Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 40 (2007) 448–452

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Letters to the Editor

Validating discovery in literature-based discovery

1. Introduction

A recent JBI article on literature-based discovery [1] claims to identify many tens of medically related potential 'discoveries' for three specific medical problems (Alzheimer's, Migraine, Schizophrenia) using a semi-automated software approach. The article presents three specific 'discovery' examples, and statistics for the remainder. In the literature-based discovery context, a discovery represents the linking of two or more concepts that had never previously been linked in order to produce novel, interesting, plausible, and intelligible knowledge, and an innovation represents the exploitation or implementation of two or more concepts that had been linked previously but were not being exploited. By showing that the concepts had been linked previously in all three cases (in major widely available literatures), this letter will demonstrate that the three specific claimed discoveries are neither discoveries nor innovations.

2. Background

The authors [1] present a literature-based discovery system called LitLinker that incorporates knowledge-based methodologies with a statistical method. The literature-based discovery begins with a starting term (e.g., Alzheimer's Disease), then uses a text mining process to find a set of linking terms (e.g., Neurotransmitters) that are directly correlated with the starting term, and uses the same text mining process to identify a set of target terms that are correlated with each linking term (e.g., Endocannabinoids). 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 (i.e., the Medline taxonomy) as the representation of the content of the documents and performs searches on them to collect the literatures. To find correlations, LitLinker calculates 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 distances 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.

The authors used the following 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 medical problems: 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 (T_i union G_i)/T_i, and recall for starting term i is defined as the ratio of (T_i union G_i)/G_i, where T_i is the set of target terms generated by LitLinker for the starting term i, and G_i 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 show precision increasing with time to about .06 for Alzheimer's Disease, .025 for Migraine, and .075 for Schizophrenia. Recall-time graphs show recall oscillating with time, with approximate mean values of about .22 for Alzheimer's Disease, .43 for Migraine, and .14 for Schizophrenia. These appear on the surface to be quite reasonable results, if in fact the target terms in the gold standard are true potential discoveries. The authors provide one specific example of potential discovery for each of the three diseases examined.

3. Analysis

I will examine the three specific 'discoveries' listed. The first example listed (for Alzheimer's Disease) was the MESH term *endocannibinoids*, and the authors referenced a recent paper by a Spanish group [2] about the possible role of the *endocannibinoid* system in Alzheimer's Disease. The authors state: "Although there has been no prior published work [before 2004-RNK] on the potential connection between *endocannibinoids* and Alzheimer disease, LitLinker could identify it by analyzing existing connections in the medical literature".

One of the references in [2] is a published work from 2003 that relates endocannabinoids to Alzheimer's Disease

[3]. Additionally, entering the query "endocannabinoids and Alzheimer's" in Medline yields three additional papers published prior to 2004 [4–6]). Both terms were in the Abstract, but not necessarily in the MESH terms. There is a well-known effect in information retrieval called the indexer effect (e.g. [7]), whereby errors in classification and/ or omission are made by third-party indexers. Additionally, there is a latency period before new Medline articles are indexed in MESH. This is one of the dangers of relying on MESH terms solely, as the authors have done, and requires extreme levels of checking if the results are to be credible.

Additionally, a check of these papers in the Science Citation Index (SCI) shows that there were seven papers published before 2004 that cited the Fernandez-Ruiz paper. Also, there was an additional paper in the SCI identified that was retrieved with the above query [8], and there were fourteen papers published before 2004 that cited the Marsicano et al. paper [8]. And this is probably the tip of the iceberg. There are obviously other ways to refer to endocannabinoids (e.g., anandamide, arachidonoyl glycerol, noladin ether) and simple searches involving these terms generated a few more papers containing links between endocannibinoids and Alzheimer's. In sum, we would not call this a discovery, or even an innovation, because the links between endocannibinoids and Alzheimer's were established well before 2004. In this particular example, these links were known by a large number of people.

Because of space limitations, I will only comment briefly on the other two purported specific 'discoveries', since even one counter-example is all that is necessary to refute a 'discovery' claim. The second specific 'discovery' example is the connection between AMPA receptors and Migraine. The authors summarized the discovery that LitLinker found as "LY293558 is promising for Migraine treatment [1]". I entered the query ((LY293558 OR AMPA) AND MIGRAINE) in both Pubmed and the Science Citation Index. Seven papers published before 2004 were identified that linked AMPA to Migraine, four of which are referenced here [9–12]. One of the pre-2004 papers [9] was published by the authors of the post-2003 paper that contained the above 'discovery' statement. The title of [9] is rather conclusive: "A double-blind, placebo-controlled study to investigate the efficacy and tolerability of 1.2 mg/kg intravenous *LY293558* versus 6 mg subcutaneous sumatriptan versus placebo in patients with an *acute migraine attack*". Again, this is the tip of the iceberg. No proxy terms were used for AMPA/LY293558 in the query, no citing papers were examined, and no patents were examined.

The third specific 'discovery' example is the use of secretin to treat Schizophrenia. The authors summarize the discovery as "LitLinker could automatically identify the potential connection between secretin and Schizophrenia. However, in a 1999 patent [13], the following links between secretin and Schizophrenia are established.

"In addition to this effect on the digestive function, secretin also appears to improve the abnormal brain activity in individuals having symptoms of autism. The increased blood flow in the brain detected during a SPECT scan after administering secretin in EXAMPLE 1 supports this theory. *While causing pancreatic secretions, secretin also stimulates the production of cholecystokinin (CCK). Deficiencies in CCK have been linked to other neurological disorders, such as schizophrenia*, and CCK production has been found to be related to levels of the neurotransmitter serotonin. Thus, secretin may be indirectly related to the body's natural production of serotonin. The increase in serotonin levels in the blood after the procedure in EXAMPLE 1 supports this relationship between secretin and serotonin."

"Accordingly, the method of treating autism by administering secretin and/or causing the body to naturally secrete required amounts of secretin corrects the secretin deficiency, improving the digestive functions in autistic patients previously experiencing intestinal difficulties and improving communication, cognition, and socialization capabilities of autistic patients. Since other neurological disorders, such as depression, obsessive–compulsive disorder, Alzheimer's, allergies, anorexia, bulimia, *schizophrenia*, also involve abnormal modulation of neurotransmitter levels, *these disorders may also be treatable with secretin*."

4. Discussion and conclusions

Literature-based discovery is a noble goal, and if it can be implemented on a large scale, it will be an enormous achievement. 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 reference [1].

References

[1] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 2006;39(6):600–11.

[2] Pazos MR, Nunez E, Benito C, Tolon RM, Romero J. Role of the endocannabinoid system in Alzheimer's disease: new perspectives. Life Sci 2004;75(16):1907–15.

[3] Benito C, Nunez E, Tolon RM, Carrier EJ, Rabano A, Hillard CJ, et al. Cannabinoid CB2 receptors and fatty acid amide hydrolase are selectively overexpressed in neuritic plaque-associated glia in Alzheimer's disease brains. J Neurosci 2003;23(35):11136–41. This article is referenced in Ref. [2].

[4] Milton NG. Anandamide and noladin ether prevent neurotoxicity of the human amyloid-beta peptide. Neurosci Lett 2002;332(2):127–30.

[5] Fernandez-Ruiz J, Lastres-Becker I, Cabranes A, Gonzalez S, Ramos JA. Endocannabinoids and basal ganglia functionality. Prostaglandins Leukot Essent Fatty Acids 2002;66(2–3):257–67.

[6] Lagalwar S, Bordayo EZ, Hoffmann KL, Fawcett JR, Frey 2nd WH. Anandamides inhibit binding to the muscarinic acetylcholine receptor. J Mol Neurosci 1999;13(1–2):55–61.

[7] Healey P, Rothman H, Hoch PK. An experiment in science mapping for research planning. Res Policy 1986;15(5):233–51.

[8] Marsicano G, Moosmann B, Hermann H, Lutz B, Behl C. Neuroprotective properties of cannabinoids against oxidative stress:

role of the cannabinoid receptor CB1. J Neurochem 2002;80(3):448–56.

[9] Sang CN, Ramadan NM, Chappell AS, Freitag FG, Smith TR, Silberstein SD, Tepper SJ. A double-blind, placebo-controlled study to investigate the efficacy and tolerability of 1.2 mg/kg intravenous LY293558 versus 6 mg subcutaneous sumatriptan versus placebo in patients with an acute migraine attack. Neurology 2001;56(8):A218. [Suppl. 3].

[10] Mitsikostas DD, Sanchez del Rio M, Waeber C, Huang Z, Cutrer FM, Moskowitz MA. Non-NMDA glutamate receptors modulate capsaicin induced c-fos expression within trigeminal nucleus caudalis. Br J Pharmacol 1999;127(3):623–30.

[11] Mitsikostas DD, Sanchez del Rio M. Receptor systems mediating c-fos expression within trigeminal nucleus caudalis in animal models of migraine. Brain Res Rev 2001;35(1):20–35.

[12] Ramadan NM. The link between glutamate and migraine. CNS Spectr 2003;8(6):446–9.

[13] Beck V, Rimland B. Method of using secretin for treating autism. United States Patent 6,197,746, March 6; 2001.

Ronald N. Kostoff *
307 Yoakum Parkway, Unit 1821,
Alexandria, VA 22304, USA
E-mail address: kostofr@onr.navy.mil

Received 19 March 2007
Available online 16 May 2007

* Fax: +1 703 696 8744.

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.05.001

Reply

Response to "Validating discovery in literature-based discovery"

Dr. Kostoff's letter illustrates how challenging it is to evaluate discovery systems, such as the LitLinker system in the paper that he critiqued. It is difficult to predict the future, and it is perhaps even more difficult to determine whether a system can predict the future accurately. In our critiqued paper, we attempted to accomplish both difficult tasks—using LitLinker to predict disease–chemical correlations that would hold in the future, and evaluating whether it could predict the future accurately. Our paper is one of the few papers that presents results from a quantitative evaluation of a literature-based discovery system because no one has devised a perfect way to perform such evaluations. One key question is particularly difficult to answer: when is something "known"? In this particular situation, we need to know when a correlation between a disease and some substance becomes "known", and thus is a candidate for discovery if a discovery system is run on material before that time. Is something known when someone suggests that there could be a connection? Is it when several researchers believe that there is a connection? Is it when most researchers believe that there is a connection? Is it when knowledge of this connection affects clinical practice? Using the best level of evidence from a medical professional's standpoint, a perceived correlation is a trusted one that should affect clinical practice if it holds in a meta-analysis of multiple, large-scale, well-designed randomized controlled trials [1]. If we require that level of evidence, then none of our "discoveries" are trustworthy even today, and they certainly were not before January of 2004, the end point for the run of LitLinker. The appropriate definition of "known" for these discovery systems is probably somewhere between the two extremes.

Thus, a main issue for any discovery system is where to draw the known vs. unknown line or where to set the discovery threshold. In our paper, we chose to define a known correlation as any correlation that occurs at least once among the MeSH terms in the MEDLINE literature. As Dr. Kostoff points out, we could have been even more conservative in defining a discovery threshold.

Let's look at each of the three discovery cases that he critiques in detail. His first example is of the correlation between endocannibinoids and Alzheimer's disease. As we stated in our article, no documents in MEDLINE before 2004 contain both MeSH terms [2]. However, as Dr. Kostoff pointed out, a search for those terms in the abstract does yield three papers [3–5]. In these documents, *Neuroprotective Agents* was the only MeSH term that appeared to link the correlations. In contrast, LitLinker suggested nine terms to link endocannibinoids and Alzheimer's disease that could provide a fruitful direction for research. Three of those linking terms were used in the example we gave to illustrate LitLinker's discovery [6]. The other document that Dr. Kostoff refers to does not mention Alzheimer's disease in the MEDLINE record [7]; thus, it would be difficult for any automated process to identify.

For the second example of AMPA receptors and Migraine, a search in MEDLINE before 2004 using the corresponding MeSH terms yields no papers. In the list of papers that Dr. Kostoff cites as evidence that this correlation was known prior to 2004, one is not in MEDLINE [8], and one does not contain either AMPA or LY293558 in the abstract [9]. Thus, it is difficult to see how an automated search could find those papers.

For the third example of secretin and Schizophrenia, there were no papers published in MEDLINE before 2004 with the corresponding MeSH terms, nor were there any that included the terms or any of their synonyms in the abstract or title. The only example that Dr. Kostoff provided was a patent. Given the poor accessibility, limited peer review, and prevalence of provisional patents, we see no need to require a patent search in addition