Research Journal of Applied Sciences 12 (7-9): 384-391, 2017
ISSN: 1815-932X

Adaptive Model for Disability Determination Decision Process Based on Natural Language Processing

[1]Eslam Amer and [2]Mohammed Abel Elfatah
[1]Department of Computer Science,[2]Department of Information System,
Faculty of Computers and Information, Benha University, Banha, Egypt

Abstract: Due to the high growth rate in claiming disability benefits, Social Security Administration (SSA) faces a real overload challenge. Disability determination process has turned out to be time-consuming, complicated and expensive. By unlocking patient's details, we can gain valuable information that could lead to improvement in the quality of healthcare, reducing time and healthcare cost. This study presents an approach to ease the process of disability determination. Our approach uses natural language processing and biomedical text mining to deal with data stored in patient's Electronic Healthcare Records (EHRs). Such data may encode significant information about the patient's case. The developed system extracts relevant medical entities and builds relations between symptoms and other clinical signature modifiers. The proposed system uses extracted information as evaluation features. Such features decide whether an applicant should gain disability benefits. Evaluations show that the proposed system accurately extracts symptoms and other laboratory marks with high F-measures (93.5-95.6%). The proposed automated system deduces right assessments to approve or reject the applicants for disability benefits.

Key words: Natural language processing, biomedical text mining, disability determination, electronic health records, reducing, quality

INTRODUCTION

Disability programs aim to provide benefits and services to a wide range of people and their families. However, programs such as the United States Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) are facing a critical challenge. The rapid growth rate of beneficiaries of disability programs became one of the major economic challenges for most countries. According to, the United States Social Security Administration (SSA) 2016 annual report (https://www.ssa.gov/oact/tr/2016/tr2016.pdf), the number of caseloads addressed by SSDI and SSI has grown substantially in recent years. At the end of 2015, SSDI and SSI programs reported that they granted disability benefits to more than 60 million people.

At the end of 2015, SSDI and SSI programs reported that they granted disability benefits to more than 60 million people. These beneficiaries were classified as 43 million workers in retirement with their dependents, 6 million survivors from crises and 11 million handicapped workers with their dependents. The total expense in 2015 was $897 billion while the total revenue was $920 billion.

Also, the rapidly increasing caseloads faced by SSA, coupled with an imbalance in administration, hamper disability determination and make it expensive and time-consuming. This results in a delay in the determination decision and may produce inaccurate results. With the lack of effective tools, applicant's cases wait for too long a time for processing. This results in issues that negatively impact the trustworthiness of SSA programs. Another factor affecting disability programs consists of the potential policy changes that followed changes in disability programs over the years. For example, policy changes due to national economic conditions can result in extra constraints. Such constraints add more limits to the applicants, reduce the levels of disability benefits and omit other initiatives taken by SSA. The previous challenges faced by the SSA programs will negatively affect its humanitarian objectives. Further, it will reflect in the adjustment of the SSA services.

Despite great efforts to strengthen the effectiveness of disability determination, processing time is severely long which causes an enormous backlog. It is shown by GAO (2004) and David *et al.* (2015) that there are inconsistencies in the eligibility criteria in different places.

Corresponding Author: Eslam Amer, Department of Computer Science, Faculty of Computers and Information, Benha University, Banha, Egypt

Besides, critical decisions used to be taken without being supported by evidence (GAO, 2012; Mann *et al*., 2014).

The real challenge for SSA is to collect accurate data and use them effectively and efficiently to manage disability programs. However, with the rapidly changing work environments and increasing numbers of applicants, the pragmatic challenge facing SSA disability programs is: why can't computers do this determination for us? Or, how to automate the whole disability determination process?

Disability program's managing systems need to evolve to cope with the rapidly changing environment which calls for a change in the current styles used to gather data. These changes would be necessary to adapt to current challenges and to any future changes in SSA's eligibility protocol. Therefore, disability programs should be improved by speeding up the determination process as well as the ability to select right candidates who deserve disability benefits and filter out false candidates (Ramampiaro 2009; Mann *et al*., 2014).

Due to the increasing healthcare costs, the quality of care services provided to patients hasn't shown any considerable improvement. Recently, a lot of research demonstrated that healthcare technologies can reduce healthcare costs dramatically. For this reason, healthcare organizations have adopted the use of Electronic Health Records (EHRs). Contemporary advances in information technology provided simple methods to gather a variety of healthcare data.

Electronic Health Records (EHRs) are the computerized version of the patient's medical history. They consist of huge relevant to patient health care like demographics, physician's observations and clinical laboratory data. EHRs may contain other information that describes observations about patient's care viewpoints or deductions. An important feature of EHRs is that they provide effective and efficient sharing methods to health care providers and organizations. These methods enable organizations to share information with one another in real time. Also, it enables authorized users to access and manipulate their data effectively. In this context, electronic health record can streamline and speed up the workflow by providing direct access to any updates to any records, in real time (Kemkarl and Dahikar, 2012). EHRs can support other health care activities such as evidence-based decision support systems. Storage and retrieval of medical records becomes more efficient with the use of EHRs which support and improve the accuracy of diagnoses and encourage patient participation the healthcare process which eventually improves the care condition and generally health outcomes (DesRoches *et al*., 2008).

Now a days, the efficiency of Electronic Health Record (EHR) systems has prompted researchers to develop Clinical Natural Language Processing (CNLP) methods. CNLP aims to understand embedded information in clinical narratives text (Meystre *et al*., 2008; Nadkarni *et al*., 2011). Through employing Natural Language Processing (NLP) and biomedical text mining on EHRs, it is possible to search and extract particular medical information. NLP provides powerful tools to process noisy, unstructured text such as medical reports. It provides an "inner locus of control" for health care professionals to manage their life.

This study introduces an approach to overcome the challenges faced by SSA through automating the disability determination process. The proposed approach uses biomedical text mining to extract particular entities from the EHR text. Such entities (ex, symptoms) and its associated clinical values will represent the new applicant's case. An important feature in this new representation is that every applicant's case becomes a set of undoubted marks or patterns. Our system uses such marks as decision parameters to enable a decision to accept or reject an applicant's case. The presented approach consequently reduces the reviewing cost and speeds up the whole determination process which in turn aids decision makers to deal effectively with the burgeoning caseloads.

Literature review: EHR systems can perform a critical role in the disability management systems from the initial determination to the final decision. However, the primary objectives of EHR systems are limited to merely supporting treatment-oriented functionalities in healthcare environments. In this study, a review of research works that utilized EHRs is presented.

Developing quantitative models to improve the precision of medicine for patients is a primary goal. These models could be used to predict or estimate patient health status. Hersh *et al*. (2007) stated that Electronic Health Records (EHRs) provide a great opportunity for speeding up clinical analytics research. Recent studies have demonstrated that most of the research works carried on EHRs focused on recognizing drug names and some signature information like dosage amount (Hakenberg *et al*., 2012; Gurulingappa *et al*., 2013; Zhao *et al*., 2014). Other research studies (Xu *et al*., 2010; Uzuner *et al*., 2010) extended the scope by including extra information about the drug. This information such as route of taking the drug and frequency (number of times the drug is taken per day) was extracted by using a regular expression-based approach. The previous systems showed improvements in extracting medical entities from patient prescription summaries. However, they focused

only on entities such as drug names and dosage sentences. The performance of such systems showed a variation in accuracy.

Some research works like the study done by Simpson and Demner-Fushman (2012), Ding and Riloff (2015), Amer and Fouad (2016) focus on extracting information that is related to identifying symptoms, chemical compounds, drug dosage and drug effects. The research done by Ding and Riloff (2015) introduces a method for building relations between diseases and symptoms based on the rate of co-occurrence between a disease and its symptoms in text articles. However, such systems didn't aim to build any predictive models that depict relations between symptoms of diseases and drugs which can be used further.

A recent research done by Gong *et al.* (2016) provided an approach to model a disease called BerMiner. The approach finds and extracts biomedical entities that are related to breast cancer, along with relations between such entities. The methodology seems promising for the generation of corpus-related entities that are related to breast cancer which can be used to classify articles or texts that are related to breast cancer. However, BerMiner didn't aim to extract critical entities that could be used as parameters in decision making.

A study research was done by Thompson *et al.* (2015) which classify EHR data from two adjacent systems that depict specific periods of cancer care episode (screening, diagnosis, treatment) and other post-treatment supervision characteristics in both the adjacent organizations. The study was about the ability to understand how both organizations provide health care trajectories and services for the patient. Thompson *et al.* (2015) assert that linking EHR data from different healthcare places can enhance information and knowledge about disease treatment services in a different organization which may result in favoring one organization over the other.

The challenges facing use of EHR data are that the data are unstructured, noisy, heterogeneous and systematically biased and therefore, hard to represent (Jensen *et al.*, 2012; Weiskopf *et al.*, 2013; Ozair *et al.*, 2015). Moreover, the clinical phenotype can be represented in different ways (Kaufman *et al.*, 2016). These challenges stand as obstacles for automatic decision support systems to identify patterns that could produce electronic decisions in real-world applications (Bengio *et al.*, 2013). A successful predictive or decision support system solely depends on the feature selection and some data representation associated with the features (Bengio *et al.*, 2013; Jordan and Mitchell, 2015). A common approach is to have a domain expert designated to tell which patterns to look for in EHRs and to specify the clinical associative data that assure selected patterns.

This study presents a framework to represent patients (caseloads) using a set of selected weighted features. The main objective is to introduce a possible solution to the problem of disability determination by automating the process of deciding whether a candidate qualifies for the disability benefits or not.

The problem can be divided into two steps: identifying relevant medication entities (e.g., drug names, diseases, certain abbreviations) and determining the relation between detected symptoms and other signature modifiers. For example, in a candidate's application with Chronic Kidney Disease (CKD), the following symptoms can occur: nausea, anemia, vomiting, loss of appetite, fatigue and weakness such symptoms can be classified as general ones and can be shared with other diseases. Other factors like age, family history, etc. could be taken into consideration; however such factors can also be viewed as signs without confidence.

In the proposed reserch, clinical lab reports are used to extract unquestionable data marks that are associated with relevant signs. Based on those signs, the decision could be taken automatically without any need for further investigation. For example, CDK candidate can be considered disabled when (some or all) previous symptoms occur associated with some laboratory impairments like Glomerular Filtration Rate (GFR) value between 15-29 mL/min per 1.73 m^2 and hemoglobin <11 g/dL at entry (Thomas *et al.*, 2008). This can be viewed as a confidence decision factors or disability criterion which affect the disability selection process for CDK disability applicant. When a match takes place between data extracted from the applicant's case with the indexed disability criterion, the applicant's case is accepted; otherwise, it is rejected.

MATERIALS AND METHODS

Extracting the right signs and marks is the main issue that decision makers depend on while making their decisions to allow or disallow disability benefits. Automating the process of selection of the right candidates who qualify for disability benefits will save a great amount of time and money.

In this research, three disability diseases namely chronic liver disease, chronic respiratory disorders and chronic heart failure are selected as an initial evaluation for the proposed system. The selected diseases have been picked up from the adult section in Social Security Administration (SSA) Web page https://www.ssa.gov/disability/professionals/bluebook/AdultListings.htm.

Initially, all symptoms for each disease are collected, using specialized medical experts from SSA blue book (http://www.ssa.gov/disability/professionals/bluebook/). All extracted symptoms and relations existing for each

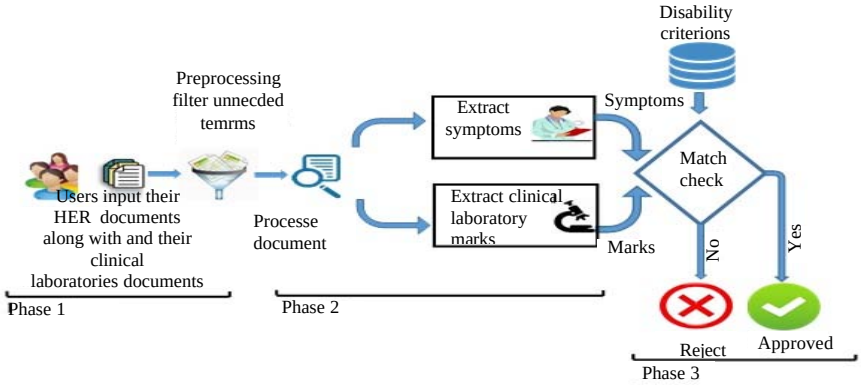

Fig. 1: The proposed system

database. SSA blue book contains disability evaluation conditions under social security. These conditions are clinical laboratory assessments; by modeling such conditions, the proposed system becomes enabled to make a decision about who can be called disabled and who cannot.

The proposed technique (Fig. 1) aims to introduce a methodology to automate the process of decision making. The technique has three main phases, namely Preprocessing of the unstructured EHR text, extraction of main entities found in the processed text and Identification of the specific entities and their associated value. Each phase will be discussed in detail.

Preprocessing of the unstructured EHR text: The purpose of document preprocessing is to organize input document for further processing. The main objective of this phase is to keep important items (e.g., nouns) and get rid of trivial, non-important ones (e.g., verbs and stop words). For this purpose, the document is initially parsed using GENIA (http://www.nactem.ac.uk/GENIA/tagger/) part-of-speech tagger. The following example elaborates the preprocessing of an input document. Consider the following sentence: "Cancer is a malignant disease". The resulting output according to GENIA part of speech tagger is the following: "Cancer/NN is/VBZ a/DT malignant/JJ disease/NN".

Where NN tag represents a noun, VBZ tag represents a verb, DT tag represents a determiner and JJ tag represents an adjective. Noun tags and adjectives tags are kept as they represent the high influence core of the sentence; however, other tags are discarded as they are of the least significance. GENIA part-of-speech tagger is designed particularly to extract information from medical text; therefore, it is a powerful tool for preprocessing medical documents such as EHRs.

For grammatical reasons, documents usually use various forms of a word such as play, playing and plays. Additionally, there are families of derivationally related words with exact or similar meaning such as democracy, democratic and democratization. In many cases, it sounds as if it would be helpful to search for one of these words and retrieve documents that contain another word in the set.

Since, words can have various morphological variants that lead to similar semantic interpretation, words are stemmed and lemmatized.

The objective of both stemming and lemmatization is to decrease inflectional forms and occasionally derivationally related forms of a word to a plain base form. For example car, cars, car's and car's will be stemmed to word car.

In our experiment, we rely on the most common algorithm for stemming English that has repeatedly been shown to be empirically very effective, viz., Porte's algorithm (Willett, 2006).

Lemmatization usually refers to performing things duly with the use of a vocabulary and the morphological analysis of words, usually intended to strip inflectional endings only and to return the original or base form of a word which is known as the lemma. In comparison with stemming if confronted with the word or token "saw", stemming might return just s while lemmatization would check to return either word see or saw based whether the part-of-speech tag of the token was a verb or a noun.

The proposed system relies on the Stanford CoreNLP (http://stanfordnlp.github.io/CoreNLP/) Natural Language Processing toolkit to perform lemmatization for words.

The output of the preprocessing step is a set of nouns which are the basic entities found in the EHR document.

Extraction of main entities found in the processed text: The aim of this step is to identify the main medical entities that are existing in the EHR document. Entity recognition or Named Entity Recognition (NER) is aimed at detecting specific terms which refer to relevant entities. In the medical domain, entities may refer to genes, proteins, diseases or drugs.

The critical challenge facing NER in the medical domain is due to the fact that the same phenomena can be referred by several forms. For example, terms like "epilepsy (http://medical-dictionary.thefreedictionary.com/epilepsy)" and "falling sickness" are synonyms that refer to the same disease named "A chronic disorder characterized by paroxysmal brain dysfunction due to excessive neuronal discharge".

To overcome this challenge it was obligatory to rely on a medical dictionary to identify the meaning of the term. In this resesrch, we used Web dictionary (http://medical-dictionary.thefreedictionary.com) as a knowledge source to identify extracted entities and their description. Web dictionary provides accurate definitions of various terms used in the health sciences like anatomy, physiology, diseases, drugs, tests and procedures. One of the greatest features of web dictionary is that it includes the named entity and its associated synonyms which affect the elasticity of the system which includes medical terms that we are going to work with. The output of this step is the set of all named entities that are found in the EHR document.

Identify specific entities and their associated value: Extracting named entities found in medical documents is of great importance; however, in clinical lab reports there are some entities that are favored over the others and considered as the main reference when a decision has to be made about the patient. The objective of identifying specific, named entities along with their associated values is to gather meaningful information from the report. To deal with this issue, regular expression rules are used to extract such entities and the closest value associated with them.

The output of this step is a set of marks. The output marks will be classified against sets of criteria initially given and stored as a disability benchmark database for given disease(s). In the proposed system, the similarity between extracted marks and stored benchmarks was calculated using the cosine similarity function and defined as:

$$\text{Similarity}(A,B) = \cos\cos(\theta) = \frac{A \times B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A^2}\sqrt{\sum_{i=1}^{n} B^2}} \quad (1)$$

where A, B are components of Vector A and B, respectively.

The marks will align to the most similar disease; then marks are evaluated against some laboratory values specifically related to the selected disease. Based on the eligibility evaluation, the candidate is granted disability benefits; otherwise, the application is rejected.

RESULTS AND DISCUSSION

In this study, we describe the dataset used in the experiment as well as the evaluation and discussion of experiments.

Dataset: Experiments were carried out on 140 different EHRs related to different diseases for patients of different ages. The categories of EHRs dataset are grouped by diseases such as kidney (45 EHR documents) respiratory (60 EHR documents) and heart failure (35 EHR documents). Privacy of patients is protected and any declaration about their names or address or any type of contacts was removed from the report. The patient's name was just replaced by a distinctive number to be identified with during the experiment.

Evaluation: Evaluation of the proposed technique has two evaluations metrics; the first is the system's ability to correctly identify terminologies found in the EHR file. The second evaluation is the accuracy of selecting whether a candidate deserves disability benefits or not.

The system's ability to retrieve terms found in EHR documents is calculated using standard retrieval measures precision, recall and F-measure metrics which are defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

As:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Where:

TP = True Positive represents the number of correct terminologies identified

FP = False Positive the number of false or incorrect terminologies retrieved by the system

FN = False Negative represents relevant terminologies that were not identified or retrieved as correct by the system

Evaluation in Table 1 showed that the proposed system can accurately extract symptoms and other laboratory associated information with high F-measures (93.5-95.6%).

Figure presented in Table 1 show promising results for identifying symptoms. However, symptoms are vague and cannot be used for decision/determination as some symptoms are common to many diseases. For example, consider the Chronic Obstructive Pulmonary Disease (COPD) and Congestive Heart Failure (CHF). Both of them have common symptoms like "short of breath", "chest pain", "coughing", though both the diseases are completely different.

The decision process needs strong, precise arguments which are considered as undoubted evidence in conjunction with symptoms.

Table 2 showed that based on the disability criteria, eligibility criteria stored in the proposed system database, coupled with the candidate input case file (EHR documents plus clinical lab reports) the proposed system performed correct assessments for candidates who qualify for disability benefits and those who do not.

Decision determination relies on Boolean criteria, either match or mismatch. However, in real assessments, decisions should be fuzzified from acceptance to rejection. The proposed model intends to introduce a possible solution to deal with the enormous number of candidate's applications. Fuzzy rules will be inserted instead of Boolean rules to guarantee the precision in accepting or rejecting a case.

The disability determination process is extremely long. A possible solution to the enormous caseloads and backlogs is to automate the determination process. In this study, we presented a system that extracts patient descriptors from text documents like EHRs. The proposed system captures critical features that can effectively aid the determination process which accelerates the whole determination process. Results obtained showed that the proposed system can accurately extract disease symptoms and other clinical associated information from EHR documents with high F-measures.

Table 1: Evaluation results for system performance in identifying terms in EHR documents

Diseases	Precision	Recall	F-measure
Kidney	0.962	0.910	0.935
Respiratory	0.945	0.925	0.935
heart	0.951	0.962	0.956

Table 2: Evaluation results for system performance in disability eligibility assessments

Diseases	Values	Granted	Rejected	Accuracy of selection (%)	Accuracy of rejection (%)
Kidney	45	6	39	100	100
Respiratory	60	12	48	100	100
Heart	35	4	31	100	100

Our system makes use of external medical knowledge source to overcome the multi-form challenge facing medical named entities. The proposed system performs correct assessments for both candidates: those who qualify for disability benefits and who don't. The system relied on Boolean rules to decide whether the candidate qualifies for disability grant or not. These rules can be viewed as facts that can be either true or false. However, the system can be more effective if it relies on fuzzy rules, to estimate the degree or ratio of disability and to decide whether the candidate's case is eligible for partial disability grants.

CONCLUSION

In this study, a new system has been proposed which deals with the challenges encountered by the disability determination process. The rapid growth in the number of applicants according to Social Security Administration (SSA) 2016th annual report, should be coupled with improvements in selection methodologies that overcome the current limitation of existing systems which have been shown to be time-consuming, complicated and expensive. The proposed system utilizes Patient's Electronic Clinical Records (EHRs) by employing natural language processing to obtain valuable information about patients. By extracting particular entities from patient EHRs along with clinical signature modifiers a decision could be taken automatically without any need for further investigation. The proposed methodology reduces disability determination process time from days or months to a couple of minutes which enhances the efficiency of the system. Experimental evaluation showed that the proposed system can accurately extract symptoms and other laboratory-associated information with high F-measures (93.5-95.6%) which proofed the effectiveness of the system. According to the causes and symptoms of

diseases along with its clinical degrees listed in SSA blue book (this is can be viewed as SSA eligibility criteria) the proposed system revealed correct assessments for all types of candidates.

SUGGESTIONS

In future research, we shall emphasize the use of fuzzy rules to evaluate whether a candidate has a partial disability and if so, to what degree and also to make a recommendation to a candidate to apply for the disability program after a certain period of time.

REFERENCES

Amer, E. and K.M. Fouad, 2016. Keyphrase extraction methodology from short abstracts of medical documents. Proceedings of the 2016 8th Cairo International Conference on Biomedical Engineering Conference (CIBEC), December 15-17, 2016, IEEE, Cairo, Egypt, ISBN:978-1-5090-2987-7, pp: 23-26.

Bengio, Y., A. Courville and P. Vincent, 2013. Representation learning: A review and new perspectives. IEEE. Trans. Pattern Anal. Mach. Intell., 35: 1798-1828.

David, H., N. Maestas, K.J. Mullen and A. Strand, 2015. Does delay cause decay? The effect of administrative decision time on the labor force participation and earnings of disability applicants (No. w20840). National Bureau of Economic Research, Cambridge, Massachusetts, USA.

DesRoches, C.M., E.G. Campbell, S.R. Rao, K. Donelan and T.G. Ferris *et al*., 2008. Electronic health records in ambulatory care a national survey of physicians. New Engl. J. Med., 359: 50-60.

Ding, H. and E. Riloff, 2015. Extracting information about medication use from veterinary discussions. Proceedings of the 2015 International Conference on North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT), May 31-June 5, 2015, University of Utah, Salt Lake City, Utah, pp: 1452-1458.

GAO., 2004. Social security administration: More effort needed to assess consistency of disability decisions. Government Accountability Office, Washington, DC. USA.

GAO., 2012. Modernizing SSA disability programs: Progress made but key efforts warrant more management focus. Government Accountability Office, Washington, DC. USA.

Gong, L., R. Yan, Q. Liu, H. Yang and G. Yang *et al*., 2016. Extraction of biomedical information related to breast cancer using text mining. Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), August 13-15, 2016, IEEE, Changsha, China, ISBN:978-1-5090-4093-3, pp: 801-805.

Gurulingappa, H., L. Toldo, A.M. Rajput, J.A. Kors and A. Taweel *et al*., 2013. Automatic detection of adverse events to predict drug label changes using text and data mining techniques. Pharmacoepidemiology Drug Saf., 22: 1189-1194.

Hakenberg, J., D. Voronov, V.H. Nguyen, S. Liang and S. Anwar *et al*., 2012. A SNPshot of PubMed to associate genetic variants with drugs, diseases and adverse reactions. J. Biomed. Inf., 45: 842-850.

Hersh, W.R., 2007. Adding value to the electronic health record through secondary use of data for quality assurance, research and surveillance. Am. J. Manage. Care, 81: 126-128.

Jensen, P.B., L.J. Jensen and S. Brunak, 2012. Mining electronic health records: Towards better research applications and clinical care. Nat. Rev. Genet., 13: 395-405.

Jordan, M.I. and T.M. Mitchell, 2015. Machine learning: Trends, perspectives and prospects. Sci., 349: 255-260.

Kaufman, D.R., B. Sheehan, P. Stetson, A.R. Bhatt and A.I. Field *et al*., 2016. Natural language processing; Enabled and conventional data capture methods for input to electronic health records: A comparative usability study. JMIR. Med. Inf., Vol. 4, 10.2196/medinform.5544

Kemkarl, O.S. and D.P.B. Dahikar, 2012. Can electronic medical record systems transform health care? Potential health benefits, savings and cost using latest advancements in ict for better interactive healthcare learning. Intl. J. Comput. Sci. Commun. Networks, 2: 453-455.

Mann, D.R., D.C. Stapleton and D.J. Richemond, 2014. Vocational factors in the social security disability determination process: A literature review. Mathematica Policy Research, Washington, DC., USA.

Meystre, S.M., G.K. Savova, K.C. Kipper-Schuler and J.F. Hurdle, 2008. Extracting information from textual documents in the electronic health record: A review of recent research. Yearb Med. Inf., 35: 128-144.

Nadkarni, P.M., L. Ohno-Machado and W.W. Chapman, 2011. Natural language processing: An introduction. J. Am. Med. Inf. Assoc., 18: 544-551.

Ozair, F.F., N. Jamshed, A. Sharma and P. Aggarwal, 2015. Ethical issues in electronic health records: A general overview. Perspect. Clin. Res., 6: 73-76.

Ramampiaro, H., 2009. Retrieving biomedical information with biotracer: Challenges and possibilities. Master Thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Simpson, M.S. and D. Demner-Fushman, 2012. Biomedical Text Mining: A Survey of Recent Progress. In: Mining Text Data, Aggarwal, C.C. and Z. ChengXiang (Eds.). Springer, Berlin, Germany, ISBN:978-1-4614-3222-7, pp: 465-517.

Thomas, R., A. Kanso and J.R. Sedor, 2008. Chronic kidney disease and its complications. Primary Care Clin. Office Pract., 35: 329-344.

Thompson, C.A., A.W. Kurian and H.S. Luft, 2015. Linking electronic health records to better understand breast cancer patient pathways within and between two health systems. EGEMs., Vol. 3, 10.13063/2327-9214.1127.

Uzuner, O., I. Solti and E. Cadag, 2010. Extracting medication information from clinical text. J. Am. Med. Inf. Assoc., 17: 514-518.

Weiskopf, N.G., G. Hripcsak, S. Swaminathan and C. Weng, 2013. Defining and measuring completeness of the electronic health records for the secondary use. J. Biomed. Inf., 46: 830-836.

Willett, P., 2006. The porter stemming algorithm: Then and now. Program, 40: 219-223.

Xu, H., S.P. Stenner, S. Doan, K.B. Johnson and L.R. Waitman *et al*., 2010. MedEx: A medication information extraction system for clinical the narratives. J. Am. Med. Inf. Assoc., 17: 19-24.

Zhao, L.L., T. Zhang, L.W. Zhuang, B.Z. Yan and R.F. Wang *et al*., 2014. Retracted Article: Uncovering the pathogenesis and identifying novel targets of pancreatic cancer using bioinformatics approach. Mol. Biol. Rep., 41: 4697-4704.