The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Aiding Remote Diagnosis with Text Mining

Rebecca Hellström Karlsson,[1] Vinutha Magal Shreenath,[2] Sebastiaan Meijer[2]

[1] KRY, [2] KTH Royal Institute of Technology
rebecca@kry.se, vinutha@kth.se, smeijer@kth.se

Abstract

Along with the increase of digital healthcare providers, the interest in diagnostic aids for remote diagnosis has increased as well. As patients write about their symptoms themselves, we have access to a type of data which previously was rarely recorded, and which has not been filtered by a healthcare professional. Knowledge of similar patients and similar symptoms is beneficial for doctors to arrive at a diagnosis. Therefore, the remote diagnostic process could be aided by presenting patient cases together with information about similar patients and their self-reported symptom descriptions. Apart from online diagnosis, such an aid could be beneficial in many healthcare settings, such as long-distance visits and knowledge gain from patient diaries.

In this paper, we present the impact of aiding remote diagnosis by presenting clusters of similar symptoms, using symptom descriptions collected from a virtual visit application by the Swedish telemedicine provider KRY. Symptom descriptions were represented using the bag-of-words model and were then clustered using the k-means algorithm. An experiment was then conducted with 13 doctors, where patient cases were presented together with the most representative words of the associated cluster, to measure how their work was impacted. Results indicated that it was useful in more complicated cases, but also that future experiments will require further instructions on how the information is to be interpreted.

Introduction

Permeation of digital technologies and availability of data in various aspects of healthcare has opened up a previously opaque domain. Data generated includes operations in healthcare from individual level to hospital management level and beyond (Herland, Khoshgoftaar, and Wald 2014). Social media also contributes to his data in terms of search queries, tweets, and discussion boards, making it possible to track spread of an infection or locate defective medications (Lu et al. 2013). There have been investigations' on managing this medical data, and reviews and reposts of how to organize this data and make it useful (Cohen et al. 2010; Jensen, Jensen, and Brunak 2012).

However, the impact of such data on the actual diagnostic process itself remains under-explored. The workings of a diagnostic process are based highly on each individual doctor. It is nonetheless clear that the more experience and the more patients the doctor has had, the more aware they are of the many differentials that could be possible for the presenting symptoms (Groopman and Prichard 2007). Symptom-diagnoses mappings have been investigated as a tool for understanding relationships between diseases (Zhou et al. 2014), and it can be claimed that this mapping is in large what the doctor gains from experience. However, only measuring changes in this mapping would not alone indicate that the changes improved the diagnostic process and accuracy (Field and Hole 2002).

Considering the recent tepid response of the medical community and the scientific community in general to proclaimed advantages of text mining and AI within healthcare, it would bode well to be cautious about the uses of similar tools within the same domain. It would also be useful to understand the impact of such tools and aids situated within already existing processes of diagnostics, especially collaborating with human experts (doctors).

The work presented in this paper aims to bridge that gap. In this paper, we present methods for mining of self-reported symptoms from patients. This dataset of texts from patients is not connected to outcomes of their cases. This limits us from measuring the accuracy of the methods. For this reason and as the information mined is of high sensitivity, we adopt validity of as an especially important measure, before the information is given to the doctors during the diagnostic process. We then conclude with the impact of this aid on the diagnosis and suggest future work in this area.

Related Work

Within healthcare, text mining has had a wide impact, but the work has primarily been done on clinical texts: the use

of text mining on patient-reported symptoms is less explored. It has been used to automate coding of health-status documents to ICD-9 codes (Kukafka et al. 2006), which are important for reimbursement purposes but very time-consuming to do manually (Raja et al. 2008). Other potential applications that have been proposed from research are: (1) to extract data for research initiatives from electronic medical records (Penz, Wilcox, and Hurdle 2007); (2) to predict adverse events such as vaccine reactions (Hazlehurst et al. 2005); and (3) to minimize differences in treatment depending on the doctor, by providing doctors with an optimized treatment plan based on previous patients in the same situation (Cerrito and Cerrito 2006).

The field has explored using both the traditional Vector Space Model (VSM) model for document representation (Näsman and Josephine 2013; Tremblay et al. 2009), as well as models that are more sensitive to the syntax of natural language (Weegar et al. 2015; Hazlehurst et al. 2005). Text mining in healthcare has been based on clinical text (i.e. the words of a health-care professional) to a large extent (Cerrito and Cerrito 2006; Hazlehurst et al. 2005; Penz, Wilcox, and Hurdle 2007; Kukafka et al. 2006; Tremblay et al. 2009). Text mining tools previously used in Swedish healthcare, for which vocabularies are impactful, are therefore heavily focused on clinical terms (Näsman and Josephine 2013). The words occurring in clinical text are quite different from how patients themselves describe their symptoms, and these tools can therefore not be directly applied on symptom descriptions.

Text Mining on Patient-Reported Symptoms

There exist a few studies which analyzed the content of patient-reported symptoms. One study gave patients questionnaires where they could tick boxes for their symptoms which were then used for clustering (Dipnall et al. 2016); another study performed a content analysis of patient-reported medication outcomes from social media data (Ru, Harris, and Yao 2015); and a third study analyzed tweets to identify latent infectious diseases (Lim, Tucker, and Kumara 2017). The third study indicates that the k-means algorithm is commonly used for clustering health related social media data , which is also the case for clustering of regular social media data as well (Rosa et al. 2011).

The patient texts collected from virtual visit applications resembles tweets in many aspects. The texts are short descriptions of their symptoms or health status, often less than two lines, and contain misspellings and jargon. Special consideration must therefore be taken when processing these texts, as the low word count can be an issue for document clustering. The desired effect of stemming, to connect words with the same meaning to one term, is also harder to achieve with jargon and misspellings (Rosa et al. 2011). Ru et al. improved this connection by using data sources for alternate drug or disease names to map these words together. Lim et al. did so as well, but for non-standard expressions for body parts, pain locations, and symptoms.

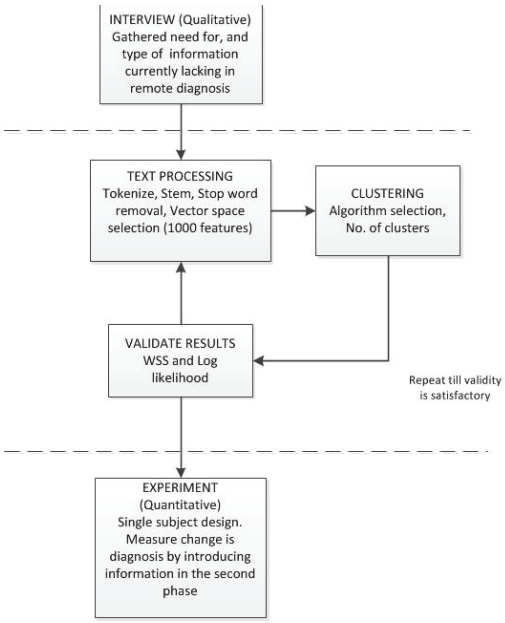

Figure 1: Methodology.

From this survey, it is apparent that while there have been many studies on text mining in healthcare, few have been conducted on the mining patient reported symptoms, though there are similar studies conducted on tweets. It is therefore of interest to investigate the impact of this data in aiding the diagnostic process.

Methodology

The research question this work addresses is: "*What is the impact of presenting additional information through text mining on the diagnostic process*?" We address this question by the methodology in Figure 1.

Initial information on the lack of information and possible presentation of the information for remote diagnosis was gathered by interviewing a doctor. This qualitative input was used during text mining to choose the method. The outcome of text mining was presented in an experiment to measure the impact of the same on the diagnostic process, by observing the changes in the differential diagnoses.

Text Mining

In this section we describe the methods used to mine patient reported symptoms. Text processing and clustering

was performed within the WEKA software application (Frank, Hall, and Witten 2016). The variations that were explored during the text mining phase are presented in Table 1.

Dataset

The data was collected from a virtual visits application by the Swedish telemedicine provider KRY. On opening the app, the user is directed to categorize their symptoms, or write free text describing their illness. This written text is the raw data that this work is built on. 11879 instances of free text were collected over a 3-month period, for text mining described in this work. The length of the free text entered was often shorter than two sentences. The data was anonymized and was treated as unlabeled data, as no outcomes were linked to the collected instances. The data was divided into train and test sets with an 80/20 ratio.

Text Processing

We used the classic model VSM (Salton, Wong, and Yang 1975) for document representation, with tf-idf as weighting scheme (Manning, Raghavan, and Schütze 2008). For every iteration each instance was tokenized with WEKAs WordTokenizer and stemmed with the (Snowball Org 2006). Several iterations of stop words and feature vector sizes were explored, which are presented in Table 1. The first iteration of stop words used a Swedish stop-word list (Bougé 2011), and was then improved upon by manual inspection on the resulting cluster centroids. Feature vector sizes between 250 and 2000 features were explored.

Clustering

Clustering was determined as the most appropriate method to structure this data, as we wanted to find homogeneous subgroups, and the k-means algorithm was selected as suitable for this problem. We experimented with the different clustering algorithms, distance measurements, and algorithm settings available in WEKA to arrive at a satisfactory combination. In order to find the optimal clustering scheme, several metrics for cluster validity were employed based on (Halkidi, Batistakis, and Vazirgiannis 2001). *Cohesion*, in terms of Within-cluster Sum of Squared errors (WSS), was used as an internal criterion. As a relational criterion, *maximum likelihood estimation* was employed (measured in log-likelihood). Lastly, as an external criterion, we employed *visual appraisal of the clusters coherency* by the doctor who was interviewed for the initial information together with the first author.

Text Mining Results

Here we present the document representation and clustering scheme that was the best fit for the data, in terms of the specified cluster validity metrics. The final document representation used a vectors space of 1000 features and an extensive stop-word list (4_3, see batch 6 in Table 1). This was used for clustering with the WEKA algorithm called SimpleKMeans with k-means++ (Arthur and Vassilvitskii 2007) for initial cluster seeding. The algorithm was wrapped into a density-based clustering algorithm, to enable the use of maximum likelihood estimation for validation of the clustering scheme. Euclidian distance was used for cohesion calculation. After trying out other cluster sizes the size k=60 was chosen as the optimal clustering scheme, after comparing WSS and log-likelihood of other k. When the test data was clustered using this scheme, 41 of the 60 clusters were formed, hence those 41 clusters were used for the final clustering model.

Table 1: The different settings for clustering algorithms and text processing that were used. When several stop-word lists or feature vector sizes are specified, one clustering run was performed for each combination.

Batch no.	Stop-word lists	Feature vector sizes	Number of clusters, K
1	1 *(original)*	1000	10-50
2	2 *(add basic stop-words)*	1500, 2000	30-55
3	3 *(add stop-words based on clustering outputs)*	500-1500 *(3 options)*	15-25
4	4_1, 4_2, and 4_3 *(3 iterations)*	250-2000 *(5 options)*	15
5	4_3	1000, 2000	40-60
6	4_3	1000	50-60

Experiment

In this section we present the experiment we designed to answer our research question. The experiment was designed with the purpose of being a single subject experiment. A questionnaire was created to execute the experiment, in which the participants were asked questions regarding the efficacy of the cluster centroids as a diagnostic aid for specific patient cases. This had no effect on any real patient. The intended participants were physicians with experience in remote diagnosis, and the participants remained anonymous.

Cluster Instances

After the optimal clustering scheme was found the test set was clustered using that model, assigning a cluster to each patient text present in the test set. Five instances belonging to different clusters were then selected from the test in-

stances, for use in the survey. Instances were chosen belonging to clusters within a range of sizes (no. o. instances), to gain insight into the efficacy for different cluster sizes. Two versions of these texts were then constructed: (1) one with only the text from the patient; and (2) one with the text from the patient and the top ten words from the cluster they were assigned to. These clusters are presented in Table 2. The two versions were then presented as clinical cases in the questionnaire.

Table 2: Top words from clusters presented to the doctors together with an instance from that cluster.

Cluster	Size	Cluster centroids (words)
A (#0)	55	magsjuk diarré lös kräkts kräk kräkning feber avföring blöja slö
B (#39)	66	halsont feber halsfluss hosta förkylning streptokocker muskelvärk snuva influensa hes
C (#53)	102	ögat vagel öga svullet röd svullnad ögonlocket rött rinner svullen
D (#51)	10	migrän huvudvärk pannan migränmedicin illamående tinning orkeslös påverka yrsel tryck
E (#33)	4679	sår springmask ångest urinvägsinfektion vattkoppor svullnad svamp ramla klåda oro

Participants

The participants for the experiments were recruited from doctors employed by KRY, who had experience in remote diagnosis. The questionnaire was open for a week, and participants were reminded of the importance of confidentiality between participants regarding details of the experiment. The experiment participants consisted of thirteen doctors, denoted as D1-D13. Their experiences ranged between 1 and 18 years, with an average of 6.8 years, and experience with remote diagnosis ranged between 0.25 and 1 year.

Questionnaire

From the initial interview it was confirmed that measuring changes in symptoms-diagnoses mappings would accurately describe changes in their diagnostic process, and therefore questions about diagnostic keywords and differential diagnoses were included. Questions regarding the benefits of the changes in the diagnosis were also constructed together with the interviewed doctor.

The questionnaire began with introductory questions about the experience of the physician, and then presented the first version of the patient texts on separate pages. For each text, the physicians were asked to identify important words in the text, and then to identify possible differential diagnoses. They were also asked to rate difficulty of diagnosis, confidence in their diagnosis, and their desired need for more information regarding that case to reach a secure diagnosis. These metrics were rated on a Likert-scale from 1 to 5. The physicians were then presented with second version of the texts along with the same questions asked for the first versions. They were also asked whether they were aided by these extra words for each patient case, and for specific comments on the aid for that case. The questionnaire ended with questions about their overall thoughts about the experiment.

Results

Effects on the Diagnostic Process

For each patient case presented with cluster centroids at least half the participants changed their diagnostic keywords, and at least half changed their differential diagnoses. The impact on these metrics for each of the clusters is presented in Figure 2. For cluster B, those who changed their keywords or differentials indicated an assumption that the cluster centroids applied to the patient case. For cluster C and cluster D, the diagnostic keywords and differential diagnoses became more aligned after the cluster centroids were presented. In particular for cluster D, one diagnosis which at first only had been part of one doctor's differential diagnoses also appeared in many other doctors' differentials.

The changes in rating for diagnostic difficulty, confidence in diagnosis, and desire for additional information indicated that clusters A, B, and C improved the diagnostic process slightly, with the exception that the degree of con-

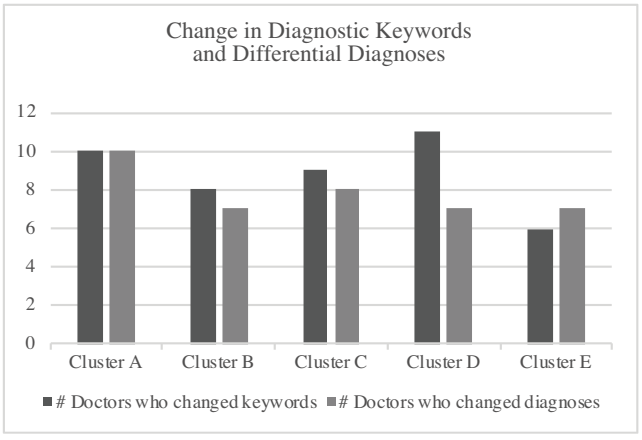

Figure 2: The impact on the participants diagnostic keywords and differential diagnoses with different clusters.

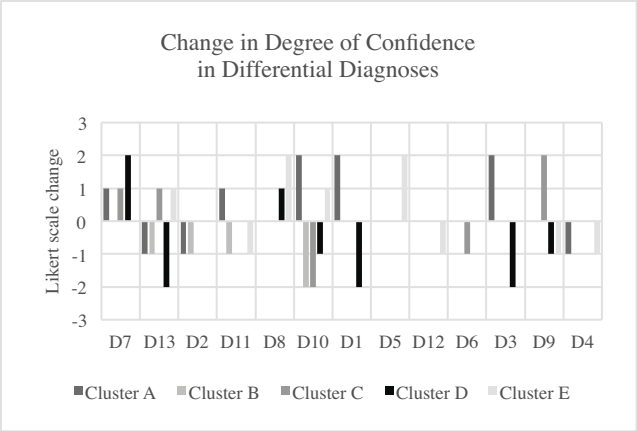

Figure 3: The perceived change in degree of confidence for each participant with different clusters. Participants are ordered in terms of experience, from left to right.

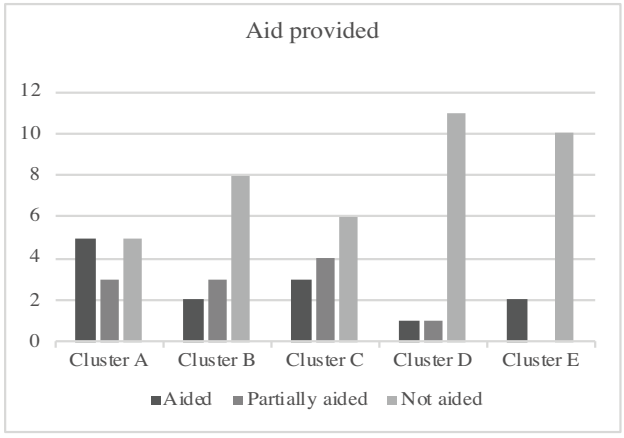

Figure 4: Aid provided with different clusters.

fidence only remained the same or decreased for cluster B. The same responses were less clear for clusters D and E. For cluster E, the average ratings indicate that the difficulty decreased and the confidence increased, but the information need increased as well. For cluster D the exact opposite was indicated for each question. The ratings were mostly consequent for each participant and cluster, e.g. that participants who had a rating improvement in diagnostic difficulty for a specific cluster often rated an improvement in the other two categories as well.

The diagnostic difficulty decreased for clusters A-C and E, but increased for cluster D. The change in confidence in diagnosis increased for clusters A, C, and E, but decreased for clusters B and D. The change in diagnostic confidence for each participant is shown in Figure 3. The change desire for additional information decreased for clusters A-D, but increased for cluster E.

Aid Provided

The participants were asked for each patient case whether the cluster centroids were useful or not. The distribution of their responses is presented in Figure 4. The participants generally felt more aided by clusters A, B, and C, which aligns well with the results of the diagnostic process. Most notable was that only 15% felt aided or partially aided by cluster D, which was the most specific cluster and the cluster which aligned the participants' diagnostic keywords and differential diagnoses the best. One reason could be that the participants thought the patient case for cluster D was very clear in-of-itself, which some comments have indicated, and therefore the additional words did not seem useful to them. Comments on cluster E were contradictory: both too little and too much information was received.

Some of the participants who didn't feel aided commented that they didn't understand the purpose of seeing what other patients had said and thought the focus should be on extracting more information on the current patient. Some of the participants who felt partially aided by the cluster centroids commented that "it's easy to get tunnel vision if you go on what the patient says" [author's translation] and that "more differential diagnoses came up, but they were somewhat in the periphery" [author's translation]. Some general comments were that the text mining aid was useful in some cases, in particular the more difficult cases. They also noted that these difficult cases are not so common in primary care, which is the setting that the data came from. One participant suggested that presenting the full texts of similar patients, and not only cluster centroids, could aid them more.

Conclusions

In this work, we have investigated the impact of presenting additional information through text mining on the diagnostic process. Patient-reported symptom descriptions were clustered and represented by their cluster centroids. The cluster centroids were then presented as a diagnostic aid in an experiment where 13 doctors judged patient cases with and without the aid. This study shows that text mining provides rich possibilities in aiding diagnosis. From the work it is evident there was an impact and it was positive in aid provided, increase in confidence and diversifying differential diagnoses; in some cases. In other cases where clusters were not specific, the opposite was reported. In general, there were also questions on how this information was generated and how it was meant to be used. This leads us to future work where interpretability issues of these techniques will need to be addressed.

However, the inclusion of doctors in the study shows new possibilities of providing doctors with more agency when being presented with more information. This will generate new ideas of how and where such aids could be

useful. The results show that cluster quality has high impact on the usefulness of the aid, and that doctors found the aid to be particularly beneficial in more rare and specific cases.

Acknowledgements

This work was conducted as master thesis at KTH. The authors thank KRY for supporting this work with data and advice.

References

Arthur, David, and Sergei Vassilvitskii. 2007. "K-Means++: The Advantages of Careful Seeding." In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035. Society for Industrial and Applied Mathematics.

Bougé, K. 2011. "Download Stop Words - Kevin Bougé." December 29, 2011. https://sites.google.com/site/kevinbouge/ stopwords-lists.

Cerrito, Patricia, and John C Cerrito. 2006. "Data and Text Mining the Electronic Medical Record to Improve Care and Lower Cost." In *Proceedings of SUGI*, 26-29.

Cohen, Aaron M, Clive E Adams, John M Davis, Clement Yu, Philip S Yu, Weiyi Meng, Lorna Duggan, Marian McDonagh, and Neil R Smalheiser. 2010. "Evidence-Based Medicine, the Essential Role of Systematic Reviews, and the Need for Automated Text Mining Tools." In *Proceedings of the 1st ACM International Health Informatics Symposium*, 376–380. ACM.

Dipnall, Joanna F, Julie A Pasco, Michael Berk, Lana J Williams, Seetal Dodd, Felice N Jacka, and Denny Meyer. 2016. "Into the Bowels of Depression: Unravelling Medical Symptoms Associated with Depression by Applying Machine-Learning Techniques to a Community Based Population Sample." *PLoS One* 11 (12): e0167055.

Field, Andy, and Graham Hole. 2002. *How to Design and Report Experiments*. SAGE.

Frank, E, M A Hall, and I H Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques",*. Morgan Kaufmann.

Groopman, Jerome E, and Michael Prichard. 2007. *How Doctors Think*. Vol. 82. Springer.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. 2001. "On Clustering Validation Techniques." *J. Intell. Inf. Syst.* 17 (2–3): 107–145.

Hazlehurst, Brian, John Mullooly, Allison Naleway, and Brad Crane. 2005. "Detecting Possible Vaccination Reactions in Clinical Notes." *AMIA Annu. Symp. Proc.*, 306–310.

Herland, Matthew, Taghi M Khoshgoftaar, and Randall Wald. 2014. "A Review of Data Mining Using Big Data in Health Informatics." *Journal of Big Data* 1 (1): 2.

Jensen, Peter B, Lars J Jensen, and Søren Brunak. 2012. "Mining Electronic Health Records: Towards Bestter Research Applications and Clinical Care." *Nature Reviews Genetics* 13 (6): 395–405.

Kukafka, Rita, Michael E Bales, Ann Burkhardt, and Carol Friedman. 2006. "Human and Automated Coding of Rehabilitation Discharge Summaries according to the International Classification of Functioning, Disability, and Health." *J. Am. Med. Inform. Assoc.* 13 (5): 508–515.

Lim, Sunghoon, Conrad S Tucker, and Soundar Kumara. 2017. "An Unsupervised Machine Learning Model for Discovering Latent Infectious Diseases Using Social Media Data." *J. Biomed. Inform.* 66 (February): 82–94.

Lu, Yingjie, Pengzhu Zhang, Jingfang Liu, Jia Li, and Shasha Deng. 2013. "Health-Related Hot Topic Detection in Online Communities Using Text Clustering." *Plos One* 8 (2): e56221.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Näsman, Markus, and S U Josephine. 2013. "Detecting Hospital Acquired Infections Using Machine Learning." KTH Royal Institute of Technology.

Penz, Janet F E, Adam B Wilcox, and John F Hurdle. 2007. "Automated Identification of Adverse Events Related to Central Venous Catheters." *J. Biomed. Inform.* 40 (2): 174–182.

Raja, Uzma, Tara Mitchell, Timothy Day, and J Michael Hardin. 2008. "Text Mining in Healthcare. Applications and Opportunities." *J. Healthc. Inf. Manag.* 22 (3): 52–56.

Rosa, Kevin Dela, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2011. "Topical Clustering of Tweets." *Proceedings of the ACM SIGIR: SWSM.*

Ru, B, K Harris, and L Yao. 2015. "A Content Analysis of Patient-Reported Medication Outcomes on Social Media." In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 472–479.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18 (11): 613–620.

Snowball Org. 2006. "Swedish Stemming Algorithm - Snowball." September 11, 2006. http://snowballstem.org/algorithms/swedish/stemmer.html.

Tremblay, Monica Chiarini, Donald J Berndt, Stephen L Luther, Philip R Foulis, and Dustin D French. 2009. "Identifying Fall-Related Injuries: Text Mining the Electronic Medical Record." *Information Technology and Management* 10 (4): 253–265.

Weegar, Rebecka, Maria Kvist, Karin Sundström, Søren Brunak, and Hercules Dalianis. 2015. "Finding Cervical Cancer Symptoms in Swedish Clinical Text Using a Machine Learning Approach and NegEx." *AMIA Annu. Symp. Proc.* 2015 (November): 1296–1305.

Zhou, Xuezhong, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. "Human Symptoms-Disease Network." *Nat. Commun.* 5 (June): 4212.