via scavenging of free radicals and iron binding; dried fruit rind of the plant Garcinia cambogia to attenuate increases in oxidative stress; Isohumulones derived from hops provide an anti-oxidative effect; antioxidative properties of brown algae polyphenolics; neuroprotective and antioxidant properties of aqueous extracts from Halimeda incrassata (Hi) and Bryothamniom triquetrum (Bt); cytoprotective properties of Magnolia officinalis and Euphorbia pekinensis against oxidative stresses; G. [Gynandropsis] gynandra extract to diminish the rate of lipid peroxidation, with a significant increase in the levels of enzymatic and non-enzymatic antioxidants; Jianpi Liqi Huoxue Decoction for anti-lipid peroxidative effect; protective effect of Lycium barbarum polysaccharides on oxidative stress in rats; M. [Marrubium] vulgare for natural antioxidants, which inhibit LDL oxidation; Alchornea cordifolia for protection against oxidative stresses; flavonoids from the flower of Rhododendron yedoense var. poukhanense for antioxidant activities; aqueous extract of G. [Gongronema] latifolium leaves for anti-lipid peroxidative properties.

4. Multiple Sclerosis (MS)

A short study on potential discovery for MS treatments [Kostoff et al, 2008f] was performed using lessons learned in the cataracts, RP, and PD studies. The query 'Multiple Sclerosis' was used to define the 'C'

literature, and semantic categories similar to those for the PD study were selected as the 'A' literature. An analysis process similar to that for the PD study was used, including the use of combination biomedical phenomena terms in the query to target records that contain similar characteristics to discovery items already found. However, unlike the PD study, the combinations of terms were limited to **intra-medical thrusts only** (where the medical thrusts were determined through clustering), based on further observations of textual patterns in potential discovery records. The citation approach described in the previous cataracts, RP, and PD sections was still used for a few cases to access the indirectly-related literatures.

Also, the quasi-ODS mode of Arrowsmith (described previously) was used to generate some potential discovery. "Multiple Sclerosis" was used as the 'C' literature, and the non-drug semantic classes were used as the 'A' literature. The key to feasibility when using Arrowsmith, however, was the use of ***clustering-generated important mechanism phrases for identifying which of the tens of thousands of 'B' literature terms to examine for potential discovery*** [Kostoff et al, 2008f, 2007b]. This clustering-based method for defining important B literature terms can be used as a supplement to, or alternative for, the logistics regression approach of Torvik and Smalheiser [2007].

Potential discovery examples from literatures related directly and indirectly to the MS core literature include:

Petaslignolide A is suggested to be a major neuroprotective agent primarily responsible for the protective action of the butanol fraction of P. japonicus extract against kainic acid-induced neurotoxicity in the brains of mice; Salvia miltiorrhiza Bunge (a Chinese herbal medicine) attenuates increased endothelial permeability induced by TNF-alpha. Inchinko TJ-135 (a Japanese herbal medicine) inhibits inflammatory cytokines and enhances production of anti-inflammatory cytokines; Kalpaamruthaa (KA), a modified indigenous Siddha preparation constituting Semecarpus anacardium nut milk extract (SA), Emblica officinalis (EO) and honey showed an enhanced antioxidant potential n the management of RA; tiliroside and gnaphaliin, two flavonoids isolated from Helichrysum italicum, are antioxidants against in vitro Cu(2+)-induced LDL oxidation in the same order of magnitude compared to that of the reference drug, probucol, Cissus quadrangularis extract (CQE) exerted inhibitory action on generation of lipid peroxidation, proinflammatory cytokines and neutrophil infiltration.

5. Water Purification Clustering

For this non-medical ODS LBD study, the SCI was the database source searched for ODS LBD improvements to present water purification processes [Kostoff et al, 2008g]. Fundamentally, the Swanson-based A – B – C process was used. One unique feature of this approach was the role played by clustering. The 'B' literatures were clustered, and clusters were used to filter the 'C' concepts to be examined. The 'B' clusters were inspected visually, and they tended to group by interesting or uninteresting with regard to potential discovery promise. This allowed the uninteresting clusters to be discarded. After interesting concepts were identified, in some cases the citation linking approach was used to identify other promising concepts. The 'vetting' process (described previously) was used for some of the potential discoveries to insure that they were indeed unique.

Some potential discoveries include: The use of plasmin to deter cell adhesion for use in non-fouling coatings for membranes; the use of plant roots to filter water, and to create new membranes based on plant physiology that can vary their permeability; the prevention of fouling on membranes based on surface bio-magnetism repelling of bacteria; the use of sterile surface materials (one end of a long-chained hydrophobic poly-cation containing antimicrobial monomers is attached covalently to the surface of a material) to prevent fouling of membranes; the use of fish/shrimp gill cleaning mechanisms as an membrane antifouling strategy; the use of triterpene glycosides (derived from Erylus formosus and Ectyoplasia ferox, two Caribbean sponges) to deter or prevent fouling of membranes . Many more are listed in Kostoff et al [2007b]; Solka et al., [2007].

For the reasons presented above, only the tip of the iceberg of potential discovery possible was observed. Additionally, it remains to be demonstrated how effectively the authors are able to identify single frequency events using a clustering approach.

Finally, an ODS LAD approach for the WP problem [Kostoff et al, 2008g, 2007b] was developed and tested. Experts from technical disciplines disparate from, but related to, WP were notified about the posting of a Broad Agency Announcement concerning WP, in the hope that these experts would submit proposals using principles and approaches from their disciplines to

solve problems related to WP. About 2/3 of the pre-proposals received from these experts proposed techniques not normally associated with WP.

II. THESAURUS /MeSH TERMS

A. WEEBER ET AL – 2001

There are two features that distinguish this LBD approach [Weeber et al, 2001] from its predecessors. First, it uses a two step model of discovery. The first step is an ODS process based on hypothesis generation, and the second step is a CDS process based on hypothesis testing. Second, it maps the text into the Unified Medical Language System (UMLS) as a thesaurus, to standardize vocabulary and reduce dimensions substantially. Additionally, given the semantic categories contained in the UMLS, filtering by category can be accomplished.

The authors attempted to "simulate the actual discovery" by which Swanson linked Fish Oil/ Eicosapentaenoic Acid to RP. But, rather than identify an intermediate literature with no prior knowledge, they started with Swanson's finding of "platelet aggregation, blood viscosity, and vascular reactivity". In doing so, they bypassed a key challenge in the discovery process!

For each of the three literatures above, they identified potential discovery candidates, using a ranking scheme based on frequency. In fact, their approach emphasized high frequency concepts that "will likely draw the expert's attention". The Fish Oil variants did not immediately stand out in the rankings, but the authors concluded: "We think that an expert user will mark them as "interesting", and therefore a hypothesis has been generated successfully".

Having "found" the Fish Oil 'discovery' by the ODS process, the authors then used a CDS process (starting with RP at one end and Fish Oil at the other end, identify the different mechanisms by which Fish Oil can impact RP) to show that a number of intermediate paths are possible. Thus, the overall process is based on high frequencies: high frequency concepts to rank the potential discovery concepts from the ODS process, then high frequencies of pathways linking the highly ranked high frequency potential discovery concept to the problem (treating RP, in the above case).

There are a number of problems with this approach. First, it requires a detailed thesaurus for feasibility. While such a thesaurus exists for the medical field, a similar resource is not available for many other fields. Second, it is high frequency-based. We have discussed problems with this requirement for identifying potential discovery previously. Third, the specific approach presented did not identify the intermediate literatures a priori. Finally, because of where Fish Oil appeared in the rankings, we question whether it would have been noticed easily as a promising discovery concept, as the authors claim. Their combining an ODS approach with a CDS approach could provide useful insights about linking mechanisms. The authors also studied potential applications of the drug Thalidomide using their ODS LBD approach [Weeber et al, 2003].

B. STEGMANN AND GROHMANN – CO-WORD – 2003

This technique [Stegmann and Grohmann, 2003] is based on co-word clustering of MeSH terms [Callon et al, 1991; Kostoff, 1993]. For the RP problem, the RP records were downloaded, and mainly the MeSH terms were extracted. Terms occurring once (one document only) were omitted from the analysis.

Then, the high Equivalence Index term pairs were clustered. Cluster properties such as density and centrality were computed, and derived properties were computed as well. Maps of density and centrality were generated.

Examination of the resulting density-centrality map shows interesting terms to be concentrated in the lower left quadrant (i.e., exhibit below-median centrality and density values). This is true for the RP intermediate literature as well as for the Fish Oil literature. However, there are interesting terms in other quadrants as well, and not all terms in the lower left quadrant are interesting. Thus, the merit of this characterization approach is to identify a starting point for exploring discovery rather than a hard roadmap. Intuitively, the lower left quadrant is in line with our previous statements about relatively rare events being more favorable for discovery than high density central events, although unitary events have been excluded.

C. SRINIVASAN – 2004

A 2004 JASIST LBD article [Srinivasan, 2004] described a semi-automated software approach for hypothesizing potential discovery. The article presented ODS and CDS results for medical topics analyzed by previous investigators. In follow-on articles that year using the same software package [Srinivasan and Libbus, 2004; Srinivasan et al, 2004], potential 'discoveries' were hypothesized for three medical problems not examined previously for ODS LBD purposes: retinal problems, lower bowel problems, and EAE/Multiple Sclerosis treated by curcumin. This section will demonstrate that the three specific claimed 'discoveries' are neither discoveries nor innovations, where a discovery (in the ODS LBD context) represents the linking of two or more concepts that had never been linked previously in order to produce novel, interesting, plausible, and intelligible knowledge. We will use our vetting approach to show the presence of prior art. We present some detail in this section, to illustrate the operational mechanics of our vetting process.

The discovery approach used in the 2004 JASIST article [Srinivasan, 2004] is based upon Swanson's initial ODS LBD concept described previously [Swanson, 1986], where 'C' is the source (problem) literature, 'B' is the intermediate literature, and 'A' is the solution literature.

In the 2004 JASIST article, the author generated a potential discovery-identifying algorithm that operated by building MESH-based profiles from MEDLINE for topics. A profile is a weighted vector of MeSH terms that together represent the corresponding topic. Additionally, MeSH terms were separated by semantic type (the MeSH vocabulary has already been classified using 134 Unified Medical Language System [UMLS] semantic types), and weights for each MeSH term were computed within the context of a semantic type. The authors used term frequency-inverse document frequency (TF-IDF) weighting and then normalized the weights.

The ODS LBD algorithm operated as follows. First, a MeSH profile was built for the initiating 'C' topic of interest (in the medical context, 'C' could be a disease for which a treatment is desired, or a substance for which potential target diseases are desired). MeSH terms in the profile had TF-IDF weights that were normalized within each semantic type. A select number of 'B' MeSH terms ('B' represents the intermediate literature that links the initiating literature 'C' with the target literature 'A') were automatically selected from the user specified semantic type (ST) 'B' vector components and their profiles were built. These were then merged to form a final profile.

The combined weight of a term was the sum of its weights in the individual 'B' profiles. In the last step, the 'A' MeSH terms were limited to those representing novel connections. An 'A' MeSH term's score was regarded as the system derived estimate of the potential value in its association with the 'C' topic. This score depended both on the number of paths connecting back to 'C' as well as the strengths of these paths. The higher the score, the stronger the recommendation made by the system. The 'A' MeSH terms within each semantic type were ranked by combined weight.

In [Srinivasan and Libbus, 2004; Srinivasan et al, 2004], the authors started with curcumin and looked for ailments this substance could benefit potentially. Three areas identified were "retinal pathologies including diabetic retinopathies, ocular inflammation and glaucoma", Crohn's Disease / Ulcerative Colitis (both members of Irritable Bowel Syndrome), and EAE/Multiple Sclerosis (MS).

We will examine the three specific claimed 'potential discoveries' listed above. Since the papers were published in 2004, and the data were taken in mid-November 2003, then potential discovery would require that no papers/patents linking curcumin and these three ailments had been published prior to November 2003. Our approach is to examine the core literature (papers/ patents) for these three ailments published before November 2003, and ascertain whether they include curcumin as a potential treatment. If they do, then potential discovery by the authors cannot be validated.

To examine the core literature, we used text terms based on the main MeSH terms used by the authors, and initially entered them (initiating topic 'C' literature AND target 'A' literature terms) into the PubMed search engine. This allowed us to retrieve MEDLINE articles that contained the initiating topic and target literature MeSH terms and/or text terms. Then, to obtain citing or reference article data, we entered the same terms into the SCI. Finally, to obtain patent data, we entered the same terms into the Derwent Innovations Index, an aggregated global patent database on the Web of Knowledge.

Using mainly MeSH terms as text terms is a very conservative approach. If we were searching for prior art to support a legal case, we would use many other proxy terms for the initiating topic and target literatures as part of our search query. Given the breadth of coverage of the average MeSH term relative to that of the average text term, many more proxy terms could be

subsumed under the average MeSH term than under the average text term. In some sense, ***the generality of MeSH terms relative to text terms opens the door wide for refutation of potential discovery by allowing for the implementation of large numbers of proxy terms in the vetting process.***

Only a few of these examples will be shown, due to space considerations.

For the MS example, Natarajan and Bright [2002] published a paper in June 2002 linking curcumin to the treatment of MS. That paper had numerous citations, five of which were published in the first half of 2003.

For the Crohn's Disease example, Sugimoto et al [2002a, 2002b] published a meeting Abstract in *Gastroentorology* in April 2002 and a research article in *Gastroentorology* in December 2002 concluding "This finding suggests that ***curcumin*** could be a potential therapeutic agent for the treatment of patients with ***inflammatory bowel disease***." The keywords in the research article record include Crohn's Disease and Ulcerative Colitis, and Colitis is in the title as well. See also Sahl et al [2003], Ukil et al [2003].

For the retinal pathologies example (where glaucoma focuses mainly on intraocular pressure and optic nerve damage), three examples are required due to topical diversity. For the diabetic retinopathy example, a 2002 paper [Okamoto et al, 2002] suggests cervistatin, pyrrolidinedithiocarbamate, or curcumin could equally serve as a treatment for proliferative diabetic retinopathy. Additionally, one of its citing papers [Balasubramanyam et al, 2003] focused exclusively on the proposed curcumin treatment for diabetic retinopathy, confirming the importance of curcumin in the cited paper. Further, a patent whose application was published in October 2002 and which was granted in May 2003 suggested a link between curcumin and both retinopathy and Crohn's Disease/Ulcerative Colitis [Babish et al, 2003].

For the ocular inflammation example, a 2001 paper described the use of commercially available herbal eye drops containing curcumin for a "variety of infective, inflammatory and degenerative ophthalmic disorders" [Biswas et al, 2001]. This formulation has existed since at least the 1990s, and almost ten clinical/laboratory papers of which we are aware have been published on its evaluation between 1998 and 2002. Finally, the patent by Babish above links curcumin to conjunctivitis and uveitis (an inflammation of part or all of the uvea, the middle (vascular) tunic of the eye and commonly involving the other tunics (the sclera and cornea and the retina)).

For the glaucoma example, a patent with 2001 application date and 2003 granting date linked curcumin directly with glaucoma [Komatsu, 2003].

These results should not be surprising. There are over 2300 papers in MEDLINE (as of mid-2007) related to curcumin (or curcuma or curcuminoid), of which over 20% directly address its role as an anti-inflammatory agent. Any disease in which inflammation plays a role and which is presently not co-mentioned with curcumin would be a candidate for potential 'discovery'. Many of Srinivasan's proposed discoveries relate to inflammation-based diseases. Unfortunately, as stated previously, with so many researchers working on the relation of curcumin to inflammation, the chances that the link between curcumin and a major inflammation-based disease would go unnoticed are probably small, as our vetting results seem to be showing.

Why did these prior 'discoveries' escape detection by the algorithm developed in Srinivasan [2004]? The algorithm works in MeSH-term space, whereas our vetting searches occurred in text space as well as MeSH space. There is not a one-to-one mapping of text terms in the Abstract/title/SCI keywords to MeSH terms. Typically, our search terms appeared in the Abstract, but not necessarily in the MESH terms. This meant that the MeSH terms used for the indexing were not as complete as possible. There is a well-known effect in information retrieval called the Indexer Effect (e.g., [Healey et al, 1986]), whereby errors in classification and/ or omission are made by third-party indexers. Additionally, there is a latency period in MEDLINE before new articles are indexed in MeSH. This is one of the dangers of relying on MeSH terms exclusively, as the authors have done, and requires extreme levels of checking/vetting if the results are to be credible. Finally, unlike Srinivasan et al, we included the SCI and patent databases in our core literature definition.

What we have presented above is probably the tip of the iceberg. There are obviously other ways to refer to curcumin or Crohn's or retinopathy or colitis, and a search using these additional proxy terms would enhance the amount of prior art. In sum, we would not call these curcumin links a potential discovery, because the links between curcumin and retinal, intestinal, or MS problems were established well before November 2003. The algorithm under discussion, with perhaps some modifications, might be

a solution for some types of semi-automated ODS LBD, but it was not demonstrated by the three examples shown.

D. YILDIZ AND PRATT - 2006

The authors present an ODS LBD system called LitLinker that incorporates knowledge-based methodologies with a statistical method [Yildiz and Pratt, 2006]. The ODS LBD begins with a starting term (the 'C' literature), then uses a text mining process to find a set of terms (linking terms – the 'B' literature) that are directly correlated with the starting term, and uses the same text mining process to identify a set of terms that are correlated with each linking term. Finally, Litlinker ranks the target terms by the number of linking terms that connect the target term to the starting term.

In searching the database, Litlinker uses MeSH terms as the representation of the content of the documents and performs searches on them to collect the literatures. To find correlations, Litlinker calculated the probability of a term appearing in a literature by dividing the number of documents of the literature in which the term appeared by the total number of documents in the literature. Those terms with differences between the probability of a MeSH term in a specific literature and the general distribution of this MeSH term in the background set of literatures larger than a pre-defined threshold are marked as the correlated terms to the starting or linking term.

All the techniques surveyed for this review have the common requirement for strong dimensional reduction (i.e., less words, obtained by eliminating terms that are too general, redundant, and in semantic classes incompatible with discovery), and Litlinker is no different. Litlinker used the MeSH hierarchy available in UMLS to eliminate broad terms, by eliminating terms higher up in the hierarchy than the term of interest. To eliminate the redundant terms, Litlinker again used the MeSH hierarchy to prune all ancestors, siblings, and children of the starting term from the list of potential linking terms. To eliminate implausible and uninteresting terms, Litlinker retained only those semantic groups in UMLS that were plausible for terms that could be correlated with a disease or a medical condition and a potential treatment.

To rank target terms from linking terms, Litlinker first eliminated any target terms that co-occurred with a starting term, then ranked the remaining target terms according to the number of linking terms that connected each target

term to the original starting term, and removed the target terms with number of linking terms below a pre-determined threshold.

The authors used a different approach to evaluate all correlations (potential discoveries) that Litlinker generated. They divided MEDLINE into two parts: a baseline literature including only publications before January 1, 2004, and a test literature including only publications between January 1, 2004 and September 30, 2005. They ran Litlinker on the baseline literature and checked the generated connections in the test literature.

They reported results for three cases: Alzheimer's Disease, Migraine, and Schizophrenia. They used precision and recall as key evaluation metrics. Precision for starting term i is defined as the ratio of (Ti union Gi)/Ti, and recall for starting term i is defined as the ratio of (Ti union Gi)/Gi, where Ti is the set of target terms generated by Litlinker for the starting term i, and Gi is the set of terms in the gold standard created from the test literature of starting term i. The gold standard was defined as the MeSH terms in the test literature not in the baseline literature and filtered for appropriate semantic group.

Precision-time graphs showed precision increasing with time to about .06 for Alzheimer's Disease, .025 for Migraine, and .075 for Schizophrenia. Recall-time graphs showed recall oscillating with time, with approximate mean values of about .22 for Alzheimer's Disease, .43 for Migraine, and .14 for Schizophrenia. These appeared on the surface to be quite reasonable results, if in fact the target terms in the gold standard were potential discoveries. The authors provided one specific example of potential discovery for each of the three diseases examined, and these three specific examples were the only potential discoveries with sufficient detail to be checked independently.

Before commenting on some of the problems we have with the Litlinker approach, we wanted to verify some of these claimed 'discoveries'. We examined the specific example of potential discovery for each of the three databases listed. As shown in Kostoff [2007a] (a published critique of the LitLinker approach and claimed discoveries), for each of these three potential 'discoveries' there was prior art published in the mainstream journal and patent literatures before 2004 that linked each potential 'discovery' with the relevant target literature.

In response to Kostoff [2007a], Pratt and Yildiz [2007] have questioned our use of the third vetting step (patent literature examination) as being overly harsh. Most of the potential 'discoveries' reported by previous ODS LBD researchers (such as Srinivasan) were Medline-based (since all the previous topics examined were medical), with perhaps some potential 'discoveries' coming from SCI supplementation. Their search techniques did not access the patent literature; why, then, should the patent literature be used in the vetting process?

We take the definition of discovery literally. We work on the assumption that every potential discovery we report in the literature could be patentable (or its equivalent in uniqueness), if we so choose. To be patentable, a potential discovery derived from ODS LBD has to meet three main conditions: no prior art; value added by the linkage; sufficiently important to justify the resource expenditures required to patent. These conditions are not a function of the databases we selected. From our perspective, claiming discovery based on no prior art from Medline search only, or SCI search only, or patents search only, or search of any other restricted database, is of limited practical consequence. If Yildiz and Pratt, and other ODS LBD researchers who claimed potential discovery, had presented their results as "no prior art based on Medline search only", we would have much less of a problem, since, in those cases in which this 'no prior art' condition held, that is a factual statement. However, ***when there is an equivalence drawn between 'no prior art in Medline' and potential discovery, we have problems with such claims***.

In vetting the results of our own ODS LBD studies [Kostoff et al, 2008c-g], we were much harsher than in the vetting we report in this report. Most of the findings of prior art (during the vetting of our own studies) occurred in the patent literature vetting step. Had we performed steps 1 and 2 only (Medline and the SCI), we would have had substantially more potential 'discoveries'. However, we believe that presenting such results as potential 'discoveries' to independent third parties would have impacted the credibility of our findings, and would have cast doubt on the credibility of our whole approach. Therefore, ***we defined discovery in the sense in which it is understood by most of the technical community***, and designed our vetting process to support that goal.

Obviously, there are always more sources that can be checked for vetting, and more types of citation linkages by which articles may be related, but we

view the above as a threshold of necessary checking before being willing to claim potential discovery. We use the first three of these 'wickets' as evaluation criteria for checking the potential discoveries claimed in this report by the authors of the articles evaluated.

As with the other ODS LBD techniques surveyed in this review so far, there do not appear to be features in LitLinker that would automatically target potential discovery. In addition, LitLinker requires use of a MeSH term-like taxonomy and a UMLS-type thesaurus. Most disciplines don't have such comprehensive supporting systems, so for all practical purposes, LitLinker is limited to the medical field. Use of term probabilities and accompanying higher occurrence frequencies for identifying correlated terms may be acceptable for identifying linking terms ('B' literatures). However, use of these term probabilities and higher frequencies for identifying target terms is similar to Swanson's first filtering condition, and we have the same problems here with LitLinker as we did with Swanson. LitLinker's use of number of links connecting the target term with the starting term in order to rank the target terms is similar to Swanson's second filtering condition, and we have the same problems here with LitLinker as we did with Swanson.

In addition, Swanson requires the intermediate literatures to be orthonormal sets for the purpose of counting pathways from the 'A' literature to the 'C' literature (to avoid double counting). There is no formal orthogonality requirement for linking literatures in LitLinker (although the elimination of redundant terms goes in that direction), and one could in theory be doing multiple counting in overlapping literatures when summing up pathways in LitLinker. For example, two of the nine linking literatures identified as part of the endocannibinoids 'discovery' are neurotransmitters and neuroprotective agents. These literatures overlap in 1493 MEDLINE articles published between 1980 and 2004.

Finally, there is one curious feature of the evaluation examples. The authors selected three diseases from Swanson's study that were analyzed by a CDS approach, and they used Litlinker to test them by an ODS approach. Why didn't they use the one example that Swanson evaluated by an ODS approach (RP), using Swanson's 1975-1985 database? They would have had twenty years of post-1985 data, for long-term statistics, and would have used the benchmark database that essentially all the other ODS LBD researchers use for comparison.

LitLinker, with perhaps some modifications, might be a solution for semi-automating literature-based discovery, but it was not demonstrated by the three examples from [Yildiz and Pratt, 2006].

E. VAN DER EIJK - 2004

This approach [van der Eijk et al, 2004] is based on mapping from a co-occurrence graph to an Associative Concept Space (ACS), where concepts are assigned a position in space such that the stronger the relationship between concepts, the closer they lie in the ACS. Potential discovery can then be obtained from strong implicit relationships, where concepts are close to each other in ACS but have no direct connections.

The text words are first transformed to concepts through use of a thesaurus, in this case the MeSH terms of MEDLINE. A list of a document's concepts is called a concept fingerprint of that document. For each identified concept, a unique concept identifier is added to the fingerprint. The concept identifier is assigned a relevance score, based on term frequency and the specificity of the term in the thesaurus. The fingerprints form compact representations of documents, because they are lists of concept identifiers.

Co-occurrence of concepts in fingerprints is a central metric. Concepts are mapped into an ACS. Concepts that are connected by frequent co-occurrence paths (either directly or indirectly) should have a small distance in the ACS, while concepts with few or no paths between them should be far apart. A Hebbian learning algorithm is used to determine an appropriate position for the concepts. As a result, the Euclidean distance between two concepts is a measure of both co-occurrence and how many co-occurring concepts the two concepts have in common. Thus, co-occurrence quantity is a driving metric of relative positioning in ACS.

The authors provide two examples [van der Eijk et al, 2004] of ACS for small sub-sets of the total MEDLINE database (<<1%), whereby concepts that were close together in ACS but not connected were predicted to have a strong implicit relationship. Searching for co-occurrence of these concepts in total MEDLINE showed a significant number of co-occurrences.

In the first example, the authors retrieve a subset of MEDLINE records (13423 records, February 9, 2003) from PubMed with the MeSH-based query "(duchenne OR DMD OR dystrophy OR limb-girdle OR LGMD OR

BMD)". According to the ACS diagram, and the authors' analysis, 'Deafness' and 'Hearing Loss' are both in close proximity to 'Macular Degeneration', but have no direct connections in this small sub-set of the total MEDLINE database. Then, the authors state that "a query of the whole of MEDLINE for articles containing both 'Deafness' and 'Macular Degeneration' yielded 28 results (June 13, 2003), some of which clearly link deafness and macular dystrophy, a condition that leads to degeneration of the macula". Thus, based on the sample results, the authors are able to predict potential 'discovery' in the remainder of the MEDLINE database.

However, as a check, we ran the query (duchenne OR DMD OR dystrophy OR limb-girdle OR LGMD OR BMD) AND ("macular degeneration" and (deafness or hearing)) in PubMed covering text and MeSH fields, which would yield articles relating macular degeneration to hearing loss in the same subset the authors downloaded. In the sample, we found thirteen pre-2003 articles that contained (macular degeneration and deafness or hearing) in the text fields and/or the MeSH fields, as opposed to the zero articles the authors claimed. All the articles linked macular degeneration/macular dystrophy to some form of hearing loss. When we re-ran the query as above minus the term 'hearing', we found eleven articles. We see no evidence of potential discovery, or even innovation. The known associations date back to the mid-1970s.

In the second example, using the same subset of MEDLINE records from PubMed, the authors state " that "Insulin" and "Ferritin," among others, are positioned closely together, without these concepts co-occurring in our set of Abstracts. Again, we might infer that they are related and a PubMed search yielded 212 articles that contain both terms (June 27, 2003)."

Our check of insulin and ferritin co-occurrence in the retrieved MEDLINE subset showed no direct connections. However, the connection between iron overload or hemochromatosis and diabetes is well known. And the relationship between serum ferritin and both insulin resistance and type 2 diabetes mellitus had been well documented in the literature prior to 2003. This 'finding' does not meet our required discovery criterion of **value added.**

III. LATENT SEMANTIC CONCEPTS

A. GORDON AND DUMAIS – 1998

Gordon and Dumais [1998] used an alternative approach to Gordon's previous ODS LBD work [Gordon and Lindsay, 1996]. They used latent semantic indexing (LSI), based on singular value decomposition (SVD), to compute document and term similarity. This approach allows articles to be accessed without exact term matching, as long as they have semantic similarities. Similar to factor analysis of phrases, where variables are transformed from phrases to more general factors, LSI uses SVD to transform phrases to more general factors, or concepts, resulting in a lower dimensional space [Deerwester et al, 1990]. The terms and documents can now be expressed as a vector of statistically independent factors in the lower dimensional space. The closeness of any two terms can be estimated by the cosine of their vector expressions.

Operationally, the LSI approach eliminates concepts found in every record and in only one record. The latter is problematic, since rare events, as discussed previously, could be promising candidates for discovery.

They reported three tests, based on the MEDLINE records' Abstracts from 1983-1985. The first was to compare LSI with their previous document frequency approach for defining the pre-discovery intermediate literatures. They found excellent agreement. Since their previous approach for identifying intermediate literatures was based on high frequency phrases, it appears the LSI results are being driven in effect by high frequency phrases as well. High frequency is appropriate for defining **intermediate** literatures (but not, in our estimation, for identifying potential discovery, as we have stated previously).

The second test assumed that 'blood viscosity' was a promising intermediate literature for discovery. LSI was applied to the blood viscosity literature. Gordon and Dumais had hoped to show that Fish Oil/ Eicosapentaenoic Acid would be identified as potential discovery through LSI, but their cosine-based rankings with the term ***Raynaud's*** would not lead one to believe a priori that these substances were promising potential discoveries. Gordon and Dumais suggested, from the results of this test, that calcium dobesilate and niceritrol are two promising candidates for potential discovery, and should be examined further.

We have done a quick check of these two substances. They are both drugs, and have high MEDLINE frequency addressing RP. We have the same

problem with high frequency phenomena for discovery as discussed previously.

Before describing the third test, we need to discuss these results further. What types of substances, mechanisms, and other findings should be viewed as potential discovery? Should drugs developed to reduce blood viscosity and platelet aggregation, and reported in many tens of MEDLINE research articles, be viewed as potential discovery when applied to the specific case of RP? Even if the two drugs suggested above were to prove successful in the RP case, we view this as a modest innovation, not a scientific discovery.

The third test applied LSI to a sample retrieval of all MEDLINE records (18499 records), and bypassed the need for defining an intermediate literature explicitly. The 1000 closest neighbors to the term 'Raynaud's' were identified and 37 hand-selected 'terminals' (terms of appropriate semantic type to be potential discovery) were chosen. Considering the sub-set of these 'terminals' that did not appear in the RP core literature, the terms were completely different from those of the second test. Most, if not all, of the terms appeared to be drugs. Examination of the first five terms recommended by Gordon and Dumais for further exploration (Perhexiline, Diltiazem Hydrochloride, Nylidrin, Lidoflazine, Nitrendipine) as potential 'discovery' showed them to be drugs with high frequencies of occurrence. The average frequencies were even higher than the two substances recommended from the second test.

In summary, the LSI variant used by Gordon and Dumais appears to replicate their previous high frequency-based approach for defining intermediate literatures and to be a reasonable alternative for this purpose. Additionally, the LSI variant was able to capture high frequency phrases as potential items of interest, but did not identify low frequency items, for the sample examined. Identification of promising low frequency terms by LSI, which we believe may be the most promising items for potential discovery, remains to be demonstrated.

B. BRUZA ET AL – 2004-2006

Bruza and co-workers [Bruza et al, 2004; Cole and Bruza, 2005; Song and Bruza, 2006; Bruza et al, 2006] have generated a semantic space approach that bears some similarities to LSI. It is based on the Hyperspace Analogue to Language (HAL) model, which produces representations of words in a

high dimensional space that seem to correlate with the equivalent human representations.

HAL takes a corpus of text as input and learns a representation of words by accumulating weighted associations of co-occurring words in the context of a fixed length window. More specifically, given a vocabulary of n words drawn from the corpus in question, HAL computes an n x n matrix by moving a window of length l over the corpus by one word increments, ignoring punctuation, sentence, and paragraph boundaries. All words within the window are considered as co-occurring with strength 1. When the counts of the sliding window are aggregated, the strength of association between words becomes proportional to the distance between the words, because words that are closer together co-occur in more windows. The row and column in the HAL matrix corresponding to a given word i are added together to produce a single vector representation for that word. Thus, the vector representation of a word will be a function of the number of times component words appear in the total corpus (word frequency) and how closely the component words are spaced to the word being represented.

Once this semantic space (the HAL matrix) is generated, then discovery is possible by finding strong associations between a 'C' vector (e.g., Raynaud's) and potential 'A' vectors (e.g., Fish Oil). In the experiments reported on the RP problem, 111603 MEDLINE articles from the 1980-1985 time frame were downloaded, and only the titles were analyzed. Analogous to the LSI approach of starting with the term-document matrix, Bruza's approach starts with the HAL matrix, and weights the matrix entries using odds ratio and Dunnings log likelihood score. The reasons for performing this weighting are very insightful, and are repeated here verbatim. We could not agree more with this philosophy!

"In the construction of a semantic space there is the tacit assumption that the frequency of co-occurrence of two words u and v gives some indication of the importance of v in establishing the meaning of u. ***In literature discovery however the value of frequency in establishing a connection between words is suspect. Highly frequent co-occurrences may be part of the background knowledge and therefore it may be the very infrequent co-occurrences that contain the surprises that convey useful information to the human.*** Therefore at the very least it is desirable to correct for the frequency bias inherent in semantic space models by term weighting."

Bruza et al's results on replicating the RP problem 'discoveries' of Swanson are presented in two parts: discovery of 'B' (intermediate) terms, and discovery of 'A'-'C' connections. For 'B'-term discovery, log-likelihood appears to be far superior for both normal stop words and the Arrowsmith stop words, and outperforms odds ratio as well.

In contrast, log-likelihood does not perform well in the discovery of 'A'-'C' connections. However, information flow ranking (a heuristic form of dot product), when primed with odds ratio weights, performs well in a contextually reduced space with both 'fish' and 'oil' at the head of the ranking. Information flow ranking substantially outperforms Cosine ranking.

Because of our previously-stated concerns about the degree of 'discovery' Fish Oil has relative to the RP problem, we would need to see further applications of Bruza et al's techniques to problems not addressed previously by ODS LBD before commenting on the efficacy of his approach.

C. KOSTOFF ET AL, 2008

LSI was used to identify potentially promising 'discoveries' for the non-medical topic of water purification [Kostoff et al, 2008g]. This LSI-based approach complements the clustering-based approach mentioned previously in I-D. Relative to the clustering approach, the LSI approach offers the promise of accessing very disparate literatures readily and in a more streamlined manner. The LSI approach has the capability of starting with a core phrase(s) or Abstract, and then ranking promising phrases and Abstracts.

This LSI-based approach differs operationally from the previously discussed LSI-based approach of Gordon and Dumais [1998]. The set of documents used for potential discovery in [Kostoff et al, 2008g] was assembled via an iterated query formulation [Kostoff et al, 1997; Kostoff, 2006] in order to obtain a literature that contained both core articles and expanded articles (articles related to the core articles). This expanded (seeded) literature provided a more fruitful "hunting ground" for the LSI-based procedure to find associations, helping in part to explain the copious LSI-based discovery results obtained on the WP problem. How well the LSI approach can

identify discovery without generating a seeded literature initially remains to be demonstrated.

Operationally, a core article query was used to tag the core articles, and then the remainder was tagged as expansion. The core articles were then clustered into thematic groups, and used terminology from these groups as 'seeds' for linking to related terms from the expanded literature.

Following Gordon and Dumais [1998], the cosine similarity score was computed between the selected core terms and the expanded terms in the projected LSI space. After the expanded core terms were sorted based on their cosine similarity to the selected core term, numbers of interesting associations were obtained. High ranking terms were examined, and much potential discovery surfaced (including potential discovery from single frequency concepts). Much more detailed results (and additional capabilities) are presented in Kostoff et al, [2008g, 2007b].

Some potential discoveries include: the use of cranberry extracts for membrane fouling deterrence; the use of brominated cyclopeptides from the marine sponge Geodia barrette as a fouling deterrence mechanism; the use of essential oils from coniferous trees as fungal/bacterial deactivation agents in water purification; the use of barriers based on the epidermal barrier as alternative water filtration membranes; and the design of self cleaning membranes based on filer feeder mechanisms. The reader is referred to [Solka et al., 2007; Kostoff et al, 2007b] for a more extensive list of discoveries obtained using this strategy.

IV. MIXED TERMS

A. WREN ET AL - 2004

Wren used the standard 'A', 'B', 'C' Swanson literature relationship structure for generating potential discovery [Wren et al, 2004]. He defined classes of objects (e.g., genes, diseases, chemical compounds, etc), extracted class members from a variety of source databases, and then studied their co-occurrences in MEDLINE records (titles and Abstracts) to generate implicit relationships. He prioritized these implicit relationships by comparing actual occurrences in a literature network against a random network model to evaluate how statistically exceptional is any given set of shared relationships.

An essential component of Wren's approach is the ranking scheme used to prioritize implicit relationships. The relatedness or strength of relationship between two objects is defined as the number of times that the two objects have been co-mentioned (co-occurrence frequency) and the probability that each co-mention represents a non-trivial relationship. Obtaining co-occurrence frequencies is straight-forward; estimating probability of a non-trivial relationship is much more subjective and difficult. Wren estimated the latter by tracking the persistence of MEDLINE relationships over time. If a relationship was observed before a specified date, and observed very infrequently after that, it was assumed to be non-important, and down-weighted accordingly. Thus, the key element to his ranking scheme is co-occurrence frequency and importance.

Characterization and Discovery

Before proceeding to the 'discovery' hypothesis Wren generates, we will discuss his basic approach further. Two key aspects of ODS LBD are characterization and discovery. Characterization is identifying and describing important patterns in text. In general text mining, characterization involves identifying the technical infrastructure in a retrieved database (e.g., key researchers, Centers of Excellence, etc) and the technical structure (pervasive technical thrusts, relationships among thrusts, etc). Scientific discovery is not involved, but identification of these relational patterns is an important output.

Characterization is a quantity-based phenomenon. An essential component of characterization is occurrence and co-occurrence frequency, since statistical patterns are primary. Low frequency phenomena, interesting though they may be, tend to be neglected in the characterization process.

Discovery is different; it is a quality-based phenomenon. It requires human judgment to link multiple concepts into a novel concept. Many of the techniques in this review, especially Wren's, assume that quantity-based relationships can be associated with quality, and by implication, discovery. There is no reason to believe that discovery can be estimated through quantity-based relationships, whether they are occurrence frequency-based, co-occurrence frequency based, or any other quantitative metric-based.

Intuitively, one would expect the opposite, especially for disciplines one link apart. As in the RP Fish Oil case, if there are large numbers of researchers examining Fish Oil for its effects on blood viscosity and blood clotting, shouldn't one expect that at least some of these researchers would be cognizant of potential applications to almost any problem dealing with blood circulation? And wouldn't this be reflected in the medical literature to some extent, as we showed to be the case with RP? For disciplines separated by a number of links (very disparate literatures), then higher frequency phenomena would have a greater chance of being candidates for real discovery, since the odds of their being recognized as applications in the target literature would be reduced because of their distance from the target literature.

Wren's Hypothesized 'Discovery'

Wren searched for potential discovery in treating cardiac hypertrophy. His ranking technique showed the drug chlorpromazine (CPZ) shared many implicit relations with cardiac hypertrophy, and he then inferred that it might be useful for reducing the progression of cardiac hypertrophy. There does not seem to be prior art in the journal literature. There is at least one patent that addresses the link [Finer and Chabala, 2002]. It covers a wide swath and addresses phenothiazine derivatives, not limited to chlorpromazine:

"Title: Use of ***phenothiazine derivatives*** for the treatment of e.g. cellular proliferative diseases

USE - For treating cellular proliferative diseases; disorder associated with KSP kinesin activity including cancer, hyperplasia, restenosis, ***cardiac hypertrophy***, immune disorders and inflammation; inhibiting KSP kinesin (all claimed)."

CPZ is a phenothiazine compound used primarily as an anti-psychotic for humans. While other phenothiazine compounds such as thioridazine have well-documented histories of strong association with cardiac arrythmias, CPZ also has a history of cardiac adverse effects on humans.

Additionally, there are a large number of potentially adverse side effects from the use of CPZ. These include **cardiac side-effects**, such as: EKG changes (particularly nonspecific Q and T wave distortions [induction of QT prolongation and torsades de pointes] - Sudden death, apparently due to

cardiac arrest, has been reported); arrhythmogenic side effects caused by blockade of human ether-a-go-go-related gene (HERG) potassium channels; simple tachycardia. **Other side-effects** include: Neuroleptic Malignant Syndrome; neuromuscular reactions (tardive dyskinesia; dystonias, motor restlessness, pseudo-parkinsonism); convulsive seizures (petit mal and grand mal); lowered seizure thresholds; bone marrow depression; prolonged jaundice; hyperreflexia or hyporeflexia in newborn infants whose mothers received phenothiazines; drowsiness; hematological disorders, including agranulocytosis, eosinophilia, leukopenia, hemolytic anemia, aplastic anemia, thrombocytopenic purpura and pancytopenia; postural hypotension, momentary fainting and dizziness; cerebral edema; abnormality of the cerebrospinal fluid proteins; allergic reactions of a mild urticarial type or photosensitivity; exfoliative dermatitis; asthma, laryngeal edema, angioneurotic edema and anaphylactoid reactions; amenorrhea, gynecomastia, hyperglycemia, hypoglycemia and glycosuria; corneal and lenticular changes, epithelial keratopathy and pigmentary retinopathy; some respiratory failure following CNS depression; paralytic ileus; thermoregulation difficulties.

Given CPZ's known history of adverse side effects, including cardiac effects, we question CPZ as a value-added discovery for potential treatment of cardiac hypertrophy or any cardiac-related problem. To validate our perceptions, we contacted two experts in cardiac hypertrophy about potential use of CPZ, and were told "there is no sufficient evidence that would support pursuing CPZ for treating cardiac hypertrophy in humans" and "link to hypertrophic cardiomyopathy was not clear".

This example illustrates the problem with using quantity-based measures to associate with quality predictions. Wren's ranking method emphasizes co-occurrences and persistence of relationships. If CPZ has a persistent and frequent history of being associated with adverse cardiac effects, both directly and as a member of a class (phenothiazines) even more strongly associated with adverse cardiac effects, then it would have a strong implicit relationship with cardiac hypertrophy. The quality of the total somatic relationship is not necessarily positive, as this example shows. While Wren ran some lab experiments showing that CPZ reduced cardiac hypertrophy in mice, the relation may reflect a local optimization on cardiac hypertrophy, and a global sub-optimization on overall somatic well-being. Proposed new therapeutic interventions need to meet several challenging hurdles including

proven safety, tolerability, and efficacy in improving clinically meaningful end-points such as overall mortality.

B. **HRISTOVSKI ET AL – 2005 - BITOLA**

This research group [Hristovski et al, 2005] uses semantic predications to enhance co-occurrence-based ODS LBD systems. The predications are produced by the combined application of two natural language processing systems, BioMedLEE and SemRep, coupled with an ODS LBD system BITOLA.

BITOLA is an ODS LBD system (presently applicable to the medical field) based on Swanson's original approach, with notable differences. BITOLA uses MeSH terms instead of title words, and uses association rules to relate medical concepts instead of word frequencies. Concept co-occurrence is used to indicate relations between concepts. If 'C' is a starting concept (e.g., a disease for which treatments are desired), then its co-occurrences with all concepts 'B' are found ('B', the intermediate concepts, could be pathological functions, symptoms, etc), followed by all co-occurrences of 'A' (potential treatments) with each 'B'. The possible number of 'C' -→ 'A' combinations can be extremely large, as is the case with all the techniques examined in this review.

To reduce the number of combinations to manageable size, BITOLA incorporates filtering and ordering capabilities. The 'B' and 'A' concepts can be limited by semantic types (e.g., 'A' concepts can only be drugs). Thresholds can be placed on the association rules (e.g., if an association rule is of the form 'C' -→ 'B' (confidence, support), where in ***confidence*** percent of articles containing 'C', 'B' is present, and there are ***support*** number of such articles, then thresholds can be placed on the ***confidence*** and ***support*** measures for filtering). The default ordering is by the decreasing association rule ***confidence***, but it is also possible to order by ***support*** or semantic type. In sum, the system is driven by co-occurrences.

BioMedLEE is based on a grammer formalism that combines syntax and semantics and uses a lexicon derived from clinical documents, the UMLS, and other online biomedical sources. For the ODS LBD application discussed in this review, BioMedLEE focuses on use of the concepts in the UMLS Metathesaurus only.

SemRep is a symbolic natural language processing system for identifying semantic relations in biomedical text. The program currently focuses on MEDLINE citations emphasizing treatment of disease. SemRep identifies a variety of semantic predications, but for the ODS LBD application, the most relevant relation (predication) is ***Treats***.

To exploit semantic predications in ODS LBD, the method introduces the notion of discovery patterns, which contain a set of conditions to be satisfied for the discovery of new relations between concepts. The focus in the published paper is on the ***Maybe_Treats*** pattern, which has two forms: Maybe_Treats1 and Maybe_Treats2. In both forms, the goal is to propose potentially new treatments. The first form Maybe_Treats1 is satisfied when there is a change in a substance, body function, or body measurement (concept 'B') associated with the starting disease 'C', and there is also an opposite change in concept 'B' associated with the concept 'A'. In Maybe_Treats2, in order to find a potentially new treatment for a starting disease 'C', a first search is made for another disease C2 that has similar characteristics (B2 substances of functions changed in the same direction. Then, the drugs (A2) already used to treat the disease C2 can be proposed as a potentially new treatment for disease 'C' if there is no evidence in the literature that A2 is already used to treat 'C'.

While the approach makes use of additional information through the associations, it is still fundamentally a co-occurrence-based concept, with all the deficiencies mentioned previously. Rare events that could lead to radical discovery would get low prioritization, and well-known relations would predominate. At best, this approach would be expected to provide a very incremental amount of potential discovery.

Two examples of potential discovery were provided. In the first, the RP-Fish Oil 'potential discovery' was replicated, although due to the location of blood viscosity and platelet aggregation in the rankings, it is not clear that Fish Oil would have been the top ranked solution concept had the results not been known a priori.

The second example was Huntington Disease (HD). Because of similarities of HD with diabetes mellitus, especially reduced levels of insulin, the authors suggest insulin treatment might be an interesting drug for HD.

These two results confirm the concerns stated relative to high frequency phenomena, and crystallize the problems with depending on high frequency correlations. In both cases, the researchers understood the linkages that constituted the potential 'discovery'. The RP case was discussed previously. In the HD case, the relationship between insulin and HD was also known to HD researchers. Abnormal glucose tolerance tests were reported in HD in 1977. In the transgenic mouse model of HD, the mice developed diabetes, and then insulin was used to treat the diabetes [e.g., Hurlbert et al, 1999].

We contacted an expert in HD research about the potential use of insulin, and were told that "the diabetes in early HD is not type 1 diabetes, but is due to insulin release problems rather than an insulin insufficiency….I do not think that treating HD patients with insulin is a good option, unless they are insulin-dependent diabetics"…. The key point here is that if two literatures are disjoint, there may be multiple reasons. It could mean that their union would produce real discovery, and no one had thought of linking them previously. Or, it could mean that their union had been considered previously, and researchers concluded that there was nothing to be gained by the linkage.

DISCUSSION AND CONCLUSIONS

Of the ODS LBD concepts that we have examined in detail, and the other papers we reviewed but did not address in this report, we believe that most, if not all, of the ODS LBD concepts have not generated potential discovery in the sense defined above. Some of the concepts may have generated potential innovation.

We believe the approaches and assumptions made by the majority of the researchers reviewed militate against discovery, and drive the results toward innovation. Almost all the techniques aim for identifying discovery in the 'BC' literature defined previously, which we view as being directly related to the starting 'AB' literature. Most of the techniques are based on correlation and/or co-occurrence phenomena, which are excellent for characterizing a literature, but are questionable for identifying potential discovery.

High occurrence or co-occurrence frequencies are required for such approaches. This translates pragmatically into many researchers working on a technique in the 'AB' literature. The high frequency approaches

therefore assume that none of these relatively large numbers of researchers in the 'AB' target literature would be aware of the applications in the directly related 'BC' literature. We believe such an assumption may be unrealistic for the directly related literatures (although exceptions are always possible), and our belief is validated by the lack of any discovery that we perceive in these papers. We believe the very lowest frequency concepts would have the highest probability for potential discovery in the directly related literatures, and these rare concepts are effectively excluded by the high frequency-based techniques. As the more indirectly-related literatures are examined (for example, if we have an 'A-B-C-D-E' system, where 'D' and 'E' literatures are related more indirectly to the starting 'A' literature), we believe that use of the higher frequency techniques may be somewhat more realistic for identifying discovery, although low frequency phenomena still offer a higher probability that knowledge of the starting literature would be unknown to researchers in the proposed 'discovery' literature. Significantly additional research is required on different techniques (both theoretical and practical) to determine which (if any) quantity-based methods are ultimately useful, and how they might be used on conjunction with the other methods evaluated in this report.

The ODS LBD researchers tend to use ranking metrics for potential discovery selection, in order to reduce the number of potential discovery candidates to be examined from the typically vast pool of potential candidates retrieved. These ranking metrics for potential discovery selection also tend to be frequency-based. Two are commonly used: 1) the frequency of a proposed 'discovery' concept in the literature sub-set of interest (e.g., the frequency of occurrence of 'Fish Oil' in the Blood Viscosity literature) is required to be much higher than the frequency of this proposed 'discovery' concept in the overall source database (e.g., MEDLINE in the RP example), and 2) the larger the number of paths between the 'BC' and 'AB' literatures, the higher the ranking (e.g., the more mechanisms by which Fish Oil can impact the major characteristics of RP) of the proposed 'discovery' concept. We believe there is no scientific basis for such ranking metrics, and their use militates against the more infrequent concepts that could represent radical discovery. This does not exclude potential discovery being identified with high values of these two ranking metrics.

Many of the more recent approaches use MeSH variables in place of text variables. While MeSH has certain advantages, such as a lower-dimensional space and the ability to include concepts that may not be represented by

specific words or phrases in text, it has some glaring disadvantages. Very recent MEDLINE articles are not yet indexed in MEDLINE, and therefore will not be accessed by a purely MeSH-based query. Also, the indexers make mistakes, and do not include all the MeSH terms that should be included in an indexing. Potential discoveries will not appear, because they could not be accessed by a purely MeSH-based query. Finally, in the vetting process, the breadth of MeSH terms opens the door to the use of many proxy terms for generating prior art and thereby refuting potential discovery.

Given:

- the length of time since Swanson's pioneering paper (two decades),
- the massive number of medical and technical problems in need of discovery,
- the relatively few articles published in the literature using existing ODS LBD approaches to generate discovery (especially articles not published by the Swanson/ Smalheiser team and not replicating the initial RP results), and
- concerns about the validity of the potential 'discoveries' reported

it is clear that improvements in the fundamental ODS LBD approach and its dissemination and acceptability are required, and the specific critiques shown identified strategic problems that need to be addressed.

Finally, the major operational problem that emerged was insufficient 'vetting' (by ODS LBD researchers and journal reviewers alike) of the published ODS LBD studies' hypothesized discoveries; in other words, either 1) not checking in detail or 2) not using sufficiently stringent conditions that these concept linkages had not been identified previously and their linking provided value added. This issue of prior art was addressed in each of the studies reviewed, and placed in the larger context of the strength of the proposed 'discovery'.

REFERENCES

Angelico F, Amodeo P. (1978). Eicosapentaenoic acid and prevention of atherosclerosis. *Lancet*. 2 (8088): 531-531.

Babish JG, Howell T, Pacioretty L, Howell TM, Pacioretty LM (2003). Composition for treating e.g. inflammation or inflammation based diseases, comprising curcuminoid species and alpha- or beta-acid. *Patent Number US2003096027-A1*. 22 May.

Balasubramanyam, M, Koteswari, A , Kumar RS , Monickaraj SF, Maheswari JU, Mohan V. (2003). Curcumin-induced inhibition of cellular reactive oxygen species generation: novel therapeutic implications. *J Biosci.* 28(6):715-21.

Bang H.O., Dyerberg J. (1981). Personal reflections on the incidence of ischemic-heart-disease in Oslo during the World-War-2. *Acta Medica Scandinavica*. 210 (4): 245-248.

Biswas NR, Gupta SK, Das GK, Kumar N, Mongre PK, Haldar D, Beri S. (2001). Evaluation of Ophthacare eye drops--a herbal formulation in the management of various ophthalmic disorders. *Phytother Res*. Nov;15(7):618-20.

Black K.L., Culp B, Madison D, Randall O.S.. (1979). Protective effects of dietary fish oil on focal cerebral infarction. *Prostaglandins and Medicine*. 3 (5): 257-268.

Bruza, P., Cole, R., Song, DW., Bari, Z. (2006). Towards operational abduction from a cognitive perspective. *Logic Journal of the IGPL*. 14 (2): 161-177.

Bruza, P., Song, DW., McArthur, R. (2004). Abduction in semantic space: Towards a logic of discovery. *Logic Journal of the IGPL*. 12 (2): 97-109.

Callon, M., J. P. Courtial, et al.. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. *Scientometrics.* 22(1): 155-205.

Cartwright I.J., Pockley A.G., Galloway J.H., Greaves M, Preston F.E. (1985). The effects of dietary omega-3 poly-unsaturated fatty-acids on erythrocyte-membrane phospholipids, erythrocyte deformability and blood-viscosity in healthy-volunteers. *Atherosclerosis*. 55 (3): 267-281.

CLUTO (2006). A clustering toolkit. http://www.cs.umn.edu/˜cluto.

Cole, RJ; Bruza, PD. (2005). A bare bones approach to literature-based discovery: An analysis of the Raynaud's/fish-oil and migraine-magnesium discoveries in semantic space. *Discovery Science*. Proceedings. 3735: 84-98.

Culp B.R., Lands W.E.M., Lucchesi B.R., Pitt B, Romson J. (1980). The effect of dietary supplementation of fish oil on experimental myocardial-infarction. *Prostaglandins*. 20 (6): 1021-1031.

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. 41 (6): 391-407.

Dyerberg J, Bang H.O., Stoffersen E, Moncada S, Vane J.R. (1978). Eicosapentanoic acid and prevention of thrombosis and atherosclerosis. *Lancet*. 2 (8081): 117-119.

Finer JT and Chabala JC. (2002). Use of phenothiazine derivatives for the treatment of e.g. cellular proliferative diseases. *Patent Number: WO200257244-A1*. 25 July.

Ganiz, M.C., Pottenger, W.M., Janneck, C.D. (2005). Recent advances in literature-based discovery. *Technical Report.* Lehigh University (LU-CSE-05-027).

Gordon MD, Lindsay RK. (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 (2): 116-128.

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

Hay CRM, Durber AP, Saynor R. (1982). Effect of fish oil on platelet kinetics in patients with ischemic-heart-disease. *Lancet*. 1 (8284): 1269-1272.

Healey P, Rothman H, Hoch PK. (1986). An experiment in science mapping for research planning. *Research Policy*. 15 (5): 233-251.

Hristovski, D; Peterlin, B; Mitchell, JA; Humphrey, SM. (2005). Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*. 74 (2-4): 289-298.

Hurlbert MS, Zhou W, Wasmeier C, Kaddis FG, Hutton JC, Freed CR. (1999). Mice transgenic for an expanded CAG repeat in the Huntington's disease gene develop diabetes. *Diabetes*. 48(3):649-51.

Jakubowski JA, Ardlie NG. (1979). Evidence for the mechanism by which eicosapentaenoic acid inhibits human-platelet aggregation and secretion - implications for the prevention of vascular-disease. *Thrombosis Research*. 16 (1-2): 205-217.

Komatsu A. (2003). Preparation of health drink, involves processing preset amount of dry turmeric powder, dry Curcuma zedoaria powder, dry Curcuma wenyujin powder and sea tangle powder with distilled white liquor at specific temperature. Patent Number(s): JP2003189819-A. July 8.

Kostoff, R.N. (1993). Co-word clustering. in *Assessing R&D Impacts: Method and Practice*. Bozeman, B. and Melkers, J., Eds. (Kluwer Academic Publishers, Norwell, MA).

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. (1997). Database Tomography for information retrieval. *Journal of Information Science*. 23:4, 301-311.

Kostoff RN. (2003). Stimulating innovation. *International Handbook of Innovation*. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 388-400.

Kostoff, R.N. (2006). Systematic acceleration of radical discovery and innovation in science and technology. *Technological Forecasting and Social Change*. 73 (8): 923-936.

Kostoff, R.N. (2007a). Validating discovery in literature-based discovery. *Journal of Biomedical Informetrics*. 40(4):448-50.

Kostoff, R.N., Block, J.A., Solka, J.A., Briggs, M.B., Rushenberg, R.L., Stump, J.A., Johnson, D., Lyons, T.J., Wyatt, J.R. (2007b). Literature-related discovery. *DTIC Technical Report Number* ?????? (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA.

Kostoff, R.N. (2008a). Literature-Related Discovery (LRD): Introduction and background. *Technological Forecasting and Social Change.* Special Issue on Literature-related discovery.

Kostoff, R.N. (2008b). Literature-related discovery (LRD): Methodology. *Technological Forecasting and Social Change.* Special Issue on Literature-Related Discovery.

Kostoff, R.N., Block, J.A., Stump, J.A., Johnson, D. (2008c). Literature-related discovery (LRD): Potential treatments for Raynaud's Phenomenon. *Technological Forecasting and Social Change.* Special Issue on Literature-Related Discovery.

Kostoff, R.N. (2008d). Literature-related discovery (LRD): Potential treatments for cataracts. *Technological Forecasting and Social Change.* Special Issue on Literature-Related Discovery.

Kostoff, R.N., Briggs, M.B. (2008e). Literature-related discovery (LRD): Potential treatments for Parkinson's Disease. *Technological Forecasting and Social Change.* Special Issue on Literature-Related Discovery.

Kostoff, R.N., Briggs, M.B., Lyons, T. (2008f). Literature-related discovery (LRD): Potential treatments for Multiple Sclerosis. *Technological Forecasting and Social Change.* Special Issue on Literature-Related Discovery.

Kostoff, R.N., Solka, J.A., Rushenberg, R.L., Wyatt, J.R. (2008g). Literature-related discovery (LRD): Potential improvements in water purification. *Technological Forecasting and Social Change.* Special Issue on Literature-Related Discovery.

Kostoff, R.N., Block, J.A., Solka, J.A., Briggs, M.B., Rushenberg, R.L., Stump, J.A., Johnson, D., Lyons, T.J., Wyatt, J.R. (2008h). Literature-related discovery (LRD): Lessons learned, and future research directions.

Technological Forecasting and Social Change. Special Issue on Literature-Related Discovery.

Kromhout D, Bosschieter EB, Coulander CD. (1985). The inverse relation between fish consumption and 20-year mortality from coronary heart-disease. *New England Journal of Medicine*. 312 (19): 1205-1209.

Moncada S, Vane J.R. (1979). Arachidonic-acid metabolites and the interactions between platelets and blood-vessel walls. *New England Journal of Medicine*. 300 (20): 1142-1147.

Natarajan C, Bright JJ. (2002). Curcumin inhibits experimental allergic encephalomyelitis by blocking IL-12 signaling through Janus kinase-STAT pathway in T lymphocytes. *Journal of Immunology*. 168 (12): 6506-6513.

Okamoto T, Yamagishi S, Inagaki Y, Amano S, Koga K, Abe R, Takeuchi M, Ohno S, Yoshimura A, Makita Z. (2002). Angiogenesis induced by advanced glycation end products and its prevention by cerivastatin. *FASEB J*. 16(14):1928-30. Epub 2002 Oct 4.

Pratt, W. and Yetisgen-Yildiz, M. (2007). Response to "Validating discovery in literature-based discovery". *Journal of Biomedical Informatics*, 40(4), pp.450-452.

Salh B, Assi K, Templeman V, Parhar K, Owen D, Gomez-Munoz A, Jacobson K. (2003). Curcumin attenuates DNB-induced murine colitis. *American Journal of Physiology-Gastrointestinal and Liver Physiology*. 285 (1): G235-G243 Jul. See also Varga C, Cavicchi M, Orsi A, Lamarque D, Delchier JC, Rees D, Whittle BJ. (2001). Beneficial effect of P54, a novel curcumin preparation in TNBS-induced colitis in rats. *Gastroenterology*. 120 (5): A691-A691 3732 Suppl. 1.

SIGN. (1998). Drug therapy for peripheral vascular disease: a national clinical guideline. *Scottish Intercollegiate Guidelines Network*. SIGN Publication Number 27. http://www.sign.ac.uk/pdf/sign27.pdf.

Smalheiser NR, Swanson DR. (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses *Computer Methods and Programs in Biomedicine*. 57 (3): 149-153.

Smalheiser NR. (2005). The Arrowsmith project: 2005 status report. *Lecture Notes in Computer Science*. 3735: 26-43.

Solka, J.L., Rushenberg, R.L., Kostoff, R.N., Tucey, N., and Bryant, A. (2007). New methods of water purification obtained using literature-based discovery, *DTIC Technical Report Number* ?????? (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA.

Song DW, Bruza P. (2006). Text based knowledge discovery with information flow analysis. *Lecture Notes in Computer Science*. 3841: 692-701.

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

Srinivasan P, Libbus B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*. 4;20 Suppl 1:I290-I296.

Srinivasan P., Libbus B. and Sehgal A. K. (2004). Mining MEDLINE: Postulating a beneficial role for curcumin longa in retinal diseases. *HLT Biolink.*

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

Sugimoto K, Hanai H, Aoshi T, Tozawa K, Uchijima M, Nagata T, Koide Y. (2002a). Curcumin ameliorates trinitrobenzene sulfuric acid (TNBS) - Induced colitis in mice. *Gastroenterology*. 122 (4): A395-A396 T993 Suppl. 1, April.

Sugimoto K, Hanai H, Tozawa K, Aoshi T, Uchijima M, Nagata T, Koide Y. (2002b). Curcumin prevents and ameliorates trinitrobenzene sulfonic acid-induced colitis in mice. Gastroenterology. Dec;123(6):1912-22.

Swanson DR. (1986). Fish oil, Raynauds Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*. 30 (1): 7-18.

Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*. 91(2). 1997.

Torvik VI, Smalheiser NR. (2007). A quantitative model for linking two disparate sets of articles in MEDLINE. *Bioinformatics*. 2007 Jul 1;23(13):1658-65.

Ukil A, Maity S, Karmakar S, Datta N, Vedasiromoni JR, Das PK. (2003). Curcumin, the major component of food flavour turmeric, reduces mucosal injury in trinitrobenzene sulphonic acid-induced colitis. *British Journal Of Pharmacology*. 139 (2): 209-218 May.

Valdes-Perez, R.E. (1999). Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence*, 107. 335-346.

van der Eijk CC, van Mulligen EM, Kors JA, Mons B, van den Berg J. (2004). Constructing an associative concept space for literature-based discovery. *Journal of the American Society for Information Science and Technology*. 55 (5): 436-444.

Weeber M, Klein H, de Jong-van den Berg LTW, et al. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*. 52 (7): 548-557.

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

Woodcock B.E., Smith E, Lambert W.H., Jones W.M., Galloway J.H., Greaves M, Preston F.E. (1984). Beneficial effect of fish oil on blood-viscosity in peripheral vascular-disease. *British Medical Journal*. 288 (6417): 592-594.

Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*. 20 (3): 389-398.

Yetisgen-Yildiz M, Pratt W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*. 39 (6): 600-611.