JMIR Publications
The Leading eHealth Publisher

Automated identification of aspirin exacerbated respiratory disease using natural language processing and machine learning

Thanai Pongdee, Nicholas Larson, Rohit Divekar, Suzette Bielinski, Hongfang Liu, Sungrim Moon

Submitted to: JMIR AI
on: November 09, 2022

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

Automated identification of aspirin exacerbated respiratory disease using natural language processing and machine learning

Thanai Pongdee[1] MD; Nicholas Larson[2] PhD, MS; Rohit Divekar[1] MBBS, PhD; Suzette Bielinski[3] PhD; Hongfang Liu[4] PhD; Sungrim Moon[4] PhD

[1]Division of Allergic Diseases Mayo Clinic Rochester US
[2]Division of Clinical Trials and Biostatistics Department of Quantitative Health Sciences Mayo Clinic Rochester US
[3]Division of Epidemiology Department of Quantitative Health Sciences Mayo Clinic Rochester US
[4]Department of Artificial Intelligence and Informatics Mayo Clinic Rochester US

Corresponding Author:
Thanai Pongdee MD
Division of Allergic Diseases
Mayo Clinic
200 First Street SW
Rochester
US

Preprint JMIR Publications

Abstract

Background: Aspirin exacerbated respiratory disease (AERD) is an acquired inflammatory condition characterized by the presence of asthma, chronic rhinosinusitis with nasal polyposis, and respiratory hypersensitivity reactions on ingestion of aspirin or other nonsteroidal anti-inflammatory drugs (NSAIDs). Despite AERD having a classic constellation of symptoms, the diagnosis is often overlooked, with an average of greater than ten years between the onset of symptoms and diagnosis of AERD. Without a diagnosis, individuals will lack opportunities to receive effective treatments such as aspirin desensitization or biologic medications.

Objective: To develop a natural language processing (NLP) algorithm to identify patients with AERD from an electronic health record (EHR).

Methods: A rule-based NLP algorithm was developed using clinical documents from the EHR at Mayo Clinic. From clinical notes, seven features were extracted that included the following: AERD, asthma, NSAID allergy, nasal polyps, chronic sinusitis, elevated urine leukotriene E4 level, and documented no-NSAID allergy by a health care provider. MedTagger was used to extract these seven features from the unstructured clinical text given a set of keywords and patterns based on the chart review of two allergy/immunology experts for AERD. The status of each extracted feature was represented as either present or absent per subject. To determine the representative combination of features to discriminate the different AERD features, we utilized the entropy approach of the decision tree classifier.

Results: The AERD-NLP algorithm achieved an accuracy of 88.00% (95% CI=[82.10, 91.74]) and 84.50% (95% CI=[78.73, 89.22]) for the training and test set, respectively.

Conclusions: We developed a promising AERD-NLP algorithm that needs further refinement to improve AERD diagnosis accuracy. Continued development of NLP and other artificial intelligence technologies has the potential to reduce diagnostic delays for AERD and improve the health of our patients.

(JMIR Preprints 09/11/2022:44191)
DOI: https://doi.org/10.2196/preprints.44191

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users. Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint JMIR Publications

Short paper

Automated identification of aspirin exacerbated respiratory disease using natural language processing and machine learning

Abstract

Background: Aspirin exacerbated respiratory disease (AERD) is an acquired inflammatory condition characterized by the presence of asthma, chronic rhinosinusitis with nasal polyposis, and respiratory hypersensitivity reactions on ingestion of aspirin or other nonsteroidal anti-inflammatory drugs (NSAIDs). Despite AERD having a classic constellation of symptoms, the diagnosis is often overlooked, with an average of greater than ten years between the onset of symptoms and diagnosis of AERD. Without a diagnosis, individuals will lack opportunities to receive effective treatments such as aspirin desensitization or biologic medications.
Objective: To develop a combined algorithm that integrates both natural language processing (NLP) and machine learning (ML) techniques to identify patients with AERD from an electronic health record (EHR).
Methods: A rule-based decision tree algorithm incorporating NLP-based features was developed using clinical documents from the EHR at Mayo Clinic. From clinical notes, using NLP techniques, seven features were extracted that included the following: AERD, asthma, NSAID allergy, nasal polyps, chronic sinusitis, elevated urine leukotriene E4 level, and documented no-NSAID allergy. MedTagger was used to extract these seven features from the unstructured clinical text given a set of keywords and patterns based on the chart review of two allergy/immunology experts for AERD. The status of each extracted feature was quantified by assigning the frequency of its occurrence in clinical documents per subject. We optimized the decision tree classifier's hyperparameters cutoff threshold on the training set to determine the representative feature combination to discriminate AERD. We then evaluated the resulting model on the test set.
Results: The AERD algorithm, which combines NLP and ML techniques, achieved an area under the ROC curve score, sensitivity and specificity of 0.86 (95% CI=[0.78, 0.94]), 80.00 (95% CI=[70.82, 87.33]) and 88.00 (95% CI=[79.98, 93.64]) for the test set, respectively.
Conclusions: We developed a promising AERD algorithm that needs further refinement to improve AERD diagnosis. Continued development of NLP and ML technologies has the potential to reduce diagnostic delays for AERD and improve the health of our patients.

Keywords

aspirin exacerbated respiratory disease; natural language processing; electronic health record; identification; machine learning

Introduction

Aspirin exacerbated respiratory disease (AERD) is an acquired inflammatory condition characterized by the presence of asthma, chronic rhinosinusitis with nasal polyposis, and respiratory hypersensitivity reactions on ingestion of aspirin or other nonsteroidal anti-inflammatory drugs (NSAIDs).[1] These reactions typically involve the upper and lower airways and may include nasal congestion, sneezing, rhinorrhea, cough, and/or wheezing.[1] The prevalence of AERD is approximately 0.3-0.9% in the general population, but the actual prevalence is unknown in practice as AERD has no unique International Classification of Diseases, Ninth Revision (ICD-9), or International Classification of Diseases, Tenth Revision (ICD-10) codes.[2, 3] In the general population, the mean age of onset of AERD is approximately 30 years[2, 4], and the prevalence of AERD is estimated to be 7-15% of individuals with asthma and 10-16% of individuals with chronic

rhinosinusitis with nasal polyposis.[5] Individuals with AERD have significant symptom burden and morbidity including severe and recalcitrant sinus disease, high rates of polyp recurrence and revision surgery, and higher asthma exacerbation and hospitalization rates.[1] Despite AERD having a classic constellation of symptoms, the diagnosis is often overlooked, with an average of greater than ten years between the onset of symptoms and diagnosis of AERD.[6], Without a diagnosis, individuals will lack opportunities to receive effective treatments such as aspirin desensitization or biologic medications.[5, 7]

One opportunity to improve diagnostic delays with AERD involves leveraging the immense volume of clinical data available in electronic health records (EHRs). By leveraging natural language processing (NLP) and machine learning (ML), extraction, processing, and analyses of medical concepts from unstructured clinical documents to aid early detection of AERD is needed.[8] In this study, we developed a combined algorithm of NLP with ML to identify individuals with AERD.

Methods

The current study was approved by the Mayo Clinic institutional review board. Patients who were evaluated within the Allergy and Immunology divisions at Mayo Clinic from January 2001 to March 2022 and met diagnostic criteria for AERD based on accepted guidelines[1] were retrospectively identified by chart review. In total, 200 subjects with AERD and 200 subjects without AERD were identified. Of these subjects, we randomly selected 100 subjects with AERD and 100 without AERD to serve as the training set, and the remaining were used for the test set.

A rule-based decision tree algorithm incorporating NLP-based features was developed to identify AERD subjects using clinical documents from the EHR at Mayo Clinic. From clinical notes, seven features were extracted using NLP techniques based on common characteristics of AERD.[1] These features included the following: prior AERD diagnosis, asthma, NSAID allergy, nasal polyps, chronic sinusitis, elevated urine leukotriene E4 level, and documented no-NSAID allergy. “Prior AERD diagnosis” is whether the patient has a diagnosis of AERD before or has suspicion of a high chance of AERD by the physician. For “Asthma,” “Nasal Polyps,” and “Chronic sinusitis,” the patient needs to have a diagnosis confirmation by the physician in the clinical documents. “Elevated urine leukotriene E4 level” indicated the patient has any record in lab results to of a urine leukotriene E4 level greater than 104 pg/mg creatinine. “NSAID allergy” was defined as a subject having had a respiratory reaction to an NSAID. Meanwhile, “documented no-NSAID allergy” indicates that a health care provider recorded “unconfirmed or no specific history of NSAID allergy up to date” in the clinical documents. Given the successful use cases of MedTagger[9] to identify disease in the different clinical domains,[10, 11] we used MedTagger to extract these features with the given set of keywords (including typos, abbreviations and acronyms) and patterns based on the chart review of two allergy/immunology experts for AERD. If the extracted features were located in particular note sections (i.e., “History of Present Illness,” “Allergies,” “Past Medical/Surgical History,” “Impression/Report/Plan,” “Diagnosis,” “Principal Diagnosis,” “Secondary Diagnoses,” and “Post Procedure Diagnosis”), they were considered valid AERD features. We collected each feature in all clinical documents per subject in the past five years from the last clinic visit since clinical characteristics of AERD can evolve over time. (i.e., development of NSAID allergy).

We counted the number of times each extracted feature appeared in the clinical documents for each subject and used this count as the numerical representation of each feature. To identify the most practical combination of features for discriminating between different presentations of AERD, we optimized the hyperparameters of the Classification And Regression Tree (CART) decision tree classifier with the identified features on the training set using Sklearn.[12] We performed hyperparameter tuning on five different parameters with one model setting: (1) criterion, with options of gini or entropy, (2) max depth, ranging from 1 to 10 with an interval of 1, (3) min sample split, ranging from 2 to 10 with an interval of 2, (4) min samples leaf, ranging from 1 to 10 with an interval of 1, (5) max features, ranging from 1 to 7 with an interval of 1, and (6) random state as a model setting, ranging from 1 to 100 with an interval of 10, as well as 1000. Furthermore, to achieve the highest AUC score, these hyperparameters were tuned for two types of feature sets: (1) quantitatively represented as numerical values per subject, and (2) binary, where '1' denotes the

presence and '0' denotes the absence or missing status of each extracted feature per subject. We constructed a decision tree using the best feature set with optimized hyperparameters and then calculated the AUC scores for a range of cutoff thresholds from 0.1 to 1.0 in intervals of 0.1 to determine the optimal cutoff threshold based on a given training set. The resulting tree with the optimized parameters and cutoff threshold converted into sequential rule sets (see Table 2) to evaluate the performance in the test set.

Results

In our cohort, the mean age of the 400 subjects was 55.5 years, and 54% were female. Table 1 displays the descriptive statistics for each feature, comparing the presence or absence of the feature in the train and test sets. Based on the training set, we obtained the sequential rule sets through the optimized decision tree (with criterion as gini, max depth as 7, min samples leaf as 7, min samples split as 2, max features as 3, random state as 20, and best cutoff threshold as 0.6 for parameter settings) using the numerical represented feature set in Table 2. The sequential rules listed in Table 2 describe several clinical factors that include: diagnosis of AERD (AERD), diagnosis of allergy to an NSAID (NSAID allergy), diagnosis of chronic sinusitis, documented history of tolerance to an NSAID (non-NSAID allergy), and a prior abnormally elevated urine leukotriene E4 level (LAB).

Table 1. Descriptive statistics of AERD features, describing the presence as 1 or absence as 0.

AERD Feature	Train		Test	
	Case (n)	% of 200 cases	Case (n)	% of 200 cases
AERD	103	52	60	30
Asthma	192	96	82	41
NSAID allergy	98	49	121	61
nasal polyps	175	88	192	96
chronic sinusitis	182	91	180	90
LAB*	93	47	179	90
Documented no-NSAID allergy	70	35	101	51

* LAB indicates elevated urine leukotriene E4 level.

Table 2. Derived sequential rules for AERD algorithm and resulting performance in the test set

Rule	Sequential rules	AERD	#Case	#Correct	#Error	Confidence%[e]
1	AERD[a]<=3.5, NSAID allergy[b]<=2.5, Chronic Sinusitis[c] <=6.5, and then documented non-NSAID allergy<=0.5	No	30	12	0	40
2	AERD[a]<=3.5, NSAID allergy[b]<=2.5, Chronic Sinusitis[c] <=6.5, and then documented non-NSAID allergy>0.5	No	9	0	0	0
3	AERD[a]<=3.5, NSAID allergy[b]<=2.5, Chronic Sinusitis[c] >6.5, and then documented non-NSAID allergy<=0.5	No	43	40	0	93
4	AERD[a]<=3.5, NSAID allergy[b]<=2.5, Chronic Sinusitis[c] >6.5, and then documented non-NSAID allergy>0.5	No	4	3	0	75
5	AERD[a]<=3.5, NSAID allergy[b]>2.5, and then Chronic Sinusitis[c]<=9.0	Yes	10	0	0	0
6	AERD[a]<=3.5, NSAID allergy[b]>2.5, and then Chronic Sinusitis[c]>9.0	Yes	74	1	0	1
7	AERD[a]>3.5, NSAID allergy[b]<=1.5, and then LAB[d]<=0.5	Yes	0	0	0	0
8	AERD[a]>3.5, NSAID allergy[b]<=1.5, and then LAB[d]>0.5	No	2	0	0	0
9	AERD[a]>3.5, NSAID allergy[b]>1.5	Yes	2	0	0	0
10	Others	Yes	16	0	0	0
		No	10	0	0	0
	The cases were not identified according to the original intended sequential rule; instead, a different sequence rule was used.	Yes	99	79	20	80
		No	45	33	12	73

[a]AERD: diagnosis of aspirin exacerbated respiratory disease; [b]NSAID allergy: diagnosis of allergy to a nonsteroidal anti-inflammatory drug; [c]Chronic Sinusitis: diagnosis of chronic sinusitis; [d]LAB: a prior abnormally elevated urine leukotriene E4 level; [e]Confidence = (numbers of correct cases / numbers of real cases in test set) * 100 for the particular rule from 1 to 9.

In Table 2, it was observed that the derived sequential rule, ranging from 1 to 9, captured 28% of the cases in the test set. However, a significant portion of the test set (56%) was not identified according to the original intended sequential rule but rather by a different sequence rule. For example, rule 6 failed to capture 73 cases, whereas rule 9, which is less strict than rule 6, captured 59 out of those 73 cases that were supposed to belong to rule 6. Similarly, rule 3 captured 15 cases out of the remaining 18 cases that should have been identified by rule 1. Therefore, the overall accuracy is 0.84.

The AERD algorithm achieved an area under the ROC Curve (AUC) score of 0.92 (95% CI =[0.93, 1.00]) and 0.86 (95% CI=[0.78, 0.94], the right side of Fig. 1) for the training and test set, respectively. The optimal cutoff point is 0.6 on the train set in left side of Fig. 1. Additional performances are in Table 3.

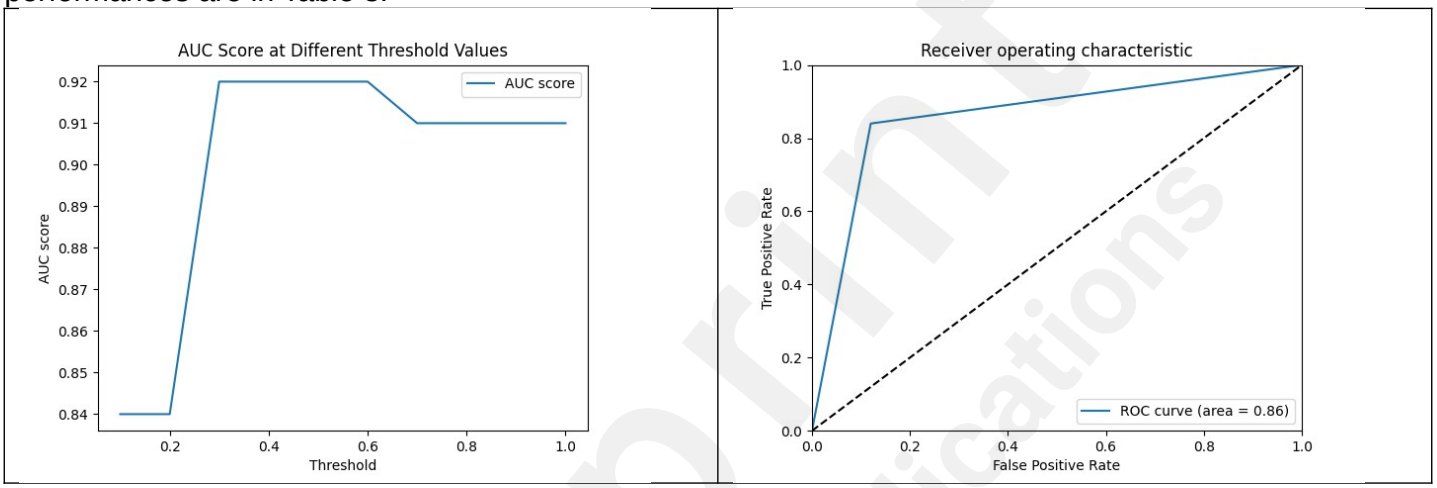


Fig 1. Area Under the Curve (AUC) scores at different threshold value on train set (Left) and Receiver operating characteristic (ROC) on test set (Right).

Table 3. Performance of the rule-based AERD algorithm.

	Sensitivity (%) [95% CI]	Specificity (%) [95% CI]	Positive Predictive Value (%) [95% CI]	Negative Predictive Value (%) [95% CI]	Accuracy (%) [95% CI]
Train	88.00 [79.98, 93.64]	97.00 [91.48, 99.38]	96.70 [90.67, 99.31]	88.99 [81.56, 94.18]	92.50 [87.93, 95.74]
Test	80.00 [70.82, 87.33]	88.00 [79.98, 93.64]	86.96 [78.32, 93.07]	81.48 [72.86, 88.31]	84.00 [78.17, 88.79]

Discussion

In our study, we demonstrate that an algorithm, which combines NLP and ML techniques, can identify patients with AERD with a positive predictive value of approximately 86.96 and a negative predictive value of 81.48. Our results are comparable to prior work[3] of automated diagnosis of AERD from EHR data using structured query language statements for data analysis that resulted in positive predictive values ranging from 78.4 to 88.7, depending on the cohort being analyzed.

Prior diagnosis of AERD presents the highest impacted feature (i.e., majority of sequential rules contain Prior diagnosis of AERD feature) to detect diagnosis of AERD. In the training and test sets, 85% and 91% of subjects with AERD had a prior diagnosis of AERD, respectively. We also extracted new clinical factors associated with AERD (“elevated urine leukotriene E4 level” and “alcohol intolerance”) that were not used previously studied. Furthermore, the “elevated urine leukotriene E4 level” feature may need to be considered as a new meaningful feature associated with AERD because the presence of the term “AERD” with an “elevated leukotriene E4 level” was a common feature of rule sets 7 and 8. Most AERD subjects in the test set were accurately identified by having had an AERD diagnosis and a documented NSAID allergy (See Table 2).Lastly, diagnosis of nasal polyps was not used to construct the optimal decision tree, which may indicate that it may

be an insignificant feature to distinguish AERD subjects from possible AERD candidates.

The test set included 32 errors from 200 subjects, which upon review, were due to either unidentified rule sets for AERD subjects (n=11) or missing/incorrect feature extraction due to unseen keywords/patterns for features (n=9) primarily. For example, the sentence: “Patient took an aspirin approximately ten years ago for headache and developed a sensation of pressure in his nose and sinuses” is an unseen pattern for prior AERD features. Based on the expression, “a sensation of pressure in his nose and sinuses,” the sentence should be a prior AERD feature; however, AERD algorithm categorized it as absent of an AERD feature because this pattern was not available in the training phrase. Six subjects had necessary feature information beyond the past five years of clinical documents from the last visit day. Six subjects had necessary information belonging to an unknown note section in the training set for feature extraction. When examining the specific rules, rule sets 2-3 resulted in very few errors (see Table 2). In contrast, the absence of terms explicitly documenting the absence of NSAID allergy and the lack of references to an elevated leukotriene E4 level resulted in more errors in the AERD algorithm.

Diagnosing and confirming AERD may be a prolonged process, as the associated clinical features may present at different times in a variety of time sequences. As a result, there is no solid ICD code (structured data) to represent AERD, and AERD-associated clinical characteristics are often undocumented in clinical texts (unstructured data) in the EHR. This lack of information regarding AERD results in the low quality of data sources and potential bias for ML models.[14] Additional efforts (e.g., standardizing routine exams for AERD) are necessary to fill these missing information gaps in practice.

This AERD algorithm has limitations in deploying to detect confirmed AERD patients in a practical setting without further refinement. We focused on identifying feature selections in the limited parameter tuning using a balanced dataset (N=200 for AERD patients and N=200 for patients without AERD), which was not a real-world situation. We used the minimum sample size due to the nature of AERD, which has a low prevalence. The rule-based algorithm is employed because the limited sample and feature set provide high interpretability and accuracy at downstream tasks rather than neural network MLs, which require a large training dataset. However, this algorithm provides a valuable contribution to capturing potential AERD patients with a pre-requisite as a large cohort in the EHR because the prevalence of AERD patients is low in clinical settings. To follow up, we plan to rank features with diverse identified feature sets and parameter tuning for the decision tree model within a large cohort. We will investigate our new feature in the EHR, which is information about urine leukotriene E4 levels in the extensive feature selections as well as explore additional features for AERD (e.g., alcohol sensitivity, anosmia, and prior sinus surgeries).

Conclusions

We developed an AERD algorithm, which combines NLP and ML techniques, to enhance AERD diagnosis in practice. On top of prior work[3], we utilized NLP with a potential feature, urine leukotriene E4 levels from EHR, which have been shown to aid AERD diagnosis.[15] Leveraging NLP and ML techniques in practice has the potential to reduce diagnostic delays for AERD and improve the health of patients.

Acknowledgments

All listed authors had substantial contributions to the design, acquisition, analysis, and interpretation of data for the work. In addition, all listed authors contributed to the drafts, revisions, and approval of the final version to be published.

Funding

This work was supported by grants from the National Heart, Lung and
Blood Institute (R01 HL136659) and the resources of the Rochester Epidemiology Project (REP)

medical records-linkage system, which is supported by the National Institute on Aging (NIA; AG 058738), by the Mayo Clinic Research Committee, and by fees paid annually by REP users. The funding sources played no role in the design, conduct, or reporting of this study. The content of this article is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health (NIH) or the Mayo Clinic.

Conflicts of interest

Authors have nothing to disclose and report no conflicts of interest.

Abbreviations

AERD:	aspirin exacerbated respiratory disease
EHR:	electronic health record
ICD:	International Classification of Diseases
NLP:	natural language processing
NSAID:	nonsteroidal anti-inflammatory drug
ML:	machine learning

References

1. Haque R, White AA, Jackson DJ, Hopkins C. Clinical evaluation and diagnosis of aspirin-exacerbated respiratory disease. J Allergy Clin Immunol. 2021;148(2):283-91.
2. Szczeklik A, Nizankowska E, Duplaga M. Natural history of aspirin-induced asthma. AIANE Investigators. European Network on Aspirin-Induced Asthma. Eur Respir J. 2000;16(3):432-6.
3. Cahill KN, Johns CB, Cui J, Wickner P, Bates DW, Laidlaw TM, et al. Automated identification of an aspirin-exacerbated respiratory disease cohort. J Allergy Clin Immunol. 2017;139(3):819-25 e6.
4. Berges-Gimeno MP, Simon RA, Stevenson DD. The natural history and clinical characteristics of aspirin-exacerbated respiratory disease. Ann Allergy Asthma Immunol. 2002;89(5):474-8.
5. Stevens WW, Jerschow E, Baptist AP, Borish L, Bosso JV, Buchheit KM, et al. The role of aspirin desensitization followed by oral aspirin therapy in managing patients with aspirin-exacerbated respiratory disease: A Work Group Report from the Rhinitis, Rhinosinusitis and Ocular Allergy Committee of the American Academy of Allergy, Asthma & Immunology. J Allergy Clin Immunol. 2021;147(3):827-44.
6. Lee-Sarwar K, Johns C, Laidlaw TM, Cahill KN. Tolerance of daily low-dose aspirin does not preclude aspirin-exacerbated respiratory disease. J Allergy Clin Immunol Pract. 2015;3(3):449-51.
7. Buchheit KM, Laidlaw TM, Levy JM. Immunology-based recommendations for available and upcoming biologics in aspirin-exacerbated respiratory disease. J Allergy Clin Immunol. 2021;148(2):348-50.
8. Khoury P, Srinivasan R, Kakumanu S, Ochoa S, Keswani A, Sparks R, et al. A Framework for Augmented Intelligence in Allergy and Immunology Practice and Research-A Work Group Report of the AAAAI Health Informatics, Technology, and Education Committee. J Allergy Clin Immunol Pract. 2022;10(5):1178-88.
9. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med. 2019;2:130.
10. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Natural language

processing of clinical notes for identification of critical limb ischemia. Int J Med Inform. 2018;111:83-9.
11. Moon S, Liu S, Scott CG, Samudrala S, Abidian MM, Geske JB, et al. Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. Int J Med Inform. 2019;128:32-8.
12. Decision trees: scikit-learn. Accessed 5/1/2023, at https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.
13. Breiman L, Friedman J, Olshen R, Stone C. Cart. Classification and regression trees. 1984.
14. Moon S, Carlson LA, Moser ED, Agnikula Kshatriya BS, Smith CY, Rocca WA, et al. Identifying Information Gaps in Electronic Health Records by Using Natural Language Processing: Gynecologic Surgery History Identification. J Med Internet Res. 2022;24(1):e29015.
15. Divekar R, Hagan J, Rank M, Park M, Volcheck G, O'Brien E, et al. Diagnostic Utility of Urinary LTE4 in Asthma, Allergic Rhinitis, Chronic Rhinosinusitis, Nasal Polyps, and Aspirin Sensitivity. J Allergy Clin Immunol Pract. 2016;4(4):665-70.

Preprint JMIR Publications

Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/b994312bdb5c6dbd6ab3262c3caa5a88.doc

Untitled.
URL: http://asset.jmir.pub/assets/a6baeece249dcbb7f0cb811fd678520d.doc

Untitled.
URL: http://asset.jmir.pub/assets/0de0e26ad5410269c8e5e43f2eb1777a.docx

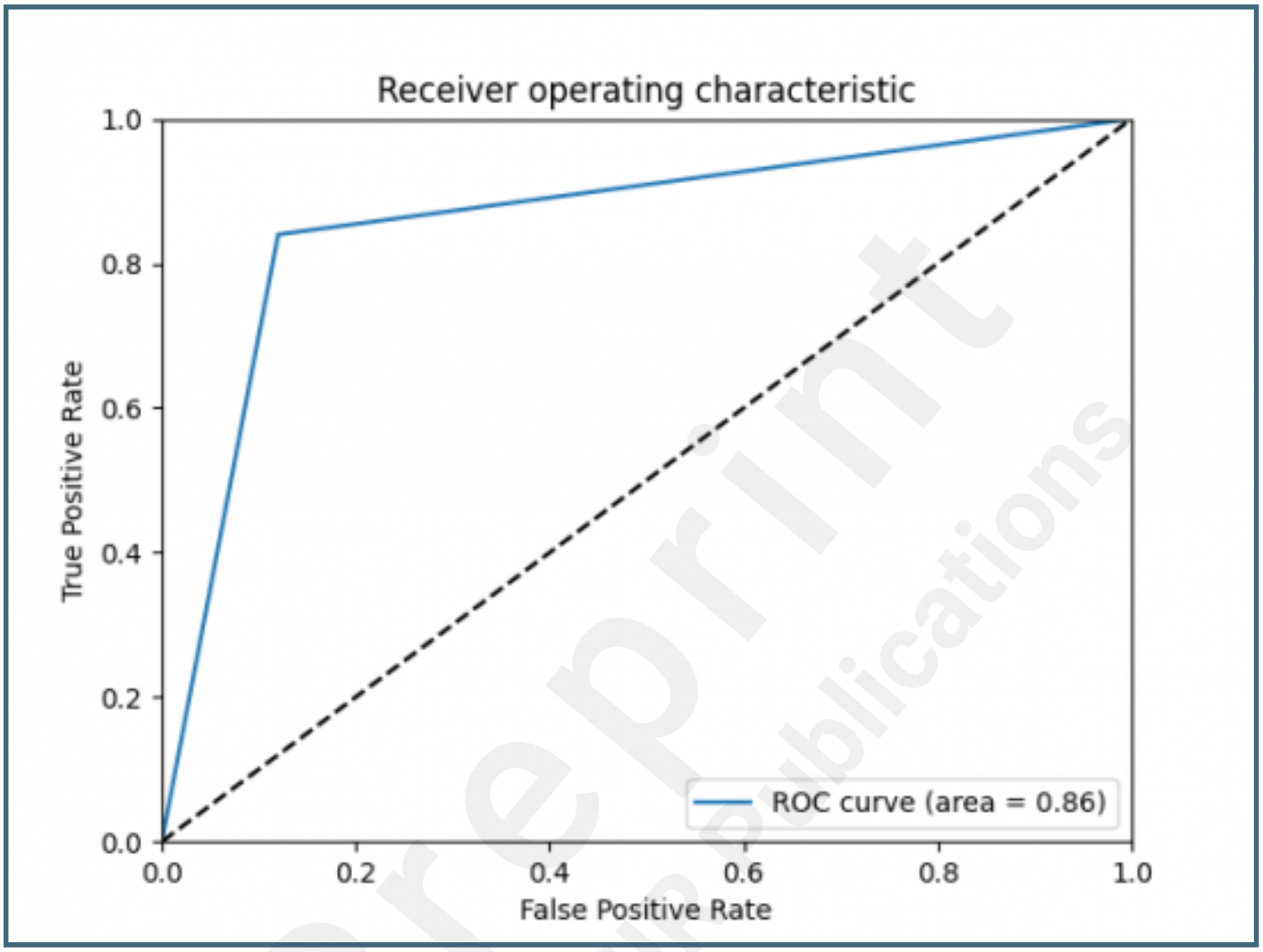
Untitled.
URL: http://asset.jmir.pub/assets/5679efc4fdf4d561d910557efffc4e11.docx

Preprint JMIR Publications

Figures

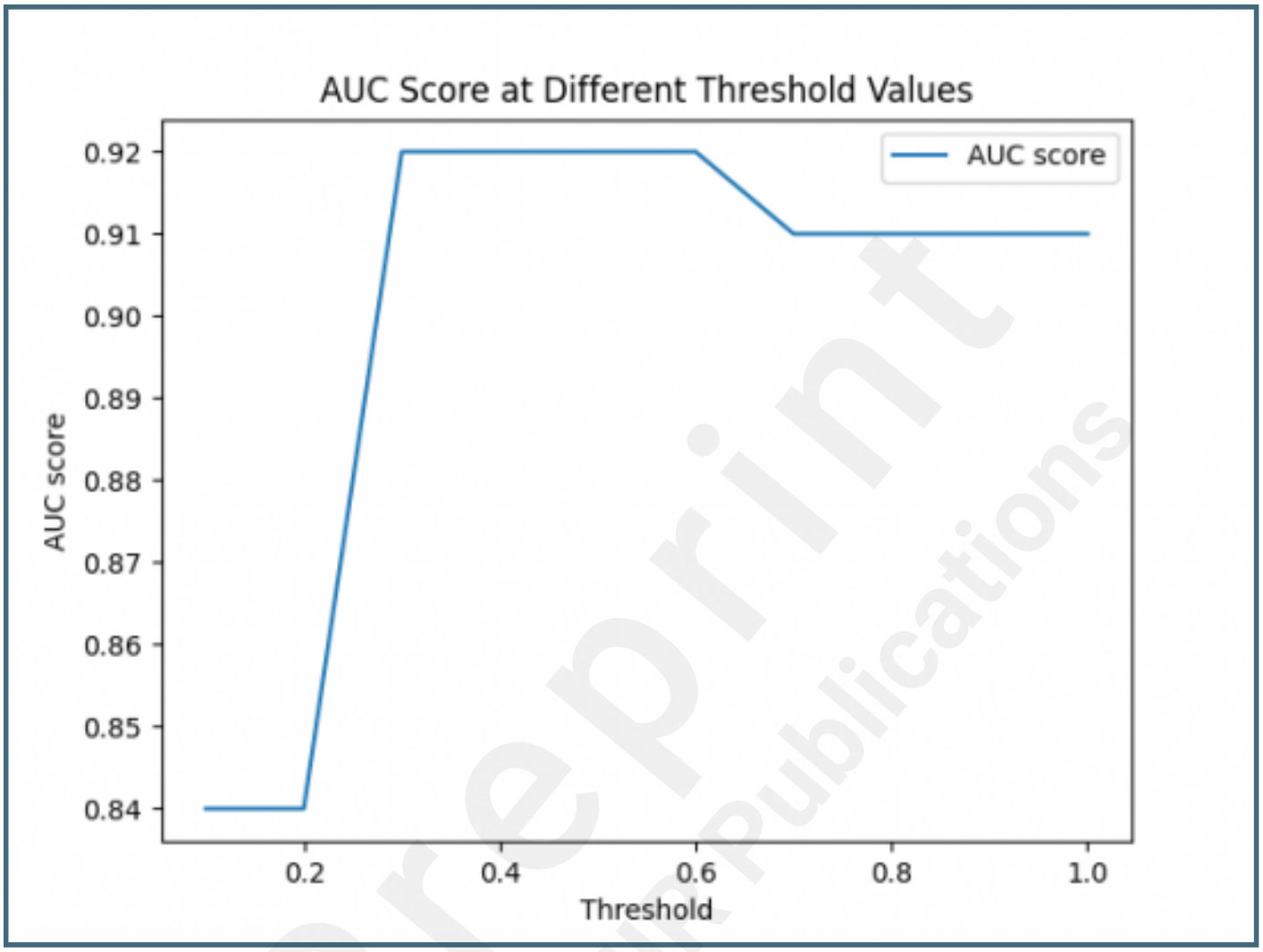
Preprint JMIR Publications

Area Under the Curve (AUC) scores at different threshold value on train set (Left) and Receiver operating characteristic (ROC) on test set (Right).



Preprint JMIR Publications

Area Under the Curve (AUC) scores at different threshold value on train set (Left) and Receiver operating characteristic (ROC) on test set (Right).



Preprint
JMIR Publications