Where is the Discovery in Literature-Based Discovery?

R.N. Kostoff

Abstract This chapter addresses the core of literature-based discovery (LBD), namely, what is discovery and how is the generation of discovery confirmed. The chapter starts with definitions of discovery and innovation, especially in the LBD context, and then proceeds to describe radical discovery and LBD. It then describes the vetting necessary to confirm the presence of discovery. Finally, the chapter concludes with a few examples where use of more comprehensive vetting techniques would have been prudent before discovery was reported. The LBD focus is on open discovery systems (start with a problem, discover a solution, or vice versa) exclusively.

1 Discovery and Innovation 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 [42]. 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) [17]. 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.

In the LBD context, discovery is linking two or more literature concepts that have heretofore not been linked (i.e., disjoint), in order to produce novel, interesting,

R.N. Kostoff
307 Yoakum Parkway, Alexandria, VA 22304, USA
rkostoff@mitre.org

P. Bruza and M. Weeber (eds.), *Literature-based Discovery,*
Springer Series in Information Science and Knowledge Management 15.

plausible, and intelligible knowledge. Thus, simply linking two or more disparate concepts is a necessary, but not sufficient, condition for LBD. 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 proposed as potential discovery) will be shown later in this chapter.

Also, in the LBD context, innovation is the exploitation of a discovery linkage, mainly the identification of a linkage that was not being exploited at a sufficient pace.

More generally, *radical discovery* and *radical innovation* depend on the source of the inspiration and/or the magnitude of the impact. The more disparate the source of ideas from the target problem discipline, the more radical the potential discovery or innovation. The greater the magnitude of change/impact resulting from the discovery or innovation, the more radical the potential discovery or innovation.

2 Radical Discovery

Discovery and innovation are the cornerstones of frontier research. One of the methods for generating radical discovery and innovation in a target discipline is to use principles and insights from disciplines very disparate to the target discipline, to solve problems in the target discipline.

The challenge has become more critical due to increasing specialization and effective isolation of technical/medical researchers and developers [16]. As research funding and numbers of researchers have increased substantially over the past few decades, the technical literature has increased substantially as a result. Researchers/developers struggle to keep pace with their own disciplines, much less to develop awareness of other disciplines. Thus, we have the paradox that the expansion of research has led to the balkanization of research! The resulting balkanization serves as a barrier to cross-discipline knowledge transfers, and retards the progress of discovery and innovation [16].

As a result, identifying these linkages between the disparate and target disciplines, and making the subsequent extrapolations has tended to be a very serendipitous process. Until now, there has been no fully systematic approach to bridging these unconnected target and disparate disciplines.

Once the principles and associated techniques have been established for producing insights from these disparate literatures, many applications are possible. These include:

1. Promising opportunities for researchers to pursue
2. Promising new Science & Technology (S&T) directions for program managers to pursue
3. Promising leads for intelligence analysts to pursue

3 Literature-Based Discovery

The pioneering LBD study was reported in Swanson's paper hypothesizing treatments for Raynaud's Disease [36]. Many subsequent open and closed system LBD studies were performed by Swanson/Smalheiser, including migraine and magnesium [37], somatomedin-C and arginine [38], and potential biowarfare agents [40]. They also developed more formalized analytical techniques for hypothesizing radical discovery [29, 39]. Other researchers have used variants of Swanson's LBD approach for hypothesizing radical discovery in open and closed discovery systems, but only open discovery systems will be addressed here.

Gordon and Lindsay [10] used an information technology-based approach to help automate the LBD process. Weeber et al. [45] used a two step model of discovery (open discovery step followed by closed discovery step) to simulate Swanson's actual discovery. Further, Weeber et al. [46] identified potentially new target diseases for the drug thalidomide. Stegmann and Grohmann [33] used a co-word clustering of MeSH terms to identify potential discovery by location on density-centrality maps. Srinivasan [30] generated a potential discovery-identifying algorithm that operated by building MeSH-based profiles from Medline for topics. Yetisgen-Yildiz and Pratt [49] use an LBD system called LitLinker that incorporated knowledge-based methodologies with a statistical method. Van der Eijk et al. [43] mapped from a co-occurrence graph to an Associative Concept Space (ACS), to identify discovery from concepts that were close to each other in ACS but had no direct connections. Gordon and Dumais [9] used latent semantic indexing, based on higher order co-occurrences, to compute document and term similarity. Bruza et al. [5,6] generated a semantic space approach based on the Hyperspace Analogue to Language to produce representations of words in a high dimensional space. Wren et al. [48] defined classes of objects, extracted class members from a variety of source databases, and then studied their co-occurrences in Medline records to generate implicit relationships. Hristovski et al. [12, 13] used semantic predications to enhance co-occurrence-based LBD systems.

The general theory behind this approach, applied to two separate literatures, is based upon the following considerations [36].

Assume that two literatures with disjoint components 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." From these combinations, linkages can be generated through the "b" themes that connect both literatures (e.g., AB $\rightarrow$ 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.

For example, as shown in Swanson's initial LBD paper, 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 Raynaud's disease (theme "c" from literature BC). There was no mention of eicosapentaenoic acid in the Raynaud's disease literature, but the acid was linked to the disease through the blood viscosity themes in both literatures [36].

A central problem with all the LBD studies that have been reported in the open literature is the absence of a gold standard that can be used as a basis of comparison. A true gold standard would allow comparisons of quality and quantity of potential discoveries. Many of the studies use Swanson's results (Fish Oil and Eicosapentanoic Acid) as a comparison standard. As I point out later, I have questions as to whether Swanson's hypotheses are true discoveries or are really innovations, and 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 a couple of main choices. If we have an exact solution to the problem, we can compare the predictive tool 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.

In LBD, we dont know the extent of discovery possible for any problem, and therefore are not able to estimate the comprehensiveness of any approach (recall). Further, we are not able to estimate the quality of any discovery until much testing has been done, and therefore cannot estimate the fraction of the potential discoveries identified that are in fact potential discoveries (precision).

For the LBD approaches reported in the literature, there appears to be an imbalance between the prediction of potential discovery and its validation. Most of the effort seems to have focused on the front end of the process (discovery candidate identification) with little effort on the back end (vetting of potential discovery predictions). As I will show, this has allowed non-discovery items to be represented as discovery.

As a result, I believe this insufficient vetting has contributed to the slowing of LBD implementation. LBD intrinsically has powerful capabilities, and one would have expected that, two decades after Swanson's initial paper, there would be treatments proposed for all the major chronic degenerative diseases, similar implementations for their non-medical equivalents, as well as major sponsored research programs on LBD throughout the world. As far as I know, no major clinical trials have been reported on LBD-driven hypotheses, and benefits resulting from these LBD studies have yet to be realized.

Given:

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

It is clear that improvements in the fundamental LBD approach and its dissemination and acceptability are required.

My text mining group has been working on improving LBD for the past few years. The general approach we have followed was reported in 2006 [18]. We have

used our specific versions of LBD on five problems (four medical, one non-medical), and have generated voluminous potential discovery for each problem. I believe we have 'cracked the code' on LBD. Our results constituted the Special Issue of the journal Technological Forecasting and Social Change, February 2008. The remainder of this chapter is focused on the potential discovery vetting procedures we have used, and includes some examples of applying our vetting procedures to discoveries that have been reported in the LBD literature.

4 Validating Potential Discovery (Vetting)

The purpose of our vetting procedures is to insure that what we report as potential discovery has not been found in the literature previously, and obeys the criteria for discovery set forth at the beginning of this chapter. If a concept has been found in the literature previously, but we believe its reporting would accelerate its development, we might report it as an innovation. We have instituted a four step vetting process that balances thoroughness with pragmatism.

The first step is to check for appearance of the potential discovery concept in the core target problem research literature. How do we define this literature? Ideally, every research document published globally in the core problem area would constitute this literature. The practical compromise we have made is to define the source literature for the core target problem literature as the Science Citation Index and Medline. While I believe this is a bare minimum core literature requirement to search for prior art/science, some examples shown in the next section illustrate that even this threshold requirement was not met before potential discovery was published.

In this first step, we operationally check for the intersection of the core target problem literature with the potential discovery literature. If the intersection is a null set, the first check is successful. Thus, if we check whether Fish Oil is a potential discovery for Raynaud's Disease, we might use the query Fish Oil (or its many specific variants) and Raynaud's Disease (or its variants), and see whether any records are retrieved. The real issue here, as will be discussed later, is how broadly or narrowly we define the core target problem literature and the potential discovery concept literature. The breadth of definition could determine whether we have generated discovery, innovation, or nothing. For example, Fish Oil may or may not be a discovery, depending on whether we define the Raynaud's Disease literature to include or exclude the Peripheral Vascular Disease literature.

The second step could be viewed as a continuation of the first step. We go beyond simple intersection to see whether there are citation linkages between the potential discovery concept and the core target problem literature that would indicate researchers were aware of the linking between these literatures previously. There are many types of citation linkages (citing papers, cited papers, papers that share common references, papers that share common citing papers, etc). Depending on how far we plan to proceed with a potential discovery (e.g., do we want to patent the potential discovery), we check at least the citing papers for linkages between the concept literature and the problem literature.

The third step is checking the patent literature. This is more difficult than the first step because of the typically wide breadth and scope of the claims in each patent.

All steps are run serially. Once the first three steps have been taken successfully, we then have the potential discovery candidate concepts examined by experts. We access two types of experts: those expert in the core target problem literature (e.g., Raynaud's Disease), and those expert in the potential discovery concept literature (e.g., Fish Oil). We ask the experts in the core target problem literature whether the potential discovery concept is indeed discovery (i.e., have they seen it before in the target problem context), and we ask the experts in the potential discovery concept literatures whether the concept could be extrapolated to the target problem. If we report potential discovery concepts that have been only partially vetted, we state that fact.

5 Examples of Validation Issues

This section presents examples of applying some of our vetting techniques to potential discoveries reported in the LBD literature.

5.1 Use of MeSH Variables

An LBD approach based on the analysis of actual text phrases is intrinsically a high-dimensional process, due to the large number of words/phrases in the literature. To circumvent this dimensionality problem, LBD researchers have used approaches that convert the problem from high-dimensional to low-dimensional. One widely used approach reported in recent LBD papers [30, 43, 49] is the use of MeSH terms instead of text terms. MeSH is a taxonomy (controlled vocabulary) in the major medical database (MEDLINE). MeSH is generated by independent indexers who read each MEDLINE article, then assign selected MeSH terms to each article. There are approximately 22,500 MeSH terms in the total MEDLINE taxonomy, orders of magnitude less than the number of text words/phrases.

The positive aspects of using MeSH terms, in addition to the reduced number of variables, are that relevant articles can be retrieved containing desired concepts but not necessarily specific text terminology. Thus, a query with a very small number of MeSH terms (e.g., lung neoplasms) can retrieve many lung cancer records that would have required perhaps hundreds of text query terms to have the same degree of coverage, and many of those retrieved records might not contain the terms lung neoplasms or lung cancer.

On the negative side, MeSH terms are restricted to the medical literature. Additionally, very recent MEDLINE records have not been indexed with MeSH terms, and would be inaccessible for LBD purposes unless text terms were added (thereby defeating one of the major reasons for selecting MeSH terms).

Further, the mapping from text terms to MeSH terms is not one-to-one, nor is it conservative like transforming from thermodynamic variables (e.g., pressure, temperature, density) to conservation variables (e.g., new variables that include combinations of the thermodynamic variables and are conserved across discontinuities, such as mass, momentum, energy) in a fluid flow system [21]. There is a well-known phenomenon called the indexer effect [11], which states essentially that indexers are fallible, and they make errors and omissions. Not all MeSH terms that should be assigned to an article are in fact assigned by the indexers. For many uses of retrievals from MEDLINE, especially where a statistical representation or a few examples are desired, the indexer effect is not overly important. However, for LBD, where any prior art/science can negate potential discovery, even one omission can prove lethal!

Thus, an algorithm that operates in MeSH space could predict discovery (where the potential discovery concept from the bc literature could not be found in the MeSH-based core ab literature), whereas the concept could be found in a text-based core ab literature. For this reason, any potential discovery made using a MeSH-based process must be vetted not only in MeSH space *but in text space as well.*

This requirement has enormous consequences! Since each MeSH term effectively represents many text terms, all these text terms have to be considered when vetting a discovery in MeSH space. Thus, *the substantive dimensional advantages that were gained in transforming from text space to MeSH space in the front end are reversed for the vetting process in the back end.* More serious is that these non-indexed or non-properly indexed records are not available for discovery using MeSH alone. To overcome this limitation, some type of text access query would be necessary.

Some examples of reported potential discoveries that were generated in MeSH space but were shown to have prior art in text space are presented in [19, 20]. To illustrate the operational mechanics of our vetting process, I will first describe in some detail one example (of many) reported in [20]. I will then summarize the single example reported in [19].

In [30], the authors generate a potential discovery-identifying algorithm that operates by building MeSH-based profiles from MEDLINE for topics. In [31, 32], the authors start with curcumin (an ingredient of the spice turmeric) and, using their algorithm, look for potential ailments this substance could benefit. Three areas identified are 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).

I will examine the three specific claimed potential discoveries listed above using vetting steps 1 and 3, and show that the claimed discoveries are neither discovery nor innovation. 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 be published prior to November 2003. My 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, I use text terms based on the main MeSH terms used by the author, and initially enter them (initiating topic C literature AND target A literature terms) into the PubMed search engine. This allows me to retrieve MEDLINE articles that contain the initiating topic and target literature MeSH terms and/or text terms. Then, to obtain citing or reference article data, I enter the same terms into the Science Citation Index. Finally, to obtain patent data, I enter 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 I was searching for prior art to support a legal case, I 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 [23] 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. [34] published a meeting Abstract in Gastroentorology in April 2002 and a research article in Gastroentorology in December 2002 [35] 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 Salh et al. [27] and Ukil et al. [41].

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 [24] suggests cervistatin, pyrrolidinedithiocarbamate, or curcumin could equally serve as a treatment for proliferative diabetic retinopathy. Additionally, one of its citing papers [3] focused on the proposed curcumin treatment for diabetic retinopathy. 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 [2].

For the ocular inflammation example, a 2001 paper describes the use of commercially available herbal eye drops (Ophthacare) containing curcumin for a variety of infective, inflammatory and degenerative ophthalmic disorders [4]. This formulation has existed since at least the 1990s, and almost ten clinical/laboratory papers of which I am aware have been published on its evaluation between 1998 and 2002. Finally, the patent by Babish above [2] 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 links curcumin directly with glaucoma [15].

These results should not be surprising. There are over 2,300 papers in Medline 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 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.

What we have presented above is probably the tip of the iceberg. There are obviously other ways to refer to curcumin or Crohn's, and a search using these additional proxy terms would enhance the prior discovery. In sum, we would not call these curcumin links a discovery, or even an innovation, because the links between curcumin and retinal, intestinal, or Multiple Sclerosis problems were established well before November 2003. The algorithm under discussion, with perhaps some modifications, might be a solution for some types of semi-automating literature-based discovery, but it was not demonstrated by the three examples shown.

In [49], the authors used MeSH terms to represent document contents. They divided MEDLINE into two parts: a baseline literature including only publications before 1 January 2004, and a test literature including only publications between 1 January 2004 and 30 September 2005. They ran their algorithm LitLinker on the baseline literature and checked the generated connections in the test literature.

They reported potential discovery for three cases: Alzheimers Disease, Migraine, and Schizophrenia. They provided statistical results for all three cases, and provided one specific example of potential discovery for each of the three cases examined.

Again, I used vetting steps one and three to search the literature for references prior to 1 January 2004. For Alzheimers Disease and Migraine, I found multiple prior references, and for Schizophrenia I found a prior patent. The details are presented in [19]. In neither of the above two cases [30, 49] did I use proxy terms for either the potential discoveries or the diseases; I used only the author's own words/phrases.

Another example is the following [43]. This approach 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 authors provide two examples in [43] of ACS for small sub-sets of the total Medline database ($\ll 1\%$), 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.

Only one of the two examples will be addressed. The authors retrieved a subset of MEDLINE records (13,423 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 author's 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, I 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, I found 13 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 I re-ran the query as above minus the term 'hearing', I found 11 articles. I see no evidence of discovery, or even innovation. The known associations date back to the mid-1970s.

In all three cases [30, 43, 49], the authors would have presented much stronger arguments for their LBD approaches had they vetted in text as well as MeSH space, and presented potential discoveries that did not appear previously in the mainline literature. Or, even if prior art/science did appear as shown, they might have reported it as innovation (if it met the criteria for innovation).

5.2 Disjointness as Sufficient Condition

In the definition of discovery, the issue of disjointness of diverse literatures was addressed as follows: In the LBD 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 LBD. 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.

Most of the LBD techniques link disparate literatures through quantity-based approaches. However, the quality of the linkages for discovery purposes requires expert judgment. The LBD community needs to be very cautious when linking a potential discovery concept from the ab concept source literature to the bc problem literature, especially in the case where there are many researchers reporting on the concept in the ab literature. What are the chances that the bc application was not

perceived by at least one or two of these researchers? If the linkage were promising, why was it not reported?

I will present two examples to illustrate the problem, but they represent the tip of the iceberg for what has been reported as LBD-based discovery. In the first example, where treatments for Huntington Disease were researched, an association rules method was used to show similarities between Huntington Disease and diabetes mellitus, especially in reduced levels of insulin [12, 13]. The authors suggested (as the potential discovery) that insulin treatment might be an interesting drug for Huntington Disease. To understand the reasons for this recommendation better, I examined literatures related to Huntington Disease, diabetes, and insulin. In the Huntington Disease (HD) case, the relationship between insulin and HD should have been obvious to the HD researchers. There were some papers where HD was induced in mice, they developed diabetes, and then insulin was used to treat the diabetes. If insulin had any impact on the HD, surely the researchers would have noticed.

To validate my perceptions, I contacted an expert in Huntington Disease research, and was told that the HD problem is not an insulin deficiency problem as in type 1 diabetes, but rather an insulin release problem as in another form of diabetes. Therefore, there is no reason to expect that administering insulin would treat the HD. The key point here is that if two literatures are disjoint, there may be multiple reasons for their disjointness. 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.

In the second example [48], the researchers searched for discovery in treating cardiac hypertrophy (defined as an increase in the size of myocites that is associated with detrimental effects on aspects of contractile and electrical function in the heart basically heart enlargement due to added physical stress on the heart muscle). Their ranking technique showed the drug chlorpromazine (CPZ) shared many implicit relations with cardiac hypertrophy, and they 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, but there may be a patent that addresses the link, although it covers a wide swath.

To understand the relationship better, I examined the medical literatures on both CPZ and cardiac hypertrophy, and found the following. 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 potential adverse side effects from the use of CPZ, including, but not limited to:

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; 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, simple tachycardia, 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.

Why, then, given this history of adverse side effects, which includes some adverse cardiac side-effects, would one highlight CPZ for cardiac hypertrophy (or any cardiac problem) as a discovery to be pursued for humans? For control of psychotic problems, CPZ may be the lesser of two evils, but does that hold true for control of cardiac problems?

To validate my perceptions, I contacted two experts in cardiac hypertrophy, and was 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. The authors 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 the authors ran some lab experiments showing that CPZ reduced cardiac hypertrophy in mice [48], the relation may reflect a local optimization on cardiac hypertrophy, and a global sub-optimization on overall somatic well-being.

I ran a shortcut LBD analysis combining some of our methods with Arrowsmith, and found a potential discovery applicable to cardiac hypertrophy. Cereal fiber has been shown to increase circulating adiponectin concentrations in diabetic men and women [25, 26]. In turn, adiponectin, an adipocyte-derived protein, has cardioprotective actions (e.g., [Adiponectin receptors] AdipoR1 and AdipoR2 mediate the suppressive effects of full-length and globular adiponectin on ET-1-induced hypertrophy in cultured cardiomyocytes, and AMPK is involved in signal transduction through these receptors). AdipoR1 and AdipoR2 might play a role in the pathogenesis of ET-1-related cardiomyocyte hypertrophy after myocardial infarction, or adiponectin deficiency leads to progressive cardiac remodeling in pressure overloaded condition mediated via lowing AMPK signaling and impaired glucose metabolism [22]. Therefore, use of cereal fiber in the diet could potentially contribute to ameliorating cardiac hypertrophy, with probably very few or no adverse side effects, and perhaps some positive side effects.

Since our previous LBD studies have generated voluminous amounts (hundreds) of potential discovery on each disease studied, I see no reason this would be different for cardiac hypertrophy, and the single potential discovery presented here would be one of very many resulting from a full study.

5.3 Definition of Prior Art

This third validation category brings us back full circle to the definition of discovery and what is prior art. As an example, many LBD studies refer to the potential discovery of Fish Oil for Raynaud's Disease [9, 10, 12, 36, 45]. Use of Fish Oil for circulatory problems was reported in the literature at least as far back as the 1970s, and possibly even earlier. Papers in the late 1970s discussed the impact of Fish Oil on atherosclerosis [1], thrombosis [8], vascular disease [14], and papers in the early 1980s also focused on vascular disease [7] and peripheral vascular disease [47]. While none of these papers mentioned Raynaud's Disease specifically, how much of a leap is it from peripheral vascular disease to Raynaud's Disease? For example, [28] 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 I examined list Raynaud's Disease under peripheral vascular diseases. Thus, depending on how broadly the core Raynaud's Disease literature is defined, Fish Oil may or may not have been a potential discovery.

6 Summary and Conclusions

In summary, this chapter has shown the importance of having rigorous definitions of discovery and innovation, and using a rigorous vetting process to insure that no prior art exists. While one can always identify further sources that could be checked for prior art, nevertheless, the sources suggested in this chapter should be viewed as a threshold before reporting potential discovery in the literature, I firmly believe that one of the major roadblocks to wide-scale acceptance of LBD by the potential user community has been the lack of real discovery reported in the literature. Until more rigorous standards for defining discovery have been implemented, and more rigorous vetting techniques used, LBD will have problems in taking its rightful place in the arsenal of discovery weapons.

References

1. F. Angelico and P. Amodeo. Eicosapentaenoic acid and prevention of atherosclerosis. *Lancet*, 2(8088):531, 1978
2. J. G. Babish, T. Howell, L. Pacioretty, T. M. Howell, and L. M. Pacioretty. Composition for treating e.g. inflammation or inflammation based diseases, comprising curcuminoid species and alpha- or beta-acid. Patent Number US2003096027-A1, May 22, 2003

3. M. Balasubramanyam, A. Koteswari, R. S. Kumar, S. F. Monickaraj, J. U. Maheswari, and V. Mohan. Curcumin-induced inhibition of cellular reactive oxygen species generation: novel therapeutic implications. *Journal of Bioscience*, 28(6):715–721, 2003
4. N. R. Biswas, S. K. Gupta, G. K. Das, N. Kumar, P. K. Mongre, D. Haldar, and S. Beri. Evaluation of ophthacare eye drops – a herbal formulation in the management of various ophthalmic disorders. *Phytotherapy Research*, 15(7):618–620, 2001
5. P. Bruza, R. Cole, D. W. Song, and Z. Bari. Towards operational abduction from a cognitive perspective. *Logic Journal of the IGPL*, 14(2):161–177, 2006
6. P. Bruza, D. W. Song, and R. McArthur. Abduction in semantic space: towards a logic of discovery. *Logic Journal of the IGPL*, 12(2):97–109, 2004
7. I. J. Cartwright, A. G. Pockley, and J. H. Galloway. The effects of dietary omega-3 polyunsaturated fatty-acids on erythrocyte-membrane phospholipids, erythrocyte deformability and blood-viscosity in healthy-volunteers. *Atherosclerosis*, 55(3):267–281, 1985
8. J. Dyerberg, H. O. Bang, E. Stoffersen, S. Moncada, and J. R. Vane. Eicosapentanoic acid and prevention of thrombosis and atherosclerosis. *Lancet*, 2(8081):117–119, 1978
9. M. D. Gordon and S. Dumais. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science and Technology*, 49(8):674–685, 1998
10. M. D. Gordon and R. K. Lindsay. 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 and Technology*, 47(2):116–128, 1996
11. P. Healey, H. Rothman, and P. K. Hoch. An experiment in science mapping for research planning. *Research Policy*, 15(5):233–251, 1986
12. D. Hristovski, B. Peterlin, and S. Dzeroski. Literature based discovery support system and its application to disease gene identification. In *Proceedings of AMIA Fall Symposium*, p. 928. Hanley and Belfus, Philadelphia, PA, 2001
13. D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2–4): 289–298, 2005
14. J. A. Jakubowski and N. G. Ardlie. 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, 1979
15. A. Komatsu. 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 JP2003189819-A, July 8, 2003
16. R. N. Kostoff. Overcoming specialization. *BioScience*, 52(10):937–941, 2002
17. R. N. Kostoff. Stimulating innovation. In L. V. Shavinina, editor, *International Handbook of Innovation*, pp. 388–400. Elsevier Social and Behavioral Sciences, Oxford, UK, 2003
18. R. N. Kostoff. Systematic acceleration of radical discovery and innovation in science and technology. *Technological Forecasting and Social Change*, 73(8):923–936, 2006
19. R. N. Kostoff. Validation of potential literature-based discovery candidates. *Journal of Biomedical Informatics*, 40(4):448–450, 2007
20. R. N. Kostoff, J. A. Block, M. B. Briggs, R. L. Rushenberg, J. A. Stump, D. Johnson, C. M. Arndt, T. J. Lyons, and J. R. Wyatt. Literature-related discovery. *ARIST*, 2008
21. P. Lax and B. Wendroff. Systems of conservation laws. *Communications on Pure and Applied Mathematics*, 13(2):217–237, 1960
22. Y. Liao, S. Takashima, N. Maeda, N. Ouchi, K. Komamura, I. Shimomura, M. Hori, Y. Matsuzawa, T. Funahashi, and M. Kitakaze. Exacerbation of heart failure in adiponectin-deficient mice due to impaired regulation of ampk and glucose metabolism. *Cardiovascular Research*, 67(4):705–713, 2005
23. C. Natarajan and J. J. Bright. 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, 2002

24. T. Okamoto, S. Yamagishi, Y. Inagaki, S. Amano, K. Koga, R. Abe, M. Takeuchi, S. Ohno, A. Yoshimura, and Z. Makita. Angiogenesis induced by advanced glycation end products and its prevention by cerivastatin. *The FASEB Journal*, 16(14):1928–1930, 2002
25. L. Qi, J. B. Meigs, S. Liu, J. E. Manson, C. Mantzoros, and F. B. Hu. Dietary fibers and glycemic load, obesity, and plasma adiponectin levels in women with type 2 diabetes. *Diabetes Care*, 29(7):1501–1505, 2006
26. L. Qi, E. Rimm, S. Liu, N. Rifai, and F. B. Hu. Dietary glycemic index, glycemic load, cereal fiber, and plasma adiponectin concentration in diabetic men. *Diabetes Care*, 28(5):1022–1028, 2005
27. B. Salh, K. Assi, V. Templeman, K. Parhar, D. Owen, A. Gomez-Munoz, and K. Jacobson. Curcumin attenuates DNB-induced murine colitis. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 285(1):G235–G243, 2003 [see also [44]]
28. SIGN. Drug therapy for peripheral vascular disease: a national clinical guideline. Technical Report IGN Publication Number 27, Scottish Intercollegiate Guidelines Network, Edinburgh, Scotland, 1998
29. N. R. Smalheiser and D. R. Swanson. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149–153, 1998
30. Padmini Srinivasan. Text mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004
31. Padmini Srinivasan and Bishara Libbus. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20(suppl 1):I290–I296, 2004
32. Padmini Srinivasan, Bishara Libbus, and A. K. Sehgal. Mining MEDLINE: postulating a beneficial role for curcumin longa in retinal diseases. *HLT BioLink*, 20(suppl 1):I290–I296, 2004
33. J. Stegmann and G. Grohmann. Hypothesis generation guided by co-word clustering. *Scientometrics*, 56(1):111–135, 2003
34. K. Sugimoto, H. Hanai, T. Aoshi, K. Tozawa, M. Uchijima, T. Nagata, and Y. Koide. Curcumin ameliorates trinitrobenzene sulfuric acid (TNBS) – induced colitis in mice. *Gastroenterology*, 122(4 suppl 1):A395–A396, T993, 2002
35. K. Sugimoto, H. Hanai, T. Aoshi, K. Tozawa, M. Uchijima, T. Nagata, and Y. Koide. Curcumin prevents and ameliorates trinitrobenzene sulfonic acid-induced colitis in mice. *Gastroenterology*, 123(6):1912–1922, 2002
36. D. R. Swanson. Undiscovered public knowledge. *Library Quarterly*, 56:103–118, 1986
37. D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988
38. D. R. Swanson. Somatomedin-c and arginine – implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33(2):157–186, 1990
39. D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2), 1997
40. D. R. Swanson, N. R. Smalheiser, and A. Bookstein. Information discovery from complementary literatures: categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology*, 52(10):797–812, 2001
41. A. Ukil, S. Maity, S. Karmakar, N. Datta, J. R. Vedasiromoni, and P. K. Das. 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, 2003
42. R. E. Valdes-Perez. Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence*, 107(2):335–346, 1999
43. C. C. van der Eijk, E. M. van Mulligen, J. A. Kors, B. Mons, and J. van den Berg. Constructing an associative concept space for literature-based discovery. *Journal of the American Society for Information Science and Technology*, 55(5):436–444, 2004
44. C. Varga, M. Cavicchi, A. Orsi, D. Lamarque, J. C. Delchier, D. Rees, and B. J. Whittle. Beneficial effect of P54, a novel curcumin preparation in TNBS-induced colitis in rats. *Gastroenterology*, 120(5 suppl 1):A691, 2001

45. M. Weeber, H. Klein, L. T. W. de Jong-van den Berg, and R. Vos. 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, 2001
46. M. Weeber, R. Vos, H. Klein, L. T. W. de Jong-van den Berg, A. R. Aronson, and G. Molema. 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, 2003
47. B. E. Woodcock, E. Smith, and W. H. Lambert. Beneficial effect of fish oil on blood-viscosity in peripheral vascular-disease. *British Medical Journal*, 288(6417):592–594, 1984
48. J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 20(3):389–398, 2004
49. M. Yetisgen-Yildiz and W. Pratt. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*, 39(6):600–611, 2006