MINING THE WEB FOR MEDICAL HYPOTHESES
A Proof-of-Concept System

Diana Maclean†, Margo Seltzer‡
† *Stanford University, Palo Alto, CA USA*
‡ *Harvard School of Engineering and Applied Sciences, Cambridge, MA USA*
malcdi@cs.stanford.edu, margo@eecs.harvard.edu

Keywords: data mining, knowledge discovery, Vioxx, myocardial infarction, UMLS, MetaMap

Abstract: As the prevalence of blogs, discussion forums, and online news services continues to grow, so too does the portion of this Web content that relates to health and medicine. We propose that everyday, medically-oriented Web content is a valuable and viable data source for medical hypothesis generation and testing, despite its being noisy. In this paper, we present a proof-of-concept system supporting this notion. We construct a corpus comprising news articles relating to the drugs Vioxx, Naproxen and Ibuprofen, that were published between 1998-2002. Using this corpus, we show that there was a significant link between Vioxx and the concept "Myocardial Infarction" well before the drug was withdrawn from the market in 2004. Indeed, within the Vioxx-related content, the concept ranks amongst the top 3.3% in terms of importance. When compared with the Naproxen and Ibuprofen control literatures, the term occurs significantly more frequently in the Vioxx-related content.

1 INTRODUCTION

There exists a wealth of data – publicly available and easily accessible – that reports individuals' experience on a diversity of medical conditions and treatments. This corpus, which we shall refer to as the "Popular Medical Literature" (PML) is the web - or rather, all medically related web content. The PML ranges in depth and focus. A blog dedicated to a patient's experience with experimental cancer treatments, for example, an online discussion forum for expectant mothers, and a news article about Asprin are all PML constituents. Although noisy, the PML is inherently valuable: it almost always contains *some* information about a given medical topic. Recent research has noted the efficacy of using the web to bring together communities of patients, families and medical practitioners on sites such as CureTogether[1] and Patients Like Me[2], and to detect disease outbreaks, as demonstrated on Google Flu[3] and Biosurveillance[4].

[1]http://www.curetogether.com
[2]http://www.patientslikeme.com
[3]http://www.flu.google.com
[4]http://biosurveillance.typepad.com/

We propose that the PML contains medical facts and connections that may be unknown, or non-obvious, to both patients, and medical professionals alike. Moreover, knowledge of such connections might assist in patient diagnosis, identification of best practice, or in highlighting fruitful problems for future research. For example, considering that the PML contains data from more individuals than a clinician might treat in a lifetime, a patient's web search for symptoms matching her own may be a viable path to diagnosis. A recent analysis of self-reported symptom data by Cure Together, for instance, found a statistically significant correlation between Asthma and Infertility - a hypothesis that has been previously tested in clinical trials (Carmichael, 2009). At the same time, conventional web search can lead to "cyberchondria," which occurs when users search for innocuous symptoms, but become drawn to rare conditions (White and Horvitz, 2008). We believe that in addition to uncovering realistic hypotheses, a well implemented medical data mining system may help to alleviate such problems.

Despite its potential, distilling medical knowledge from PML poses considerable challenges. First, the

data is unstandardized and unstructured, as contributors tend to have no professional medical training. Second, web pages are not always time stamped, making time series analysis difficult. Finally, the PML is not scientific data: irrelevant or incorrect "fads" may obfuscate true signals.

We present a proof-of-concept prototype, demonstrating that it is both possible and productive to mine the PML for medical hypotheses. We pose the following scenario: Imagine that it is the year 2002 and that your doctor has suggested prescribing Vioxx[5]. The drug is relatively new and there is little in the published literature discussing possible adverse reactions. However, several thousand people are already using Vioxx. What if you could find out what they are experiencing? Can we develop general purpose techniques utilizing the PML that may have suggested, as early as 2002, a link between Vioxx and heart attacks? We answer this question by testing the following two hypotheses:

1. The concept "myocardial infarction" (the official term for "heart attack") was more relevant in the Vioxx-related PML than in comparative control literatures.

2. The concept "myocardial infarction" was significant *within* the Vioxx PML, compared with other medical concepts.

Strong support for both hypotheses will suggest that we may have been able to predict that Vioxx was far more dangerous than initially proposed, well before its withdrawal from the market.

The contributions of this paper include a technique for distilling medical concepts from web articles; a technique for using the TF-IDF metric to rank descriptive concepts within an entire corpus; and a prototype system providing a proof of concept for the viability of medical knowledge and hypothesis mining from the PML.

The rest of this paper proceeds as follows: Section 2 discusses related work. Section 3 imparts an overview of the story of Vioxx, highlighting what was and was not known about it at various points before it was withdrawn from the market. In Section 4, we describe the data set that we built for our analyses. Section 5 presents our experiments as well as results in testing the two hypotheses stated above. Section 6 concludes.

[5]Vioxx was released in 1999 and removed from circulation in 2004 due to a large number of adverse cardiovascular effects, in particular, heart attacks.

2 RELATED WORK

We are not the first to suggest mining the web for previously unknown relationships between medical concepts. Our work lies in the intersection of three literatures: medical knowledge discovery, contagious disease outbreak monitoring, and web-based collaborative hypothesis testing. We discuss these in turn.

2.1 Medical Knowledge Discovery

Medical Knowledge Discovery (MKD) focuses on discovering conceptual paths (or B-concepts) between a source concept (A) and target concept (C). For example: Fish Oil (A) → Vascular Health (B) → Reynaud's Syndrome (C) is a famous MKD result (Swanson, 1986). The field was pioneered by Don Swanson in the 1980's, in his search for what he called "undiscovered public knowledge" in the medical literature (Swanson, 1986). "Complementary but disjoint noninteractive structures in the literature of science do exist," he wrote, "and can lead to novel scientific hypotheses that are worth testing" (Swanson, 2001). His work resulted in a number of useful findings, including the fish oil result mentioned above and the effectiveness of magnesium as a treatment for migraines (Swanson, 1988). Most importantly, Swanson's discoveries highlighted the *efficacy* of an automated solution to hypothesis discovery and verification.

Swanson's approach was: given medical terms A and C, generate a B-list of linking concepts, and then filter that B-list until only the most relevant links remain (Smalheiser and Swanson, 1998). Most state of the art MKD systems are based on Swanson's original method (Hu et al., 2005; Gordon and Lindsay, 1996; Pratt and Yetisgen-Yildiz, 2003; Weeber et al., 2001), some with several improvements (Hu, 2005). A limitation of MKD, however, is that it requires both the source and the target terms as input.

While our work follows the spirit of MKD, two notable differences are that we utilize the PML (rather than academic literature) and that we do not require a priori identification of source and target concepts. Closest to our work in the MKD literature is the LitLinker project, with which we share several methodological techniques (Pratt and Yetisgen-Yildiz, 2003). Given a source concept, A, LitLinker uses MEDLINE[6] to retrieve related articles. It then extracts medical *concepts* from the result set titles using MetaMap[7], and filters these concepts to remove 1.) extremely common concepts, 2.) concepts highly

[6]http://www.medline.com

[7]http://mmtx.nlm.nih.gov/

similar to A, and 3.) concepts highly dissimilar to A. The latter are filtered by restricting the *semantic type* of the result set, a technique also used by Hu et al. (Hu et al., 2005). This process yields a set of B-concepts and is repeated on these set elements to acquire a set of C-concepts. LitLinker successfully replicated Swanson's migraine/magnesium discovery with a small result set of relevantly-ranked concepts (Pratt and Yetisgen-Yildiz, 2003).

2.2 Contagious Disease Outbreak Monitoring

While the previous section detailed work in extracting *linkages* in the medical literature, recent work has focused on utilizing the PML to predict *disease outbreaks*. Work in this area illustrates the richness and efficacy of the PML as a data source for medical anomaly detection. Google Flu, for example, operates on the premise that individuals experiencing influenza symptoms will engage in health-seeking behavior on the internet. By aggregating this data over geographic regions, Google Flu Trends claims to improve the Center for Disease Control's influenza outbreak predictions by 2 weeks (Ginsberg et al., 2008).

Taking a similar approach, HealthMap (Brownstein et al., 2008) uses international news articles to predict the outbreak of any disease, anywhere. HealthMap has been highly successful in predicting epidemics in real time. Although the domain of HealthMap is strongly restricted to outbreak detection, the research underlying the project considers text-mining methods designed for PML, article relevance scoring, semantic disambiguation, and several other topics of relevance to our work.

2.3 Online Medical Knowledge Sharing

Some prior work on online health communities (OHCs) supports the proposal that PML is a viable data source for medical hypothesis generation. Sites such as CureTogether[8], Patients Like Me[9], MedHelp[10], and others, cater to OHC participants primarily by providing online tools for recording and analyzing personal health data. The most common form of interaction is the discussion forum, through which OHCs accumulate data about personal illness, symptoms and treatments. Prior work indicates that this data is a viable source for medical hypothesis discovery. We point to the example mentioned in Section 1, in which CureTogether discovered a correlation between Asthma and Infertility that had been previously studied in the academic medical literature.

[8] www.curetogether.com

[9] www.patientslikeme

[10] www.medhelp.com

3 BACKGROUND

In 1998, Merck filed for FDA approval of Vioxx as a treatment for arthritis. The application included data from a drug study conducted on approximately 5400 osteoarthritis patients (Solomon et al., 2002; Prakash and Valentine, 2007), showing no difference in adverse cardiovascular effects between patients treated with Vioxx, placebo, and comparative Non-Steriodal Anti-Inflammatory Drugs (NSAIDs), such as Ibuprofin (Gilmartin, 2004). In May 1999, the FDA approved Vioxx as a treatment for osteoarthritis and acute pain; Merck released Vioxx onto the market.

Earlier that same year (January), Merck initiated the the Vioxx Gastrointestinal Outcomes Research (VIGOR) study, comprising approximately 8000 patients. The goal of the study was to show that Vioxx was safer on the gastrointestinal tract than a competing arthritis treatment, Naproxin (Berenson et al., 2004; Gilmartin, 2004; Prakash and Valentine, 2007; Reuters, 2005). This was an important feature of Vioxx: several similar painkillers (such as aspirin) cause ulcers and other gastrointestinal side effects that result in the deaths of thousands of Americans every year (Berenson et al., 2004).

Eleven months into the VIGOR study 79 of the 4000 patients on Vioxx had suffered heart attacks compared with 41 of the 4000 patients taking Naproxin. Although the VIGOR results were published in the New England Journal of Medicine in November 2000, significant data detailing the adverse cardiovascular effects were excluded (Prakash and Valentine, 2007). Not until April 2002 did a warning appear on Vioxx packaging about adverse cardiovascular effects (Gilmartin, 2004; Prakash and Valentine, 2007; Reuters, 2005). Finally, two and a half years later (September 2004), Merck voluntarily withdrew Vioxx from the market.

We use this historical timeline to select a date range in which to conduct our study. We want our end date to be well in advance of the media hype surrounding Vioxx's dangerous side effects. We pick 1998-2002, leaving a window of 1.5 years before Vioxx is withdrawn from the market.

4 THE CORPUS

In building a data corpus, we limit ourselves to PML news articles related to Vioxx and two control drugs: Naproxen and Ibuprofen. All three are painkillers, and all belong to the same "drug family" known as NSAIDs (Non-Steroidal Anti-Inflamatory Drugs). We chose Naproxen as a control drug, because it was used as the control drug against Vioxx in the VIGOR study (Berenson et al., 2004; Reuters, 2005). We chose Ibuprofen because it is one of the most common painkillers on the market. Both control drugs lack significant cardiac side effects.

We constructed our corpus by issuing a Google News[11] search on the three drugs in question. We rely on Google News' time categorization of the search hits to return articles published between 1998 and 2002[12] To retrieve the Vioxx-related PML, we searched for articles that contained the terms "Vioxx", "Ceoxx" or "Rofecoxib" (two generic names for Vioxx), but that did not contain "Ibuprofen" or "Naproxen". We proceeded similarly for the control articles. After scraping the search results and discarding the "bad" articles[13], we had 603 Vioxx articles, 141 Naproxen articles, and 500 Ibuprofen articles.

Confronted with the challenge of extracting the relevant, medical content from the HTML source code, we used the NIH's MetaMap system (Aronson, 2006), which maps text to medical concepts from the Unified Medical Language System (UMLS) Metathesaurus[14]. MetaMap was developed by the National Library of Medicine for the express purpose of extracting biomedical concepts from text. The Metathesaurus component of UMLS comprises a giant database of medical concepts drawn from several source vocabularies. Each concept is tagged with at least one semantic type from the Semantic Network, yielding a broad but consistent concept categorization. Metamap is capable of sophisticated semantic parsing and term disambiguation, including the refining of *semantically-equivalent* terms. In our case, note that "heart attack", "myocardial infarction" and other equivalent terms will *all* be mapped to the concept "myocardial infarction".

Finally, we find that restricting the semantic mapping types available to MetaMap not only makes results more meaningful, but also eliminates some "junk" from the source HTML[15]. After translating each article from text to "medicalese", we use Apache Lucene[16], an open-source search engine library, to index our corpus.

5 RESULTS

To present a viable proof-of-concept for medical hypothesis generation and testing on the PML, we test two hypotheses against our corpus. We want to know both whether the concept "Myocardial Infarction" (MI) is more significant in the Vioxx-related PML than in the control-drug related PML, as well as whether the concept MI is significant within the Vioxx-related PML itself. Support for the first hypothesis indicates that Vioxx and MI have a different relationship than than one might expect. Support for the second suggests that this relationship is meaningful.

Note that we consider only the simplest PML subsets: those that mention *only* Vioxx, *only* Naproxen or *only* Ibuprofen. In future work we plan to analyze articles containing combinations thereof. While subscribing to this level of simplicity does lose information, it also allows us to make stronger assumptions of independence between the article subsets.

5.1 H1: The concept "Myocardial Infarction" is more significant than expected in Vioxx-related articles

We use frequency of MI term occurrence as a proxy for significance *between corpora*. If the concept MI holds no particular significance in the Vioxx-related PML, then we expect that the MI term will have similar frequency distributions in the Vioxx-related and the control PML. We take this as the null hypothesis, with the alternative being that the MI term occurs significantly more frequently in the Vioxx-related PML. Table 1 summarizes the mean and variance of the MI frequency in each corpus segment.

To test our null hypothesis we use a one-sided Mann-Whitney test of the MI frequency counts for each document in the Vioxx-related PML against the MI frequency counts for each document in the control drug-related PML. We assume that by restricting

[11] http://www.news.google.com

[12] While misclassifications occur, manual inspection suggests these are rare. As web data is difficult to date accurately, we decided that this small error was acceptable.

[13] Page errors, empty articles etc.

[14] http://www.nlm.nih.gov/research/umls/

[15] A corner case is presented when text from article advertisements is parsed by MetaMap as relevant, medical information. While such text could comprise only noise, targeted advertisement text might actually increase the page information.

[16] http://lucene.apache.org/

our corpus segments to single drug mentions (as discussed above) in the given timeline that the samples are independent. The p-value for the test is 0.04453, and thus we reject the null hypothesis at the 0.05 significance level.

Drug	# Articles	Mean Freq.	Variance Freq.
Vioxx	603	0.14	0.59
Control	641	0.08	0.34

Table 1: Summary of MI term frequencies in the drug-segmented PML.

5.2 H2: MI is a Significant Term Within the Vioxx-related PML

There are a total of 4696 unique terms and 603 documents in our Vioxx-specific PML. The MI term occurs a total of 82 times, ranking in the top 250 (5%) most frequent terms in the corpus. Usually we might consider these top ranking terms irrelevant because of their high frequency; however, many common stopwords have already been removed from the text by the MetaMap UMLS mapping. Despite this, frequency is a coarse measure of importance in text. In addition, we construct a ranking of important terms within the Vioxx-related PML. Our ranking is based on the term-frequency inverse-document-frequency (tf-idf) metric:

$$tf\text{-}idf(t,d) = tf(t,d) * log\left(\frac{|D|}{|D:t \in D|}\right)$$

where $tf(t,d)$ is t's normalized frequency in a document $d \in D$, where D is the set of all corpus documents. Terms that occur frequently in a document, but infrequently in the rest of the corpus, will get a high tf-idf score, while common terms will get a low score. Within a document d, tf-idf is intended to rank highly those terms that "best describe that document".

To obtain a set, S, of terms that are highly relevant to the Vioxx-related *corpus*, we add the top 10% of terms for each document by tf-idf ranking to S. For example, if a document contained 50 terms, we would add the top 5 tf-idf scored terms to S. Each element of S is scored according to the number of documents for which it was a top-10% tf-idf term. Although simple, the technique is intuitive. While rare words will likely score within the top 10% tf-idf ranked terms of *some* document, only words truly descriptive of corpus segments should score within this range for *several* documents. Conversely, terms that occur so frequently as to be meaningless should score within that range very infrequently.

In the 3066 unique terms contained in S, MI is ranked within the top 110 (3.3%) of highly relevant terms in the corpus, and within the top 2.3% terms in the corpus. In the same company are terms that we would expect to rank highly, such as "arthritis", "pain", and "NSAID". There are also several other suggestive terms. "Diethylstilbstrol" is a non-steroidal drug that was withdrawn from the market; "duodenal ulcer" relates to the problem that Vioxx was supposed to solve (gastrointestinal complications). Finally, a quick search for "vioxx" and "deet" uncovers several articles comparing the danger of the two drugs. However, the results also contained irrelevant terms. "Text", "document" and "stock", for example, are likely artifacts of HTML junk that the MetaMap parser did not discard. Other terms, such as "activity", "wanted" and "include", could likely be filtered semantically.

6 DISCUSSION AND FUTURE WORK

The results presented in Section 5.1 support our first hypothesis: "Myocardial Infarction" is more significant than expected in Vioxx-related articles. The MI term occurs on average almost twice as often in the Vioxx-related articles than in the control articles, as shown in Table 1. Moreover, a non-parametric statistical test indicates with high confidence that the frequency distribution of the MI term in the Vioxx-related PML is significantly skewed to the right when compared to that of the Naproxen and Ibuprofen-related PML.

In light of the inter-drug PML significance of the MI term, the results presented in Section 5.2 fortify both our first and second hypotheses. A simple method that, in essence, counted the number of documents for which a word was highly descriptive, ranked the MI term in the top 3.3% of the most relevant words in the corpus. Cast in this light, our first result, the fact that the difference in MI term distribution between the corpora segments is statistically significant, becomes even more relevant. Given these results, we can claim with reasonable confidence that the concept "Myocardial Infarction" was a distinctive term in the Vioxx-related PML both *within* that literature itself, as well as *across* control PMLs, well before Vioxx was withdrawn from the market. That is, not only could MI could be tied to "Vioxx" as an important, descriptive concept, but it could also be labeled as an anomaly in that general class of PML. These results are heartening for a proof-of-concept system. We do note, however, that while 3.3% is an impres-

sive margin, 110 terms is still too many for a user to browse through. Implementing an effective *search* interface for potentially relevant links is one of the most important components of future work.

Further improvements include expanding our analyses to incorporate data sets containing overlapping drug terms. Incorporating additional drugs into the control corpus would provide more comparison points. Finally, an important goal is to cultivate Web content from alternative sources, such as Twitter feeds and blog posts, into the corpus. Developing methods for attaining, cleaning and analyzing these data will prove challenging, but we believe that results will be rewarding.

7 CONCLUSIONS

We conclude by noting that well-implemented medical knowledge discovery systems based on the PML have enormous potential. Early predictions of unanticipated drug side effects and early warnings of disease outbreaks could improve health care quality and intervention response times. Hypothesis generation with predicted return values from hypothesis confirmation could streamline medical research. But most importantly, we live in a data-driven age in which digitization of medicine is inevitable: the future of medicine will depend not only on our ability to retrieve and synthesize information from a wide array of sources, but more importantly on our ability to extract significant patterns from that information.

REFERENCES

Aronson, A. (2006). MetaMap: Mapping text to the UMLS Metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*.

Berenson, A., Harris, G., Meier, B., and Pollack, A. (2004). Despite Warnings, Drug Giant Took Long Path to Vioxx Recall. Retrieved from: http://www.nytimes.com/2004/11/14/business/14merck.html?pagewanted=2&_r=1.

Brownstein, J., Freifeld, C., Reis, B., and Mandl, K. (2008). Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Med*, 5(7):e151.

Carmichael, A. (2009). Crowdsourced Health Confirms Infertility-Asthma Finding. Retrieved from: http://curetogether.com/blog/.

Gilmartin, R. (2004). Vioxx Timeline: Key Dates for VIGOR and Long-term, Placebo-controlled Studies Implemented to Provide Cardiovascular Safety Data. Retrieved from: news.findlaw.com/hdocs/docs/vioxx/111804gilmartin.pdf.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Gordon, M. and Lindsay, R. (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.

Hu, X. (2005). Mining novel connections from large online digital library using biomedical ontologies. *Library Management*, 26(4/5):261–270.

Hu, X., Yoo, I., Song, M., Zhang, Y., and Song, I. (2005). Mining undiscovered public knowledge from complementary and non-interactive biomedical literature through semantic pruning. In *ACM CICM*, pages 249–250.

Prakash, S. and Valentine, V. (2007). Timeline: The Rise and Fall of Vioxx. Retrieved from: http://www.npr.org/templates/story/story.php?storyId=5470430.

Pratt, W. and Yetisgen-Yildiz, M. (2003). LitLinker: capturing connections across the biomedical literature. In *ACM K-CAP*, pages 105–112. ACM Press New York, NY, USA.

Reuters (2005). A Timeline of Vioxx. Retrieved from: http://www.nytimes.com/2005/08/19/business/19vioxx.timeline.html?ref=business.

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

Solomon, D., Glynn, R., Levin, R., and Avorn, J. (2002). Nonsteroidal Anti-inflammatory Drug Use and Acute Myocardial Infarction. http://archinte.ama-assn.org/cgi/content/abstract/162/10/1099?view=abstract.

Swanson, D. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18.

Swanson, D. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–57.

Swanson, D. (2001). On the fragmentation of knowledge, the connection explosion, and assembling other people's ideas. *Bulletin of the American Society for Information Science and Technology*, 27(3):12–14.

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

White, R. and Horvitz, E. (2008). Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM TOIS*, 27(4).