J Pathol Inform

Original Article

The utility of including pathology reports in improving the computational identification of patients

Wei Chen[1], Yungui Huang[1], Brendan Boyle[2], Simon Lin[1]

[1]Department of Research and Development, Research Information Solutions and Innovation, Nationwide Children's Hospital, 575 Children's Crossroad, Columbus, Ohio 43215, [2]Department of Gastroenterology, Nationwide Children's Hospital, 700 Children's Dr, Columbus, Ohio 43205, USA

E-mail: *Dr. Wei Chen - wei.chen@nationwidechildrens.org
*Corresponding author

Received: 22 January 2016 Accepted: 08 October 2016 Published: 29 November 2016

Abstract

Background: Celiac disease (CD) is a common autoimmune disorder. Efficient identification of patients may improve chronic management of the disease. Prior studies have shown searching International Classification of Diseases-9 (ICD-9) codes alone is inaccurate for identifying patients with CD. In this study, we developed automated classification algorithms leveraging pathology reports and other clinical data in Electronic Health Records (EHRs) to refine the subset population preselected using ICD-9 code (579.0). **Materials and Methods:** EHRs were searched for established ICD-9 code (579.0) suggesting CD, based on which an initial identification of cases was obtained. In addition, laboratory results for tissue transglutaminse were extracted. Using natural language processing we analyzed pathology reports from upper endoscopy. Twelve machine learning classifiers using different combinations of variables related to ICD-9 CD status, laboratory result status, and pathology reports were experimented to find the best possible CD classifier. Ten-fold cross-validation was used to assess the results. **Results:** A total of 1498 patient records were used including 363 confirmed cases and 1135 false positive cases that served as controls. Logistic model based on both clinical and pathology report features produced the best results: Kappa of 0.78, F1 of 0.92, and area under the curve (AUC) of 0.94, whereas in contrast using ICD-9 only generated poor results: Kappa of 0.28, F1 of 0.75, and AUC of 0.63. **Conclusion:** Our automated classification system presented an efficient and reliable way to improve the performance of CD patient identification.

Key words: Celiac, classification, machine learning, natural language processing, patient registry

Access this article online
Website:
www.jpathinformatics.org
DOI: 10.4103/2153-3539.194838
Quick Response Code:

INTRODUCTION

Celiac disease (CD) is a common autoimmune disorder estimated to affect approximately 1% of the US population.[1] Many studies have evaluated methods for identifying high-risk populations of patients in need of proactive screening for case identification.[2,3] As with many chronic diseases, quality improvement approaches can be applied to populations of patients with CD to

This article may be cited as:
Chen W, Huang Y, Boyle B, Lin S. The utility of including pathology reports in improving the computational identification of patients. J Pathol Inform 2016;7:46.
Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2016/7/1/46/194838

standardize care and track outcomes.[4-6] Effectively identifying and tracking outcomes of these cohorts of patients requires the development of an accurate CD patient registry.

However, the development of CD patient registries often requires significant resources including practitioners, informatics teams, and administrative support. These are barriers that cannot be overcome at many centers. Our center has developed and maintained a patient registry including all newly diagnosed patients since early 2012.

Prior studies have evaluated the reliability of International Classification of Diseases-9 (ICD-9) searches to identify patients with CD with only moderate success yielding only 17% of sensitivity.[7] Contrast experiments have been done using manually defined keywords to classify high-risk CD cases generating much more accurate results at 73% of sensitivity.[7] Few efforts combining different aspects of available information to further improve classification accuracy including the use of data from visits, laboratory tests and pathology reports have been described.

Pathology reports have been previously used in building natural language processing (NLP) systems for automated chart review of patients. Here, we define NLP broadly as methods to process textual data. One study found pathology reports to be insufficient for detecting breast cancer when using NLP approach compared with manual approaches[8] while others found the use of NLP for analyzing pathology reports to be effective on detection of colon cancer[9,10] and manifestation of prostatectomy details.[11] Currently, a lack of research exists on investigating the usefulness of pathology reports on automated CD detection.

Based on a subset of patients' prescreened using ICD-9 code 579.0, we evaluated the effectiveness of using natural language features from pathology reports to improve the identification of CD patients. This paper was among the first to adopt a machine learning approach to further increase the accuracy of CD patient identification based on a combination of pathological and clinical metrics. In contrast to prior research, we conducted contrast experiments using both natural language features and clinical features under different machine learning configurations for celiac identification.

MATERIALS AND METHODS

We extracted data from our Epic (EPIC Systems, Inc. USA) clinical system on patients who visited our hospital from 2012 to 2015 and had ICD-9 code of 579.0 assigned indicating concern for CD. Our subsequent analysis was, therefore, based on patients with concerns rather than general patient population. Laboratory results for tissue transglutaminase (tTG) were searched and reviewed. A laboratory test result was marked with a nominal value of either high (≥20 EU/mL) or normal (<20 EU/mL). An upper endoscopy (EGD) may also be performed then resulted in a text-based pathology report, which could be used for NLP analysis.

The registry that we developed are all confirmed diagnosis of CD based on confirmation with both abnormal laboratory/serology testing and biopsy confirmation after detailed chart review. For each of our patients in the registry, CD could be considered a primary diagnosis. Most confirmed CD patients (confirmed by our experts for having CD) only had one pathology report while false-positive CD patients (confirmed by our experts for not having CD) had either one or zero pathology report. Retrieved patient were restricted by having at most one pathology report. Twenty-nine patients who had two or more pathology reports were excluded from our experiment as they presented a challenge to our analysis if inconsistent information about CD diagnoses existed at different periods of time. In patients with multiple tTG laboratories, the most elevated laboratory result obtained within the past 6 months of the EGD was chosen to represent the laboratory status of that CD patient.

Variables used for celiac classification are listed in Table 1. For the report category, the total number of pathology reports was used with one indicating

Table 1: Variables for celiac disease classification

Category	Variable description	Variable name	Variable set ID
Report	Total number of pathology reports	Number of reports	V1
	User defined high-risk phrases	High-risk phrases	V2
	Automatically extracted n-grams	n-grams	V3
Lab	Worst laboratory results in the past 6 months	Laboratory results	V4
	Total number of laboratories	Number of laboratories	V5
ICD-9	Total number of ICD-9 codes	Number of ICD-9	V6
All-1	All variables from V1 to V6 Without feature selection	All-1	V7
All-2	All variables from V1 to V6 With feature selection	All-2	V8

ICD-9: International Classification of Diseases-9 patients having EGD done and 0 otherwise (V1). Both user-define phrases (high-risk phrases, or V2) and automatically extracted phrases (n-gram phrases generated from our machine learning program, or V3) were chosen to evaluate the effectiveness of utilizing pathology reports. For the laboratory category, both the laboratory results (V4) and total number of labs done were used (V5). For the ICD-9 category, the total number of CD ICD-9 codes assigned (V6) was used for classification.

To evaluate the effectiveness of using different combinations of variables and feature selection methods, we further developed two additional set of variables based on the aforementioned six sets of variables above (V1–V6). Combining all first six categories of variables created the variable set 7 (V7). Automatically selecting features out of V6 using a feature selection algorithm resulted in the variable set 8 (V8) [Table 1].

Feature Extraction from Pathology Reports for Classification

For expert-knowledge-driven feature selection, we identified a list of nine key phrases described by clinicians as commonly used in pathology reports as indicators of high-risk CD. These nine key phrases, which were all converted to lower cases, included Brunner's gland hyperplasia, flattening villi, intraepithelial lymphocytes, Marsh gland stage, Marsh lesion, Marsh s3 lesion, shortened villi, villous blunting, and villous atrophy. Each key phrase was first converted to a binary variable with the value of one indicating the phrase existed in the pathology report and 0 otherwise.

Automatically extracted n-gram features were phrases generated using Weka, an open source Java-based machine-learning program, developed by The University of Waikato, New Zealand.[12] The maximum number of words included in n-gram feature was chosen to be three resulting in a large set of text features consist of uni-, bi- and tri-grams. We only collected phrases up to trigrams because the length of the longest user-provided high-risk phrase was also three. We kept the first 4000 text features with all stop words (i.e., a list of common words such as "the," "of," "is," etc.) skipped.

During the process of converting documents into n-gram features, term frequency-inverse document frequency (TF-IDF) was calculated on each n-gram feature. TF-IDF is a data transformation method commonly used in text classification to show the importance of a word to a document in a collection of documents. It has been found useful in several recent classification and information retrieval studies using medical documents.[13,14] Finally, all n-grams were converted to lowercase to avoid duplications during TF-IDF transformation.

Building Celiac Disease Classifiers

Classification model

There are many classification models available to use in Weka. Experimenting with all possible models is beyond the scope of this study. Therefore, we only chose the three most representative classifiers from four major categories to experiment with in this paper.

The four classifier categories included Bayes , function-based , lazy model and tree classifiers. In each category, we further selected three classification models [Table 2]. Previous studies on classifying medical documents showed that the performance variation among different models could be relatively small within the category compared with across the categories.[15] This was another main reason we only chose to experiment with three classifiers rather than all available classifiers from each category.

There are different ways to configure a classifier. In most cases of our experiments, the default configuration was chosen as long as the setting was compatible with the type of data we classify. For example, for LibSVM the standard algorithm of regularized C-support vector classification was selected as well as the default radial kernel type [Table 2].[16]

Variable configuration

For each one of the twelve classification models [Table 2], we also experimented with the use of the eight different variable configurations (V1–V8) [Table 1] to find the best variable configuration to produce the best possible classifier. Figure 1 is an overview of the architectural design of our experiments. Each variable configuration was used as the inputs to each one of the twelve classification models. Resulting in twelve experiments for each variable configuration. The classification model with the best performance was chosen as the best model for the underlying variable configuration used. The eight best performing models were reported [Figure 1].

Feature selection

Feature selection (or variable selection) is critical for optimizing a classification model. Appropriately selected features could minimize the chance of overfitting as well as reduce feature redundancy and training time in the face of high-dimensional variables.[17] A typical feature selection process included two separate steps available

Table 2: Selected classification models for experiments

Category	Algorithm
Bayes	BayesNet, NaiveBayes, NaiveBayesUpdatable
Function	LibSVM, Logistic, SMO
Lazy	IBK, KStar, LWL
Tree	ADTree, J48, RandomForest

in the Weka tool we used. The first step is to define a feature evaluation method using criteria based on which features are selected. The second step is to define a search method where features will be searched and selected into the final feature set. There is not a single rule of thumb for choosing the best possible feature selection method. After performing several experiments, we found the default feature selection method in Weka worked the best for us.

The feature selection model we used implemented a correlation-based feature evaluation model called CfsSubsetEval. The CfsSubsetEval feature evaluation model ranks the features based on their correlation with the class while minimizing the redundancy of variables.[18] The underlying feature search model was the BestFirst feature search model which searches the attribute space by allowing backtracking.[18,19] Detailed discussion on CfsSubsetEval and BestFirst is beyond the scope of this paper and we recommend researchers to refer to the original work from the original authors.

Classifier Evaluation

Each classifier was validated using a 10-fold cross-validation method on our entire dataset. Confirmed cases and controls were provided by our clinicians. The best model with the highest area under the curve (AUC) for each variable configuration was reported. AUC was a commonly used metric for machine learning-based evolution as a chance corrected measure and independent of class distributions.

Chi-square test was conducted to test the statistical association between each of the variables and celiac status. In addition to AUC, we also reported precision, recall, F1, and Kappa. Precision, recall, and F1 were classic machine learning measures for evaluating overall accuracy with 1 being the best and 0 being the worst for predicting positive cases. The Kappa score measured the level of agreement between the user and the classifier with 1 being complete agreement and 0 being complete disagreement.

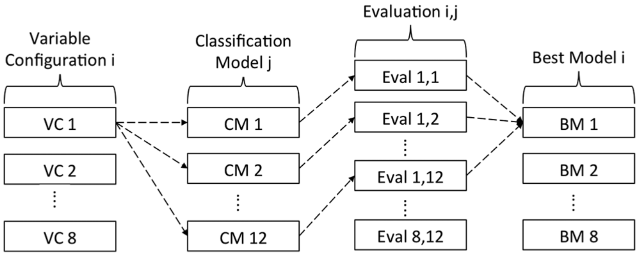

Figure 1: Overview of the architectural design of experiments

RESULTS

Our data extraction resulted in identification of 1498 unique patients with ICD-9 codes suggesting possible CD. We found 363 confirmed cases (positive celiac cases [Table 3]) matching the previously establish celiac registry and 1135 false positive cases (negative celiac cases [Table 3]). Demographic data are shown in Table 3. Chi-square test found statistically significant associations between age, gender, race, and celiac status. However, ethnicity was not significant factor for differentiating CD. Classification variables such as those about visit, laboratory and pathology report counts and laboratory results were all found to be significant for differentiating CD patients. As this study mainly focused on using clinical variables for celiac classification we did not include any demographic variables in building classifiers [Table 3].

Table 4 listed the results summarizing the best classification model out of each variable configuration. Models were ranked based on AUC score from the

Table 3: Patient characteristics and their association with celiac disease

Group	Characteristics	Positive celiac (363 cases)	Negative celiac (1135 cases)	$P^{a,b}$
Age[a]	Age	11 (7, 15)	14 (10, 17)	<0.001
Count[a]	Number of ICD-9	4 (3, 7)	2 (1, 6)	<0.001
	Number of laboratories	2 (2, 3)	1 (0, 2)	0.001
	Number of reports	1 (1, 1)	0 (0, 0)	0.001
Gender[b] (%)	Female	230 (27)	612 (73)	0.002
	Male	133 (20)	523 (80)	
Hispanic[b] (%)	Nonhispanic	361 (24)	1121 (76)	0.215
	Hispanic	2 (13)	14 (87)	
Race[b] (%)	White	334 (25)	980 (75)	0.002
	Nonwhite	29 (16)	155 (84)	
Laboratory results[b] (%)	High	250 (81)	59 (19)	<0.001
	Normal	22 (32)	46 (68)	
	Number of laboratories[c]	91 (8)	1030 (92)	no labs

[a]For continuous variables in the groups of age and count, we reported medium and interquartile range. The *P* value is the significance level based on two-sample *t*-test, [b]For categorical variables in the groups of gender, ethnicity, race and lab results, we reported the number of cases and in-group percentages. The *P* value is the significance level based on Fisher's exact test, [c]The patient has not done any labs in the past 6 months. ICD-9: International Classification of Diseases-9

highest to the lowest. Among all twelve classification models, we experimented with the LibSVM was found to be the top performer in five experiments (experiments 3, 4, 5, 6, 8), followed by Logistic model in two experiments (experiments 1, 7) and Naïve Bayes in one experiment (experiment 2). The best performing model was the Logistic model using all available features filtered by an automatic feature selection process. It achieved AUC of 0.94, Kappa of 0.78, and F1 of 0.92 [Table 4].

Among experiments using single category of variables (experiment 2 and 4–8), Naïve Bayes model based on automatically selected n-gram features generated the best results: AUC of 0.92, Kappa of 0.73, and F1 of 0.90, which was also the overall second best performing model. Here, we also conducted feature selection on executing the Naïve Bayes model as we aimed to find the best possible performance of using text features. Despite the slightly lower performance, the Naïve Bayes model was much faster than other models in terms of model building and model evaluation time. In our case, it only took Naïve Bayes a few seconds to build and evaluate the model in 10-fold cross-validation while it took a few minutes for LibSVM and Logistic model to finish in the case of using all 4000 features.

Using only high-risk phrases, number of laboratories and number of ICD-9 (experiment 6, 7, 8) resulted in the lowest model accuracy. Using laboratory resulted in higher accuracy than the previous three. It was also intriguing that using the number of pathology reports alone only slightly underperformed using report text features alone. However, additional features resulted in higher accuracy given the results based on a combination of different features.

Table 5 listed some of the automatically selected n-gram phrases in contrast to user-provided high-risk phrases. It was found that automatically created n-gram features not only covered the majority of user-provided high-risk phrases but also had the potential of being more accurate and specific in finding relevant information. For example, automatically generated n-grams such as celiac and abdominal pain were directly relevant features to CD but were not initially on our clinicians' list.

Some of the provided high-risk features were not included in our automatically selected feature list such as the phrase of Marsh lesion (converted to lower case for processing). We found this may be due to the extreme low frequency of the usage of such phrases in our pathology reports (only 1 out of all 363 confirmed cases had this exact phrase). In Table 5, the italicized phrases highlighted some of the overlaps between the two lists: user-provided and automatically extracted. The automatically extracted list only represented a subset of most relevant n-gram features [Table 5].

DISCUSSION

Our machine learning approach could quickly and effectively choose the most relevant variables from various clinical sources for automated CD classification. The level of accuracy we achieved was higher compared with previous studies on the same topic.[7,15] N-gram features performed well in our CD detection experiment which agreed with and further improved the results from previous studies using text features.[7,15] In contrast to previous studies, our automated feature generation and selection approaches were more efficient in finding relevant classification features in a short period of time and more effective than user-provided high-risk phrases. This result showed the automated list was more effective in providing a comprehensive list of relevant keywords than manual methods.

Table 4: Statistics of the best classification models for each variable configuration

ID	Feature set	AFS[a]	Number of features	Best model	Kappa	Precision	Recall	F1	AUC
1	All-2[a]	Yes	70	Logistic	0.78	0.93	0.92	0.92	0.94
2	N-gram[a]	Yes	60	Naive Bayes	0.73	0.91	0.90	0.90	0.92
3	All-1	No	4000	LibSVM	0.74	0.91	0.91	0.91	0.89
4	Number of reports	No	1	LibSVM	0.70	0.90	0.88	0.89	0.88
5	Lab results	No	1	LibSVM	0.68	0.89	0.89	0.89	0.83
6	High-risk phrases	No	9	LibSVM	0.67	0.88	0.89	0.88	0.81
7	Number of labs	No	1	Logistic	<0.01	0.62	0.75	0.71	0.71
8	Number of ICD-9	No	1	LibSVM	0.28	0.74	0.76	0.75	0.63

[a]Automatic feature selection. ICD-9: International Classification of Diseases-9, AUC: Area under the curve

Table 5: Comparison of user-defined high-risk phrases and automated key phrases

User-provided high-risk phrases	Automatically selected features
Flattening villi, gland hyperplasia, intraepithelial lymphocytes, Marsh gland stage, Marsh lesion, Marsh s3 lesion, shortened villi, villous blunting, villous atrophy	Abdominal pain, abnormal, acute, blunting, celiac, correlation, disease, edema, elevated, hyperplasia, increased intraepithelial, intraepithelial lymphocytes, lymphocytes, villous, villous atrophy, villous blunting, with biopsy, with villous

The number of laboratories done was as effective as the actual lab results reported in our case for CD classification (AUC of 0.71 vs. 0.83). This could be because an increased number of labs done or the fact that the laboratory was ordered may suggest a higher likelihood of abnormal laboratory result. Our experiments also agreed with previous studies on the fact that ICD-9 code alone produced very poor results (Kappa of 0.28, F1 of 0.75 and AUC of 0.63).[7] Therefore, we suggested researchers only use ICD-9 code as a prescreening tool for subsetting patients as in this study.

In addition, our results added further evidence to previous studies that Naïve Bayes model generated the most accurate results even if only pathology report features were used (Kappa of 0.73, F1 of 0.90, and AUC of 0.92).[15] This further indicates that pathology report, as an important pierce of synoptic reporting of Celiac status, contains invaluable information for Celiac classification. In addition, by combining both clinical and text features altogether, classification accuracy and Kappa score could be further improved (Kappa of 0.78, F1 of 0.92, and AUC of 0.94). This suggests the complementary effect of using both structured data and unstructured data for Celiac classification.

Finally, feature selection was found to be a crucial step in celiac detection for further improving the efficiency and accuracy of classification. Feature selection dramatically removed redundant and noninformative terms (from about 4000 to 60). Although new to the task of CD classification, feature selection methods have been found widely used in other disease classification studies including Alzheimer's disease[20] and asthma.[21]

In our case, some variables (i.e., number of pathology reports and lab results) could individually predict celiac at a reasonably high accuracy. However, this result may be problem dependent. Other studies implementing the same metrics may or may not result in similar performance. Even that, we still observed performance gain by adding NLP features with both accuracy and agreement increased by a reasonable amount.

The automated tool we developed could speed up the process of refining the subset population initially identified based on ICD-9 code 579.0. This refining process can automatically confirm the patient as either positive or negative CD case based on the knowledge of the patient obtained from the laboratories and pathology notes. The classification accuracy was over 90% of correctness (F1) in our case while it greatly eliminated the need of often time-consuming process of manual review.

Limitations

This study has several limitations. First, since the CD classification task was performed on a subset of patients prescreened by ICD-9 code rather than on the general patient population, the performance of final classification results may depend on the accuracy of initial ICD-9 assignment. For example, if actual CD patients were not assigned an ICD-9 code 579.0 initially, they will not have the chance of being further identified by our machine learning system. It is, therefore, important to make sure the initial ICD-9 code assignment will include as many high-risk cases as possible.

Second, although feature selection method was found to be critical in yielding high performance, we have not conducted systematic experiments with all possible feature selection methods provided in the Weka machine learning system. Given our experience, it was likely that some methods may further improve accuracy while others may even lower the accuracy.

Finally, for most machine learning algorithms we experimented, only the default configuration was used for classification. It was likely different configurations of parameter setting may lead to very different results of accuracy. Therefore, researchers are encouraged to experiment with different classifier settings to achieve the best possible results for their tasks.

CONCLUSION

In this paper, we compared results from 96 machine-learning experiments for CD identification based on 12 classification model variations and eight feature set variations. The logistic model that used a combination of clinical and pathology report features generated the best results: Kappa of 0.78, F1 of 0.92, and AUC of 0.94. Our results were in agreement with previous studies on the insufficiency of using ICD-9 codes alone and the merits of using pathology report features for CD identification. This study provides new evidence on adopting feature selection techniques to further improve classification efficiency and accuracy based on a subpopulation of prescreened patients using ICD-9 code.

This study demonstrated a viable computational approach to automatically reviewing and confirming prescreened CD patients based on ICD-9 code at the state of the art accuracy much improved from previous research on the same topic.

Financial Support and Sponsorship

Grant UL1TR001070 from Center for Clinical and Translational Science at the Ohio State University.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. Rubio-Tapia A, Ludvigsson JF, Brantner TL, Murray JA, Everhart JE. The prevalence of celiac disease in the United States. Am J Gastroenterol

2012;107:1538-44.

2. Gidrewicz D, Potter K, Trevenen CL, Lyon M, Butzner JD. Evaluation of the ESPGHAN celiac guidelines in a North American Pediatric Population. Am J Gastroenterol 2015;110:760-7.
3. Trovato CM, Montuori M, Anania C, Barbato M, Vestri AR, Guida S, *et al.* Are ESPGHAN "biopsy-sparing" guidelines for celiac disease also suitable for asymptomatic patients? Am J Gastroenterol 2015;110:1485-9.
4. Wooldridge JL, Mason S, Brusatti J, Albers GM, Noyes BE. Improvements in cystic fibrosis quarterly visits, lung function tests, and respiratory cultures. Pediatrics 2015;136:e1611-6.
5. Schechter MS, Fink AK, Homa K, Goss CH. The cystic fibrosis foundation patient registry as a tool for use in quality improvement. BMJ Qual Saf 2014;23 Suppl 1:i9-14.
6. Crandall WV, Margolis PA, Kappelman MD, King EC, Pratt JM, Boyle BM, *et al.* Improved outcomes in a quality improvement collaborative for pediatric inflammatory bowel disease. Pediatrics 2012;129:e1030-41.
7. Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, *et al.* Use of computerized algorithm to identify individuals in need of testing for celiac disease. J Am Med Inform Assoc 2013;20:e306-10.
8. Wieneke AE, Bowles EJ, Cronkite D, Wernli KJ, Gao H, Carrell D, *et al.* Validation of natural language processing to extract breast cancer pathology procedures and results. J Pathol Inform 2015;6:38.
9. Raju GS, Lum PJ, Slack RS, Thirumurthi S, Lynch PM, Miller E, *et al.* Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointest Endosc 2015;82:512-9.
10. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin Gastroenterol Hepatol 2013;11:689-94.
11. Kim BJ, Merchant M, Zheng C, Thomas AA, Contreras R, Jacobsen SJ, *et al.* A natural language processing program effectively extracts key pathologic findings from radical prostatectomy reports. J Endourol 2014;28:1474-8.
12. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD Explor Newsl 2009;11:10-8.
13. Wu ST, Liu H, Li D, Tao C, Musen MA, Chute CG, *et al.* unified medical language system term occurrences in clinical notes: A large-scale corpus analysis. J Am Med Inform Assoc 2012;19:e149-56.
14. Zuccon G, Koopman B, Nguyen A, Vickers D, Butt L. Exploiting Medical Hierarchies for Concept-based Information Retrieval. Proceedings of the Seventeenth Australasian Document Computing Symposium; 2012. p. 111-4.
15. Tenório JM, Hummel AD, Cohrs FM, Sdepanian VL, Pisa IT, de Fátima Marin H. Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease. Int J Med Inform 2011;80:793-802.
16. Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, *et al*. WEKA Manual for Version 3-7-8 2013. Available from: http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf. [Last accessed on 2016 Oct 08].
17. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, *et al.* Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc 2015;22:993-1000.
18. Hall MA. Correlation-based Feature Selection for Machine Learning: The University of Waikato; 1999. Available from: https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf. [Last accessed on 2016 Oct 08].
19. Rich E, Knight K. Introduction to Artificial Intelligence. New York City, United States: McGraw-Hill; 1991.
20. Zhang D, Shen D; Alzheimer's disease neuroimaging initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 2012;59:895-907.
21. Wu W, Bleecker E, Moore W, Busse WW, Castro M, Chung KF, *et al.* Unsupervised phenotyping of severe asthma research program participants using expanded lung data. J Allergy Clin Immunol 2014;133:1280-8.