Implicit Text Linkages between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery

DON R. SWANSON AND NEIL R. SMALHEISER

ABSTRACT

THE PROBLEM OF HOW TO FIND INTERESTING but previously unknown implicit information within the scientific literature is addressed. Useful information can go unnoticed by anyone, even its creators, if it can be inferred only by considering together two (or more) separate articles neither of which cites the other and which have no authors in common. The two articles (or two sets of articles) are in that case said to be complementary and noninteractive. During the past twelve years, this project has uncovered and reported numerous complementary relationships in the biomedical literature that have led to new information of scientific interest. Several of these literature-based discoveries subsequently have been corroborated through clinical or laboratory investigations. We describe how to use software that can create suggestive juxtapositions of Medline records, the purpose being to help biomedical researchers detect new and useful relationships. This software, called Arrowsmith, has also proved valuable as a tool for investigating patterns of complementary relationships in natural language text (Arrowsmith can be used free of charge at http://kiwi.uchicago.edu).

INTRODUCTION

The juxtaposition of certain natural language text passages from different biomedical journal articles can reveal or suggest new information not contained in the original passages considered separately. For example,

Don R. Swanson, Division of the Humanities, The University of Chicago, 1010 E. 59th St., Chicago, IL 60637
Neil R. Smalheiser, Department of Psychiatry, University of Illinois, 1601 W. Taylor St., Chicago, IL 60612
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 48-59

one article might report an association or link between substance A and some physiological parameter or property B while another reports a relationship between B and disease C. If nothing has been published concerning a link between A and C via B, then to bring together the separate articles on A-B and B-C may suggest a novel A-C relationship of scientific interest. There are now about 9 million records in the Medline database, and hence about 40 trillion (40,000,000,000,000) possible pairings of records. Clearly the vast majority of record pairs and article pairs have never been considered together. It is plausible to think that there are many undiscovered implicit relationships within the biomedical literature, at least some of which might be important (Swanson, 1993, pp. 611-19). It is important, therefore, to develop systematic methods for finding them.

The possibility of literature-based discovery implied by the above model underscores two important properties of sets of scientific articles—complementarity and noninteractivity. Two sets of articles are defined here as complementary if together they can reveal useful information not apparent in the two sets considered separately; two sets are defined as noninteractive if they are disjoint and if no article in either set cites, or is co-cited with, any member of the other set (Swanson, 1987, 1990a, 1991).

The first three examples of "undiscovered public knowledge" (Swanson, 1986a, 1986b, 1988, 1990c) demonstrated that complementary noninteractive structures actually do exist within the biomedical literature and can lead to the discovery of apparently new and interesting implicit relationships. In at least two of these cases (Swanson, 1986a, 1988) the hypothesis was subsequently corroborated experimentally by medical researchers. We have cited and discussed these corroborations elsewhere (Swanson, 1993; Smalheiser & Swanson, 1994). The hypothesis advanced in Swanson (1990c)—that the anabolic effects of arginine are brought about by systemic or local release of somatomedin C—has also received direct supporting evidence in three recent studies (see Kirk, 1993; Hurson, 1995; Chevalley, 1998); a fourth study by Corpas (1993) reported negative results. Gordon and Lindsay (1996) re-examined, replicated, and extended Swanson's work (1986a).

The above structures were found through innovative, partially systematic, database search strategies (Swanson, 1989a, 1989b). Computer-assisted processing of the downloaded output enhanced the user's ability to discover novel implicit relationships (Swanson, 1991). This software evolved into a system called Arrowsmith that processes article records downloaded from large bibliographic databases such as Medline. Text passages within database records provide the raw material that suggests or points to underlying linkages (such as A-B and B-C above) between separately published scientific findings or arguments. Our goal has been to create a research tool for studying complementary noninteractive structures in the scientific literature and at the same time to create a working system useful to biomedical scientists (Swanson, 1991; Swanson & Smalheiser, 1997; Smalheiser & Swanson, 1998b).

With the help of Arrowsmith, we have developed five additional examples of complementary noninteractive literature structures (Swanson & Smalheiser, 1997; Smalheiser & Swanson, 1994, 1996a, 1996b, 1998a), each of which led to a novel, plausible, and testable medical hypothesis. One of these studies (Smalheiser & Swanson, 1998a) elicited publication of a concurring letter from an author whose work was the basis for a new hypothesis that we proposed (Ross, 1998).

The Process of Inferring Text Linkages

Given two Medline titles that appear to be linked, the process of inferring a biologically meaningful linkage may be more subtle than it seems at first sight. We consider here examples taken from Swanson (1988):

1. "The Relation of Migraine and Epilepsy" (p. 551)
2. "Preliminary Report: The Magnesium-Deficient Rat as a Model of Epilepsy" (p. 556).

The two titles taken together appear to provide a link, via epilepsy, between migraine and magnesium deficiency (epilepsy being just one of the eleven links reported). The role of Arrowsmith in this example is only to bring the two titles together in order to create a suggestive juxtaposition. Whether the relationship thus revealed might merit further investigation then depends on human judgment. Such judgment in general would be difficult to replace by a computer procedure, for it almost inevitably entails certain background knowledge, context, and presuppositions that are commonly, though perhaps not always consciously, brought to bear by the user. For example, the word "model" in the second title is understood against a substantive background of information about animal models of human disease, and in that context implies that magnesium deficiency causes a disorder resembling epilepsy in the rat. Several hundred analogous title pairs were examined in the course of the migraine-magnesium study, for most of which the linkage was less obvious than in the case above. The user often must make just an educated guess as to which leads are most promising (Swanson, 1991).

The problem we identify in this example therefore is not how or whether to draw an inference about the possible effect of magnesium on migraine, given the above two titles, but rather how these two titles (or Medline records), and other pairs analogous to them, could have been found and brought together in the first place without knowing in advance about any specific link such as epilepsy. That task cannot be done using only a conventional Medline search. However, if one first uses Medline to form a local file consisting of all titles with "migraine," and a second file that consists of all titles with "magnesium," then a straightforward computer procedure can produce a list of all words common to the two sets of titles. "Epilepsy" would be on the list. One can think of this procedure, which Arrowsmith takes as its point of departure, as a "higher order Medline search." Arrowsmith then automatically filters out noninteresting words (by means of an exclusion list, or stoplist, compiled in advance and built into the system), makes certain morphological transformations (such as plural to singular), constructs and matches phrases, and otherwise exploits information from the Medline record to juxtapose pairs of text passages for the user to consider as possibly complementary. (Arrowsmith can process abstracts as well as titles but, for files of more than 1,000 or so records, it is more efficient and more effective to search, download, and subsequently examine just titles. The restricted context makes it easy to see and assess the A-B relationships when both A and B are in a title and similarly for B-C.) Any inferences about the significance or nature of the linkage between the above two titles, once they have been brought together, are left to the user. Arrowsmith, by creating suggestive juxtapositions of database records, is an aid to scientific discovery but not in itself a mechanism of scientific discovery.

Automatic Generation of a Candidate List for A

Arrowsmith can also do more than help uncover linkages between an initially given A and C. Assume that at the outset only C, the disease under investigation, is given, and the user does not have in mind a specific hypothesis for A (an agent that might act as cause or cure). Then, instead of a specific A, a broad category (AA) may be chosen; such a choice can be simple and effective. In general, categories of exogenous substances that may enter the body and might conceivably have beneficial or adverse effects on C are of interest. Especially important are dietary factors (or deficiencies), toxins, and categories of pharmaceutical agents or their targets (Swanson, 1991). Arrowsmith can then begin with Medline files for C and AA and from these derive a list of specific candidates for A. For example, Arrowsmith was able to start with pre-1988 literature on "migraine" as C, use a category based on dietary or deficiency factors (AA), and produce "magnesium" as a top-ranking candidate for A (Swanson, 1991; Swanson & Smalheiser, 1997).

Direct A-C Search as First Step

It is important for the user who wishes to investigate indirect or implicit connections between A and C to understand that the first step—prior to using Arrowsmith—is to find all articles that are explicitly about both A AND C by means of a conventional or "direct" Medline search. Insofar as indirect linkages are already known (i.e., published), one would expect to find a discussion of them in articles belonging to the A-C intersection. Failure to understand the contents of the A-C intersection may result in failure to distinguish new from old in the Arrowsmith output.

To conduct a good direct search, some skill and experience with Medline searching is required and in particular familiarity with the medical subject heading (MeSH) hierarchical structure, the superimposed subheading structure, and the organization of the Medline record. Searching of other major biomedical databases, including BIOSIS, EMBASE, and the Science Citation Index, is also important. In some cases, the existence of a sizable direct literature does not necessarily imply that the A and C literatures are well-integrated. For example, in our study of magnesium in the central nervous system, we found a substantial direct literature. But a citation analysis revealed a highly fragmented structure, not at all characteristic of researchers investigating a common problem who cite each other, and are co-cited, extensively (Smalheiser & Swanson, 1994, pp. 5-8). In other cases, we encountered small direct literatures that have never been cited at all in one or the other of the A or C literatures, indicating that new connections were published but then ignored. Our experience underscores the importance of conventional database searching and citation analysis prior to using Arrowsmith for a literature-synthesis study.

In any event, in the more straightforward case in which a well-constructed direct search turns up little or nothing in any of the major appropriate databases, a conventional database search cannot then go any further toward discovering unknown indirect links such as epilepsy in the above example. Arrowsmith is designed to solve that problem. We next explain what Arrowsmith does and how to use it on the Internet.

Arrowsmith on the Web

Arrowsmith may be used free of charge at the Web site: http://kiwi.uchicago.edu. The input to Arrowsmith consists of two files that the user first creates by searching Medline and downloading the resulting records to the user's local computer. We refer to these two local files as File A and File C, both of which must then be transmitted to the server kiwi.uchicago in order to be processed by Arrowsmith. Uploading local files to a remote server can be implemented using Netscape.

Preparing the Input Files

The user begins with some problem (which may be a medical disorder of unknown cause, such as migraine) and conducts a Medline search for records about that disorder (a title-word search is preferable for large files), then downloads the resulting records or titles to a local File C. Similarly, a second Medline search creates a target literature, A (such as magnesium), or some broader category (AA), that is downloaded to File A.

The intersection A AND C is presumed to have been investigated beforehand as noted above. The Arrowsmith software operates in five stages. The user normally will exit after each stage and reconnect at a later stage when results are ready (e-mail addresses are used to identify individual files and results).

Stage 1: Transmitting the Two Input Files to the Server kiwi.uchicago

The kiwi Web site is designed to accept large files transmitted by Netscape. The user provides the local pathname/filename. After Arrowsmith receives File C and File A, it creates a list of all "important" words and phrases common to the two files. This list of terms provides the source for intermediate linkages (B) between A and C. The distinction between words that are "important" and words that are not is implemented by means of a large stoplist (words to be excluded) compiled in advance by applying human judgment and then built into Arrowsmith for all applications. Certain variant word forms are also matched. The output of this stage is a preliminary list of B-terms made available to the user (at Stage 2) five to thirty minutes (depending on file sizes) after Files C and A are received.

Stage 2: Editing the B-List

The preliminary B-list may contain several hundred terms and should be edited by the user. Notwithstanding the stoplist filter, the B-list often contains many terms that the user would not consider of potential interest as linkages in light of the particular problem at hand. At the Web site, the preliminary B-list appears in a scrollable "option" window that permits multiple selection of terms. The selected (highlighted) terms are then automatically deleted from the B-list.

Stage 3: Organized Display of Medline Records as Output

The edited B-list is displayed in a window in which each B-term is a pointer to the subset of Medline records from File A containing that term, a subset called the "AB" records. Each AB display contains a pointer to the corresponding set of BC records, thus facilitating a systematic, organized process of point-and-click browsing of Medline records. For each B-term, the corresponding AB records are, in effect, juxtaposed with BC records to help the user notice a possible A-C relationship. Successful use of Arrowsmith depends on the user's subject knowledge, ingenuity, and ability to see promising connections suggested by comparing AB records with BC records for each B, as illustrated earlier in comparing a magnesium-epilepsy title with an epilepsy-migraine title.

An online example of Arrowsmith title-browsing has been prepared as an interactive demonstration (dem2) at Stage 3 of the kiwi Web site. The example is based on 2,800 migraine titles and 8,000 magnesium titles (all pre-1988, the time frame of the original study [Swanson, 1988]). The computer-produced B-list consisted of 260 terms and was edited manually to about 100 terms. The user may click on any term in the B-list to see the corresponding magnesium titles, then click on BC to see the migraine titles for that same B-term. The next two stages show what can be done if the user had not considered magnesium at the outset as a possible solution to the migraine problem.

Stage 4: Ranking Individual "A" Terms

Stages 4 and 5 do not apply if File A, above, was based on a specific substance (such as magnesium). However, if File A was created by searching a broader category (such as dietary substances), then we refer to it here as File AA and it becomes of interest to identify more specific A-terms that occur in the records within the AA category. Arrowsmith derives, from the AAB records, a list of words and phrases that become candidates for these more specific terms. The list of candidates is called the A-list. Each term on the A-list is associated with all B-terms that co-occur with it in the AAB records.

The A-list terms are then ranked by the number of their associated terms from the B-list. This method is a simplified version of the ranking method discussed in Swanson and Smalheiser (1997). Thus, the output of Stage 4 is a (preliminary) ranked A-list. Returning to our example using "migraine" to create File C and a dietary/deficiency category to create File AA, the word "magnesium" appeared at the top of the resulting A-list.

Stage 5: Editing and Grouping Terms on the A-List

As was the case for the B-list, the A-list may contain many terms of no interest that should be manually deleted, and it may contain synonyms or related terms that should be grouped together for purposes of ranking. Stage 5 presents the A-list within a scrollable option window that permits multiple selection. Two modes of operation are offered—a deletion mode and a grouping mode. In the first mode, all terms selected are deleted just as in Stage 2. In the second mode, all terms selected by the user are grouped together and treated as synonymous for the purpose of ranking. For example, the A-list might contain ascorbate, ascorbic acid, and vitamin C. In one pass through the window, clicking on these three terms will create a group in which all associated B-terms from each of the three are combined into a single new total; repeating the ranking procedure then gives the group a higher rank than any of its component A-terms. Or the user may choose to form a broader grouping such as all terms that refer to antioxidants, which would include the vitamin C terms above. Alternation between the deletion mode and the grouping mode is permitted using each mode as many times as desired. The final A-list is then reranked.

Nothing in the foregoing process determines whether any term on the A-list does or does not co-occur directly with C in Medline records;

such co-occurrence should be separately determined by means of a conventional Medline search. Extensive co-occurrence probably indicates that the relationship with C is already well known, and so the A-term in question may not be of further interest (however, see the earlier discussion of the direct search and the possibility of encountering fragmented structures).

The sole purpose of the A-list is to offer some automatically generated promising choices of specific A-terms for the user's consideration. Once the user has chosen a single specific A (such as magnesium) that seems promising as the basis for File A, then the next step is to re-run Arrowsmith beginning again at Stage 1. The category restriction may be omitted altogether (thus leading to the largest B-list for the A and C under consideration) or it (or perhaps a revised version) may be included as part of the Medline search that creates File A.

Synonym Recognition and the Role of Medical Subject Headings (MeSH)

The heart of Arrowsmith is the computerized process of finding and matching words and phrases that occur in both input files (Files A, C) as an approach to helping the user identify complementary passages of text from titles or abstracts. In addition to matching identical terms, Arrowsmith also matches certain morphological variants, including most cases of singular versus plural, and it can identify synonyms insofar as they are indexed by a common subject heading (MeSH). To take advantage of the synonym matching capability, MeSH terms must be included for each record in the input files A and C.

The output of the matching process consists of a list of terms (the B-list) that itself may contain synonyms or context-dependent equivalencies that the user may wish to take into account. A future version of Arrowsmith will provide more assistance by presenting to the user a list of word (and phrase) pairs that are candidates for synonyms or "surrogate synonyms" (sometimes called "searchonyms") that could serve as an aid to editing (Stage 2, 5), browsing (Stage 3), and forming groups (Stage 5). Words will be paired if they tend to appear in similar contexts as defined with the help of statistics based on second order co-occurrence data. Two words that are synonymous or equivalent tend not to co-occur in a highly restricted context such as a title and so do not have a strong first order title co-occurrence correlation. But their tendency to occur in similar contexts gives rise to relatively stronger second order title co-occurrence correlation.

Synonyms, searchonyms, variant word forms, and co-occurrence statistics can at best provide only a partial solution to the difficult problems of detecting complementary or suggestive pairs of text passages, but Arrowsmith is especially valuable for developing and testing improved approaches and techniques.

PATTERNS OF COMPLEMENTARITY AND SUGGESTIVITY

"A causes B, B causes C; hence A causes C" can be taken as a paradigm for complementarity, but it is an idealization. As we have gained experience using Arrowsmith, it has become clear that transitivity is almost never assured, and we have to settle for the less formal and less tidy idea of suggestibility (Swanson, 1991). The problems of suggestivity and complementarity as expressed in natural language text are complex and subtle. Nonetheless, Arrowsmith is now able to produce large numbers of suggestive juxtapositions of Medline titles or records, and it is reasonable to expect further improvement with the accumulation of additional inelegant ad hoc empirical rules with little else to recommend them except that they seem to work.

In studying links that actually occur in the natural language text of title words and phrases, we have identified a few regularities or patterns that may become the basis for useful rules. For example, the A-B and B-C relationships largely fall into three groups that can be called "influence," "similarity," and "focus." The concept of "influence" (of A on B or B on C) can be expressed by many different words, including: increases, decreases, attenuates, reduces, promotes, inhibits, ameliorates, exacerbates, enhances, causes, accelerates, facilitates, triggers, catalyzes, competes with, interferes with, or acts synergistically. The direction of influence may also be reversed, with B influencing A. The software is indifferent and symmetric with respect to the direction of any relationship. The concept of "similarity" can be important either alone (A is similar to B and B is similar to C) or in conjunction with "influence": A influences B and B is similar to C, thus suggesting that A might influence C (e.g., magnesium deficiency triggers or exacerbates epileptiform seizures; migraine in some respects is similar to epilepsy, suggesting therefore that magnesium deficiency may trigger or exacerbate migraine attacks). The category "focus" refers loosely to a cluster of relationships between some disease and its manifestations in specific cell types, processes, mechanisms, pathways, markers, and organs, or at any anatomic locale at which a focal pathology is a characteristic feature. "A" may be a drug or other substance that is active at such a focus, B, in which case the "influence" category probably applies. The relationship between a disease and its manifestations (e.g., pathologic markers for it) may be more difficult to categorize, so "focus" is used simply as a tentative collective name for possibly several types of relationship.

For example, indomethacin inhibits a variety of cholinergic responses; cholinergic deficits are characteristic of Alzheimer's disease. (Thus indomethacin, which is thought to have a protective effect in Alzheimer patients on the basis of clinical trials, might also have unexpected adverse effects [Smalheiser & Swanson, 1996a].)

The foregoing regularities notwithstanding, natural language is richly expressive, and the variety of ways in which meaningful biological linkages can be suggested to the expert human observer may be so great as to defeat any attempt to formalize and automate the recognition and inference process. Arrowsmith in its present form does not attempt to do so but instead is designed to organize and display records so as to facilitate human recognition of implicit connections. Investigating patterns of complementarity, however, may lead to richer and improved displays of information to the user and so perhaps to improved stimulation of hypotheses. Arrowsmith is not only a practical tool that can aid the biomedical researcher, it is also a research tool for investigating the problems of finding and identifying natural language text linkages.

The Role of Human Intelligence

At several points in the procedure, Arrowsmith receives a boost from human input that helps it perform as if it were intelligent. The first boost is the choice of the problem and its literature C, plus the choice of A as a specific target, or AA as a more general target category. Using A, AA, and C to construct a good Medline search also requires knowledge, experience, and judgment at the outset. The second boost is the stoplist filter, which greatly reduces the number of useless connections that otherwise would clutter the output. The stoplist is compiled using human judgment (and guesswork) concerning which words probably could not play any useful role in forming biologically meaningful and helpful linkages. It is intended as a one-time compilation, not ad hoc for each Arrowsmith application, but the stoplist does grow as the human compiler gains experience with Arrowsmith and now includes about 7,000 words. The remaining boosts come from the user in editing the B-list and A-list and in forming groups within the A-list. Finally, given the juxtaposed AB-BC titles or abstracts, any identification of promising implicit linkages of biological importance depends on the knowledge and perspicacity of the user.

Assessments of Arrowsmith by Others

This project has been analyzed, enhanced, and extended in a number of recent papers (Chen, 1993; Cory, 1998; Davies, 1989; Finn, 1998; Garfield, 1994; Gordon & Lindsay, 1996; Gordon & Dumais, 1998; Kostoff, 1998; Rikken, 1998; Spasser, 1997). Analogous work on computer-generated discovery in chemical reaction pathways has also been reported (Valdes-Perez 1994). Valdes-Perez (1999) has assessed four successful computer-assisted discovery programs in chemistry (MECHEM), medicine (ARROWSMITH), mathematics (GRAFFITI), and linguistics (MPD/KINSHIP). He explains why he believes that each of them has produced results that are novel, interesting, plausible, and intelligible.

REFERENCES

Chen, Z. (1993). Let documents talk to each other: A computer model for connection of short documents. *Journal of Documentation, 49*(1), 44-54.

Chevalley, T. H.; Rizzoli, R.; Manen, D.; Caverzasio, J.; Bonjour, J. -P. (1998). Arginine increases insulin-like growth factor-I production and collagen synthesis in osteoblast-like cells. *Bone, 23*(2), 103-109.

Corpas, E.; Harman, S. M.; Blackman, M. R. (1993). Human growth hormone and human aging. *Endocrine Reviews, 14*(1), 20-39.

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

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation, 45*(4), 273-301.

Finn, R. (1998). Program uncovers hidden connections in the literature. *Scientist, 12*(10), 12-13.

Garfield, E. (1994). Linking literatures: An intriguing use of the *Citation Index. Current Contents, 21*, 3-5.

Gordon, M. D., & Lindsay, R. K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science, 47*(2), 116-128.

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

Hurson, M.; Regan, M. C.; Kirk, S. J.; Wasserkrug, B. A.; & Barbul, A. (1995). Metabolic effects of arginine in a healthy elderly population. *Journal of Parenteral and Enteral Nutrition, 19*(3), 227-230.

Kirk, S. J.; Hurson, M.; Regan, M. C.; Holt, D. R.; Wasserkrug, H. L.; & Barbul, A. (1993). Arginine stimulates wound healing and immune function in elderly human beings. *Surgery, 114(2)*, 155-160.

Kostoff, R. N. (1998). *Science and technology innovation.* Retrieved January 1, 1999 from the World Wide Web: http://www.dtic.mil/dtic/kostoff/Swanson2.txt.

Rikken, F. (1998). *Adverse drug reactions in a different context: A scientometric approach towards adverse drug reactions as a trigger for the development of new drugs.* Unpublished doctoral dissertation, Rijksuniversiteit Groningen.

Ross, B. M. (1998). In reply. *Archives of General Psychiatry, 55*(8), 753.

Spasser, M. A. (1997). The enacted fate of undiscovered public knowledge. *Journal of the American Society for Information Science, 48*(8), 707-717.

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

Smalheiser, N. R., & Swanson, D. R. (1996a). Indomethacin and Alzheimer's Disease. *Neurology, 46*(2), 583.

Smalheiser, N. R., & Swanson, D. R. (1996b). Linking Estrogen to Alzheimer's Disease: An informatics approach. *Neurology, 47*(3), 809-810.

Smalheiser, N. R., & Swanson, D. R. (1998a). Calcium-independent phospholipase A-sub-2 and schizophrenia. *Archives of General Psychiatry, 55*(8), 752-753.

Smalheiser, N. R., & Swanson, D. R. (1998b). Using Arrowsmith: A computer-assisted approach to forming and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine, 57*(3), 149-153.

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

Swanson, D. R. (1986b). Undiscovered public knowledge. *Library Quarterly, 56*(2),103-118.

Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science, 38*(4), 228-233.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine, 31*(4), 526-557.

Swanson, D. R. (1989a). Online search for logically-related noninteractive medical literatures: A systematic trial-and-error strategy. *Journal of the American Society for Information Science, 40*(5), 356-358.

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

Swanson, D. R. (1990a). The absence of co-citation as a clue to undiscovered causal connections. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 129-137). Newbury Park, CA: Sage.

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

Swanson, D. R. (1990c). Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine, 33*(2), 157-186.

Swanson, D. R. (1991). Complementary structures in disjoint science literatures. In A. Bookstein (Ed.), *SIGIR '91* (Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, Illinois, USA, October 13-16, 1991) (pp. 280-289). New York: Association for Computing Machinery.

Swanson, D. R. (1993). Intervening in the life cycles of scientific knowledge, *Library Trends 41*(4), 606-631.

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

Valdes-Perez, R. E. (1994). Conjecturing hidden entities by means of simplicity and conservation laws: Machine discovery in chemistry. *Artificial Intelligence, 65*(2), 247-280.

Valdes-Perez, R. E. (1999). Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence, 107*(2), 335-346.