PRINT ISSN 1119-8362
Electronic ISSN 1119-8362

Full-text Available Online at
https://www.ajol.info/index.php/jasem
http://ww.bioline.org.br/ja

J. Appl. Sci. Environ. Manage.
Vol. 24 (6) 1027-1033 June 2020

Information Extraction from Electronic Medical Records using Natural Language Processing Techniques

*[1]OGBUJU, E; [2]OBUNADIKE, GN

[1]Department of Computer Science, Federal University Lokoja, Nigeria
[2]Department of Computer Science and Information Technology, Federal University Dutsinma, Nigeria
*Corresponding Author Email: emeka.ogbuju@fulokoja.edu.ng

ABSTRACT: Patients share key information about their health with medical practitioners during clinic consultations. These key information may include their past medications and allergies, current situations/issues, and expectations. The healthcare professionals store this information in an Electronic Medical Record (EMR). EMRs have empowered research in healthcare; information hidden in them if harnessed properly through Natural Language Processing (NLP) can be used for disease registries, drug safety, epidemic surveillance, disease prediction, and treatment. This work illustrates the application of NLP techniques to design and implement a Key Information Retrieval System (KIRS framework) using the Latent Dirichlet Allocation algorithm. The cross-industry standard process for data mining methodology was applied in an experiment with an EMR dataset from PubMed to demonstrate the framework. The new system extracted the common problems (ailments) and prescriptions across the five (5) countries presented in the dataset. The system promises to assist health organizations in making informed decisions with the flood of key information data available in their domain.

DOI: https://dx.doi.org/10.4314/jasem.v24i6.13

Dates: Received: 27 April 2020; Revised: 22 May 2020; Accepted: 15 June 2020

Keywords: Electronic Medical Record, BioNLP, Latent Dirichlet Allocation

In recent times a large amount of textual data are available in many hospitals' repositories because of computerization of medical processes and these data are increasing every day (WHO, 2010). Data available in health care organizations are mostly in textual form; this kind of data is not usually friendly for decision support and statistical summarization. Non-textual data are easier to analyze and summarized compared to textual data. In contrast to non-textual data, textual medical data is flexible since it captures the patient's narrative which can be told from a different perspective and allows expression of feelings. The volume of textual medical data available shows that a manual review of such data will be time-consuming and labour intensive. Thus automatic extraction of information from the textual medical data makes vital information usually buried within more accessible. The world today had provided many sources of medical information like the general web, social media, journal articles, and hospital records (Goeuriot *et al.*, 2016). The hospital records include Hospital Information System (HIS), Electronic Medical Record (EMR), Electronic Patient Record (EPR), medical diagnosis systems as well as medical image systems (Chou *et al.,* 2008). The HIS contains information about hospital inventories and administrative and management functions. The EPR manages the medical histories of patient's vis-à-vis their interactions with the hospital. While both the HIS and EPR may be specific on their roles, the EMR is all-encompassing, managing all medical records with detailed doctors' prescriptions on all the patients' medical history including treatment notes (documents) and medical images. Retrieving and making effective use of the large dataset provided by these systems especially the textual data is the purpose of this work. Insights from these datasets are of interest to both the medical practitioners and patients, and as well as the general public. To deal with textual datasets in EMRs, researchers had leveraged on the field of Natural Language Processing (NLP) and Machine Learning. Machine Learning has developed algorithms to analyze natural language text. However, applying these algorithms to EMRs have proved difficult for two reasons: 1) patients' privacy and confidentiality issues have made it difficult to obtain such data, 2) the unstructured nature of the medical text which makes it difficult to directly apply Machine Learning algorithms because most of the algorithms are trained on the structured or edited dataset (Jensen *et al*., 1012; Dorr *et al.*, 2006; Meystre *et al.*, 2010; Kalra and Ingram, 2006; Resnik *et al.*, 2008; Carroll *et al.*, 2012).

*Corresponding Author Email: emeka.ogbuju@fulokoja.edu.ng

To bypass these challenges, the application of NLP comes into play (Duna *et al.*, 2009). NLP is a subfield of artificial intelligence concerned with the intelligent processing of human language (Manning, 1999). It is an automatic approach to text analysis to process human-like language. In recent times most NLP applications are making use of Machine Learning and statistical approaches. It has been positively affected by a recent increase in textual data generation as also done in the EMRs. The booming internet or web data has favoured the researches in NLP since knowledge in the vast amount of textual data available need to be exploited. Although the process of understanding and manipulating language is extremely complex, the application of the basic techniques of NLP helps in the extraction and processing of languages in textual formats.

Among recent works in biomedical information retrieval includes graph inference retrieval model (Koopman *et al.*, 2016), radiation dosage monitoring (Kovacs *et al.*, 2016), biomedical terminology extraction system in English, Spanish and French (Lossio-Ventura *et al.*, 2016), relevance feedback on biomedical images (Markonis *et al.*, 2016) and other works that focus on users experience with the medical information systems (Palotti *et al.,* 2016; Soldaini *et al.,* 2016). None of the works had demonstrated the extraction of key information from electronic medical records using the Latent Dirichlet Allocation (LDA) algorithm. Although Zheng and Yu (2016) explored the use of LDA to extract EHR for specific queries, their work was aimed at helping patients in understanding their own electronic health records. Again, most of the works above require considerable resources for development and implementation. There is a need to develop simpler systems to overcome this challenge and provide a system that would be more focused on specific or key information extraction for better performance in the health management sector. Systems that would effectively extract key information from EMR are still on-going research efforts. It is with this backdrop that we propose a framework that would extract key information from EMRs. The aim of this work is to design and implement a Key Information Retrieval System (KIRS framework) for use in the medical information extraction domain.

MATERIALS AND METHODS

First, we gave a description of the KIRS framework which is presented in Fig. 1. The KIRS was designed to operate in three phases. The first phase would retrieve textual datasets from multiple EMRs in either single or multiple organizations. In the second phase, the NLP techniques of pre-processing, denoising, and normalization would be employed on the dataset alongside visualizations in every step. The last phase would build an extraction model or engine to generate knowledge from the processed dataset. The extraction engine will have the capacity to perform patients' opinion mining, epidemic alert/surveillance, and drug safety feedback analysis.

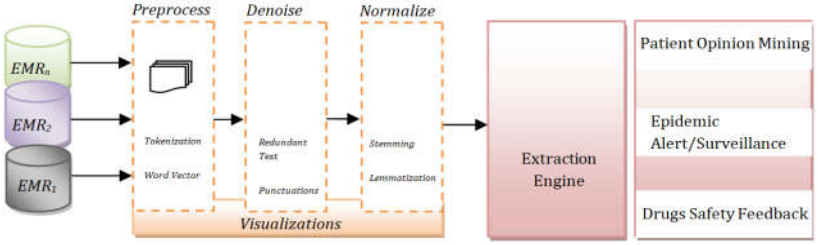

Fig. 1: KIRS framework

The engine would be implemented in this work using an LDA algorithm. LDA, as a probabilistic model, assigns a score to a word in order to group it to the most likely topic it could possibly belong to. This will help to utilize many EMRs which have the potential to reveal a range of disease conditions, discovery of unknown correlation in diseases, and classification of diseases to generate topics or key information for both health management and policy formulation. In practice, the LDA implementation generally follows the three phases in this framework in the following steps, namely: (i). load the dataset (ii) pre-process the dataset, and (iii) create the model using the document term matrix (a sparse matrix containing your terms and documents as dimensions), and visualize with a word cloud to see the terms which belong to a certain topic. The LDA Topic Model has three outputs (Jones, 2019a), they include:

$$\theta = P(\text{topic}_k|\text{document}_d) \quad - (1)$$

$$\Phi = P(\text{token}_v|\text{topic}_k) \quad (2)$$

$$\gamma = P(\text{topic}_k|\text{token}_v) \quad (3)$$

Where P=probability, k = no of topics, d = no of documents, and v = no of tokens

The model also returns log_likelihood, R-squared, and a Coherence score. The evaluation of the model is done for overall goodness of fit using the log_likelihood which is *P(tokens|topics)* at each iteration and the R-squared which is the proportion of variability in the data explained by the model (Jones, 2019a). Coherence gives the probabilistic consistency of each topic. Jones (2019a) noted that for each pair of words {a,b} in the top M words in a topic, probabilistic coherence calculates $P(b|a)-P(b)$, where {a} is more probable than {b} in the topic. It is the measure of the topic quality in essence. The overall steps in the modelling activities are presented in Fig. 2. It employs six (6) steps to generate key information or terms/words from the corpus.

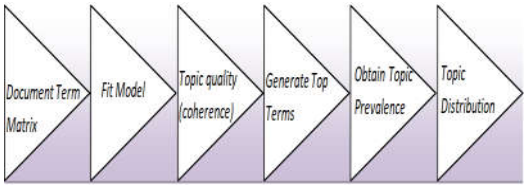

Fig. 2: Overall workflow

It starts from the creation of a document term matrix with the term frequency matrix, which shows the number of times a word (unigram) or phrase (n-gram) is occurring in the entire corpus of text. It proceeds to the creation of the extraction model and determination of the topic quality. Thereafter, a list of top terms or keywords is generated with relevant prevalence scores. Finally, the topic distribution for new documents is generated from the model.

Secondly, we describe the methodology adopted for experimentation of this system which is the cross-industry standard for data mining (CRISP-DM) depicted in Fig. 3. It is the standard methodology for data science researches.

Business Understanding: The conventional method of extracting key information from medical records requires a lot of time and effort and also gives inaccurate results especially from large EMRs. This work applies NLP on a collection of medical records from PubMed, a free online database for life science and biomedical contents. The experiment will extract key information from the dataset exactly as it would do when a typical EMR is involved. The overall essence of mining the EMRs is to assist health organizations to make sense out of the EMR data and also help in making an informed decision.

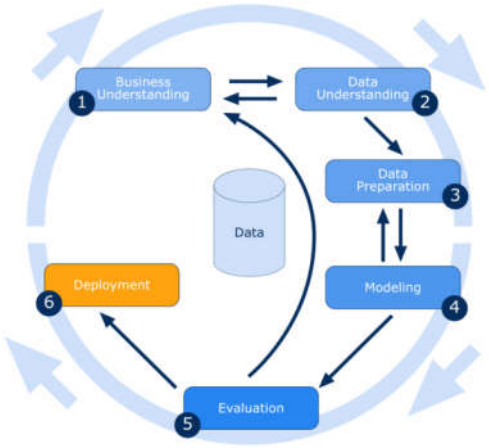

Fig. 3: Cross-Industry Process for Data Mining Methodology

The 6-phase/steps of the CRISP-DM are applied in the work alongside the implementations in R programming in the following order:

Data Understanding: The dataset contains clinical information from five (5) countries. It has 13,340 sentences and 8 features. However, since the focus of this experiment is on the extraction of key information from medical records, only two attributes were considered namely: sentences and country. Each row of the data represents a medical statement from a country while the columns hold two variables *(sentences and country).* The overview of the dataset is as shown in Table 1.

Table 1: Dataset Overview

Sentence	Country
For the treatment of uncomplicated cervical, URETHRAL OR RECTAL GONORRHEA CDC and others recommend IM ceftriaxone or oral cefixime; IM CEFTRIAXONE is the drug of choice for pharyngeal infections.	*NLD*
Diagnosis specific malignancies available for evaluation included ALL, acute myeloid leukaemia (AML), Hodgkin's disease, NHL, rhabdomyosarcoma, neuroblastoma, retinoblastoma, OSTEOSARCOMA Wilms' tumour, RETINOBLASTOMA Ewings' sarcoma, central nervous system (CNS) tumours, and hepatoblastoma.	*CAN*
Acute steady-state moderate exercise significantly altered circulating IgE concentrations in volunteers with known ALLERGY while IgE concentrations in NON-ALLERGY sufferers did not change.	*GBR*
SALMETEROL prevented EXERCISE INDUCED ASTHMA in all 13 children studied, at 1, 5, and 9 hours.	*USA*
PREECLAMPSIA AND ECLAMPSIA PREECLAMPSIA is pregnancy-induced hypertension plus PROTEINURIA	*AUS*

The sentences are medical statements or doctor's prescription for an ailment; the country is the country from which the medical statement was collected. Due to the limitations with computing resources, a sample of 100 instances was used for the experiment.

Data Preparation: The implementation of this experiment was done using the R/RStudio Programming. Prior to building the extraction engine, preliminary text pre-processing, and exploration such as word count, word cloud (a technique used to show most frequent words among a given text), noise removal, and normalization was performed on the dataset. Data pre-processing is usually used to explore datasets for insight and clean-up.

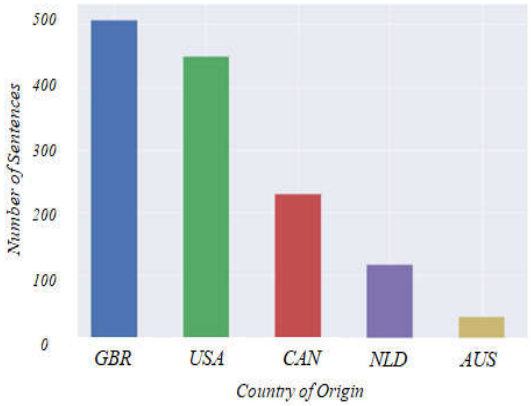

Fig. 4: Pair Plot of Sentences and Country

Fig. 5: Word Cloud for the Medical Sentences

Text normalization was employed to convert letters to lower or upper case, convert number to words or remove numbers, white spaces, stop words, and punctuations. Stop words are the most common words in a text such as "the", "a", "on", "is". These words are not usually important and do not tell any story, thus they were removed from the main text. The medical sentences were tokenized to unigrams and normalized for easy analysis and the word vector was also created using the Term Frequency Inverse Document Frequency (TF-IDF) technique. Once the dataset was cleaned up, the next step was Exploratory Data Analysis (EDA). EDA is used to find out patterns, relationships, or anomalies through visualizations which may give insight for subsequent analysis. The pair plot for sentences and country is as shown in Fig. 4. It can be observed that most of the medical statements are from GBR country. Furthermore, we generated a word cloud to show the most frequent words used in the medical sentences (see Fig. 5).

Modeling: This phase employs the use of the LDA to achieve key information extraction tasks. It is implemented by modeling the top twenty (20) topics within our experimental dataset and discussed in the Results section.

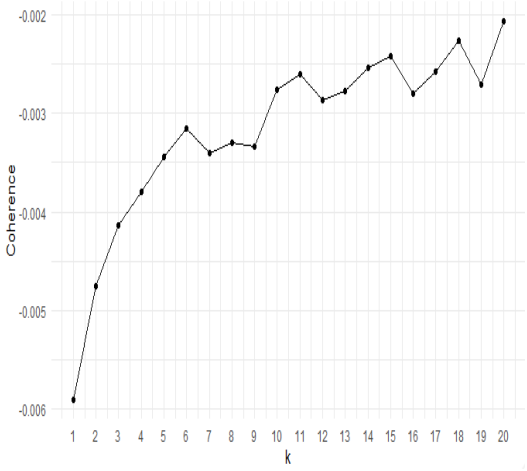

Fig. 6: Best key information by coherence score

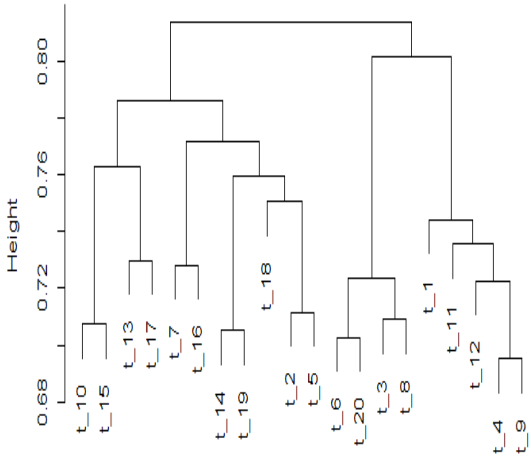

Fig. 7: Key Information shown in a dendrogram

The engine extracted the common problems (ailments) and prescriptions across the five (5) countries.

Technically, the common libraries that assisted the experiment include *"dplyr", "topicmodels", "ldatuning", "wordcloud", "ggplot2", "textmineR" and others. After PubMed's EMR data were read into the system, pre-processed, and explored, the term frequency techniques were applied to prepare the LDA model. In addition, words appearing less than 2 times or in more than half of the documents were eliminated. In a total of 500 iterations over the dataset, a sequence of twenty topics (k) was extracted alongside their coherence score. The extraction was based on the phi (Φ) function (see equation 2) which has rows representing a distribution of words over topics. Fig. 6 shows the best topics by coherence score while Fig. 7 shows the topics in a dendrogram using a Hellinger distance (distance between two probability vectors) to group related topics together.*

Evaluation and Deployment: Being a probabilistic model, R-squared, the standard metric used when constructing topic models (Jones, 2019b) was used to perform the evaluation. It returned 42%. The performance of the model was also validated by an expert knowledge through the help of a physician. The framework was made ready for consultation and use in Electronic Medical Records of hospitals across Nigeria.

RESULTS AND DISCUSSION

The result of this work is the Key Information Retrieval System (KIRS framework). The framework has an extraction engine or model which was implemented with the LDA algorithm. Using the experiment performed in the methodology section, the top key information across the countries was grouped into three major topics as follows: maintenance airways, postoperative inflammation, and intensive duration. The actual terms extracted from the EMR are presented in Table 2 alongside their coherence score which justifies their placement and their prevalence within the EMR. Apart from the key information under t_6, t_19, t_18, and t_14 which had low coherence scores, all other key information returned a high coherence score indicating that their topic quality was properly grouped.

The significance of this work is to achieve insight generation for medical personnel who use EMRs. For instance, a working knowledge can be deduced from t_19 that patients with osteoarthritis, knee, and arthritis conditions would require intensive duration care while patients with bronchial and chest_ray will require a maintenance airway care according to t_13.

Table 2: Top 20 Key Information

Topic	Label	Coherence	Prevalence	Top terms
t_9	maintenance airway	0.91	11.881	joints, protected, limb, joints affected, affected limb
t_6	maintenance airway	0.56	7.86	calcium, hypotension, maintenance airway, infection decreased, breathing
t_4	postoperative inflammation	0.95	7.782	leprae, hanseniasis-caused, leprosy, hanseniasis, infection-mycobacterium
t_19	intensity duration	0.563	7.618	osteoarthritis, trial, knee, arthritis, randomized
t_3	maintenance airway	0.93	7.173	st, angina, beat, beat-progressed, st-upsloping
t_18	maintenance airway	0.404	7.117	effective, data, model, restoration-da, effects
t_10	maintenance airway	0.782	5.277	Mg-week, week, mg, patients, objective
t_14	maintenance airway	0.578	4.841	im, im-ceftriaxone, ceftriaxone, treatment, occur
t_7	intensity duration	0.96	4.356	retinoblastoma, ewings, central, neuroblastoma, neuroblastoma-retinoblastoma
t_13	maintenance airway	0.96	4.349	general, bronchial, cambodian, general-status, chest-ray
t_17	maintenance airway	0.952	4.271	concentrations, ige-concentrations, allergy, ige, exercise
t_16	maintenance airway	0.96	4.225	involving_liver, biliary, carcinoma-possibly, discriminate, marker-hepatocellular
t_11	postoperative inflammation	0.97	3.456	overdose, skin, obtain, give, pupils
t_20	intensity duration	0.97	3.384	Day-cefaclor, cefaclor, mg-kg, ten-days, received
t_5	intensity duration	0.97	3.373	Hirsutism-monodactyly, generalized-hirsutism, eyelashes small, upturned nose, physical
t_1	maintenance airway	0.97	3.34	saccadic, myasthenia gravis, diagnostic, velocity amplitude, quantitatively assessed
t_12	postoperative inflammation	0.97	3.294	pheochromocytoma, difficult-control, case-demonstrates, differential-diagnosis, emergency-difficult
t_2	postoperative inflammation	0.766	2.365	transcatheter, anticancer, patients, arterial-embolization, embolization
t_8	postoperative inflammation	0.811	2.093	hepatic, tae, carcinoma, artery, evaluated
t_15	postoperative inflammation	0.95	1.947	arterial, tae-transcatheter, hepatic-artery, absence_gelfoam, microspheres_arterial

Maintenance airway cases are the most prevalent in the EMR. This in itself is another knowledge discovery which is very important to tracking epidemic alert and mounting necessary surveillance. The system also has the capacity to extract drugs used in the treatment of certain cases. Drugs like ceftriaxone and cefaclor were identified for the treatment of cases under t_14 and t_20 respectively. Generally, the system was able to extract thirty eight (38) drug prescriptions over forty three (43) ailments presented in Table 3. They were grouped in the Table as key information from their various topic labels (key_info_1 from t_1, key_info_2 from t_2 and so on). The evaluation from an expert medical physician confirmed that the drugs in each group have a fair chance in the treatment of the ailments in that group.

Table 3: Extracted key information from EMR

Drugs	Ailments
Key_info_1	***Key_info_1***
Opioids	*Tumours*
Chondroitin	*Hodgkins*
Ibuprofen	
Atropine	
Key_info _2	***Key_info _2***
Atropine	*Osteoarthritis*
Tizanidine	*Osteosarcoma*
Clonidine	*Monodactyly*
Ceftriaxone	*Malignancies*
Key_info _3	***Key_info _3***
Amoxicillin	*Retinoblastoma*
Key_info _4	***Key_info _4***
Phenothiazines	*Myasthenia*
Penicillin	*Gonorrhoea*
	Rhabdomyosarcoma
	Prinzmetal
	Bradydysrhythmias
	Angina
Key_info _5	***Key_info _5***
Flumazenil	*Hemoptysis*
Glucosamine	*Arthritis*
Cholinergic	*Parkinsons*
Key_info _6	***Key_info _6***
Barbiturates	*Carcinoma*
Key_info _7	***Key_info _7***
Etanercept	*Hepatocellular*
Antidepressants	*Angina*
Gluconate	*Bradycardia*
Opioid	*Arthritis*
Key_info _8	***Key_info _8***
Oxymetazoline	*Rheumatoid*
Salmeterol	*Hypotension*
	Prinzmetal
	Leprosy

This work had shown advancement over the key phrase identification system by Li and Wu (2006), Phrase-vector space model by Mao and Chu (2007), and the biomedical terminology extraction system by Lossio-Venturaa et al. (2016). It has demonstrated that beyond its extraction capacity, it can group related information for further decisions and diagnostic actions. This is a significant improvement over the Linguistic String Project Medical Language Processor system by Sager et al. (1968) which can enable the extraction and summarization of symptoms but without grouping functions for the drugs information. It performed better than the Conceptual graph by Chu and Cesnik (2001) which could capture the structure and semantic information contained in medical documents but could not extract key information as presented in this work.

It goes a long way to demonstrate that Artificial Intelligence had play a key role in providing support for medical practice and the contributions of a diagnosis support system are becoming very useful in the medical system especially the ones with the capacity for Information Retrieval. The KIRS system we implemented in this work would be useful in the construction of diagnosis support system in any medical records situation. Although it cannot replace the role of qualified medical personnel, it can be relevant to practitioners for comparing existing doctor's prescriptions from different countries in a particular ailment. Instead of traversing through huge EMRs to do the comparism, the system would discover the required knowledge through its extraction and grouping capability. In future work, we hope to explore approaches for automated information extraction from medical data using Machine Learning algorithms like KNN and Naïve Bayes with the application of NLP techniques exploited in this work as pre-processing steps. Also, we shall undertake the application of patients' opinion mining and drug safety analysis (allergies detection) with this framework in future work.

*Conclusion***:** This work had shown that NLP plays a key role in mining EMRs. It has equally shown that EMRs have the potential to reveal a range of disease conditions, discovery of unknown correlation in diseases, and classification of diseases to generate key information. The usefulness of this work covers drugs prescriptions, healthcare management, treatment comparism, and even policy formulation across countries.

*Acknowledgement***:** The authors appreciate the contributions of Dr. Magnus Ogaraku of the Federal University Lokoja Health Services for providing some expert validation functions for this research.

REFERENCES

Carroll, J; Koeling, R; Puri, S (2012). Lexical acquisition for clinical text mining using distributional similarity. Computational Linguistics and Intelligent Text Processing. Springer, Berlin Heidelberg; 232-246.

Chou, S; Chang, W; Cheng, C; Jehng, J; Chang, C (2008). An information retrieval system for medical records & documents. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 2008:1474-7. DOI: 10.1109/IEMBS.2008.4649446

Chu, S; Cesnik, B (2001). Knowledge representation and retrieval using conceptual graphs and free text document self-organisation techniques. *International Journal of Medical Informatics,* 62: 121-133.

Dorr, DA; Phillips, WF; Phansalkar, S; Sims, SA; Hurdle JF (2006). Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods Inform Med. 45*(3): 246–252.

Duna, DF; Wendy, W; Chapman, C; McDonald J (2009). What can NLP do for Clinical Decision Support?, *Journal of Biomedical Informatics*, Elsevier Inc.

Goeuriot, L; Jones, GJF; Kelly, L; Müller, H; Zobel, J (2016). Medical information retrieval: Introduction to the special issue. *Information Retrieval Journal, 19*: 1–5.

Jones, T. (2019a). Topic modeling. Retrieved from https://cran.r-project.org/web/packages/textmineR/vignettes/c_topic_modeling.html

Jones, T. (2019b). A coefficient of determination for topic models. Retrieved from https://arxiv.org/pdf/1911.11061.pdf

Kalra, D; Ingram, D (2006). Electronic health records. Information Technology Solutions for Healthcare. Springer, London; 135-181.

Koopman, B; Zuccon, G; Bruza, P; Sitbon, L; Lawley, M (2016). Information retrieval as semantic inference: A graph inference model applied to medical search. *Information Retrieval Journal.* doi:10.1007/s10791-015-9268-9.

Lossio-Ventura, JA; Jonquet, C; Roche, M; Teisseire, M (2016). Biomedical term extraction: Overview and a new methodology. *Information Retrieval Journal.* doi:10.1007/s10791-015-9262-2.

Li, Q; Wu, YB (2006). Identifying important concepts from medical documents. *Journal of Biomedical Informatics, 39*: 668-679.

Manning, CD; Schutze, H (1999). Foundations of Statistical Natural Language Processing. MIT press, Cambridge MA.

Mao, W; Chu, WW (2007). The phrase-based vector space model for automatic retrieval of free-text medical documents. *Data and Knowledge Engineering, 61*: 76-92.

Markonis, D; Schaer, R; Müller, H (2016). Evaluating multimodal relevance feedback techniques for medical image retrieval. Information Retrieval Journal. doi:10.1007/s10791-015-9260-4

Meystre, SM; Friedlin, FJ; South, BR; Shen, S; Samore, MH (2010). Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Med. Res. Methodol. 10*:70

Peter, B; Jensen, LJ; Soren, B (2012). Mining EHRs towards better research applications and clinical care. *Nature Reviews.*

Palotti, J; Hanbury, A; Müller, H; Kahn, CE (2016). How users search and what they search for in the medical domain. *Information Retrieval Journal.* doi:10.1007/s10791-015-9269-8.

Resnik, P; Niv, M; Nossal, M; Kapit, A; Toren, R (2008). Communication of clinically relevant information in electronic health records: A comparison between structured data and unrestricted physician language. Perspectives in Health Information Management, CAC Proceedings.

Sager, N; Chi, E; Friedman, C (1986). The analysis and processing of clinical narrative. *Medinfo;* Elsevier.

Soldaini, L; Yates, A; Yom-Tov, E; Frieder, O; Goharian, N (2016). Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal.* doi:10.1007/s10791-015-9258-y.

Uzuner, O; Luo, Y; Szolovits, P (2007). Evaluating the state-of-the-art in automatic de-identification. *JAMIA* 2007.

World Health Organization (2010). International statistical classification of diseases and related health problems 10th revision, edition 2010. Geneva, Switzerland.

Zheng, J; Yu, H (2016). Methods for linking EHR notes to education materials. *Information Retrieval Journal.* doi:10.1007/s10791-015-9263-1.