Journal of Pain Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Using Natural Language Processing and Machine Learning to Identify Opioids in Electronic Health Record Data

Sean P McDermott, Ajay D Wasan

Division of Pain Medicine, Department of Anesthesiology and Perioperative Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

Correspondence: Sean P McDermott, UPMC Pain Medicine, 5750 Centre Ave, Suite 400, Pittsburgh, PA, 15232, USA, Tel +1 412-665-8030, Fax +1 412-647-2993, Email seanmcd@umich.edu

Purpose: This study evaluates the utility of machine learning (ML) and natural language processing (NLP) in the processing and initial analysis of data within the electronic health record (EHR). We present and evaluate a method to classify medication names as either opioids or non-opioids using ML and NLP.

Patients and Methods: A total of 4216 distinct medication entries were obtained from the EHR and were initially labeled by human reviewers as opioid or non-opioid medications. An approach incorporating bag-of-words NLP and supervised ML classification was implemented in MATLAB and used to automatically classify medications. The automated method was trained on 60% of the input data, evaluated on the remaining 40%, and compared to manual classification results.

Results: A total of 3991 medication strings were classified as non-opioid medications (94.7%), and 225 were classified as opioid medications by the human reviewers (5.3%). The algorithm achieved a 99.6% accuracy, 97.8% sensitivity, 94.6% positive predictive value, F1 value of 0.96, and a receiver operating characteristic (ROC) curve with 0.998 area under the curve (AUC). A secondary analysis indicated that approximately 15–20 opioids (and 80–100 non-opioids) were needed to achieve accuracy, sensitivity, and AUC values of above 90–95%.

Conclusion: The automated approach achieved excellent performance in classifying opioids or non-opioids, even with a practical number of human reviewed training examples. This will allow a significant reduction in manual chart review and improve data structuring for retrospective analyses in pain studies. The approach may also be adapted to further analysis and predictive analytics of EHR and other "big data" studies.

Keywords: retrospective, classification, machine learning, natural language processing, translational, opioids

Introduction

With advances in healthcare, the quantity of available clinical and research data is growing rapidly.[1–3] Utilizing these large datasets in "big data" studies may allow for powerful statistical analyses but can also be fraught with problems.[1–3] For example, it can often be difficult to organize and extract useful information from the massive amount of unprocessed, unstructured and sometimes invalid data within the electronic health record (EHR).[2,3] Ideally, collection and storage of data should be carefully considered prospectively to avoid these problems and to complement research investigations.[4] However, this can be difficult to achieve prospectively and even more so for the large amounts of data that currently exist. Alternative solutions include manually curated and validated data, however in practice this can become a monotonously time-consuming, highly impractical, or impossible task.

Algorithmic methods are promising alternatives to reduce or eliminate the need for human validation, allowing researchers more time to spend on more meaningful research tasks.[3] Machine learning, whereby an algorithm "learns" to make predictions or associations from input data, has been applied to medicine and anesthesiology, with impressive results.[5–9] Other artificial intelligence approaches such as natural language processing, which can be defined as analysis and processing of human

Received: 20 September 2022
Accepted: 27 April 2023
Published: 20 June 2023

language by an algorithm, and may be combined with machine learning as an approach to text-based problems.[1] To date, there are a dearth of studies and documented approaches applying machine learning as methods in pain medicine research.

This methodological study applies machine learning and natural language processing to aid in the processing and initial analysis of data obtained from the EHR. Clinical data collected from our institution's pain management clinics was combined with patient-reported outcome data collected using the Collaborative Health Outcomes Information Registry (CHOIR) system.[10] We developed an approach to classifying medications as opioids or non-opioids for further use in pain medicine research.[11] This study reports the method and evaluates the accuracy and practicality of the approach. An eventual goal is to minimize time-intensive chart review or manual review of data if these artificial intelligence methods are successful and accurate.

Materials and Methods

Input Data

Medication data was extracted from our EHR (Epic Corporation, Madison, WI) for all patients seen at our multi-disciplinary pain clinics between March 2016 and August 2019 and comprised a list of 4216 distinct entries. Entries had a wide range of formats and included medications and non-medications (ie, braces, orthotics, glucometers, supplements). Two independent human reviewers (author SM, acknowledged contributor AG) each reviewed the entire list and identified all opioid analgesic medications, defined as medications that achieve a therapeutic analgesic effect through their action at an opioid receptor. Mixed agonist-antagonists such as pentazocine and buprenorphine, alone or in combination with other medications, were considered as in the opioid group. A random subset of the medication entries and corresponding classifications were used as input to train the automated algorithm, and the remainder were used to test its effectiveness. The University of Pittsburgh institutional review board granted approval for this study and determined it was exempt from full IRB review due to the retrospective nature and use of de-identified data.

Algorithm Overview

An algorithm was created in MATLAB version r2019a (The MathWorks, Inc., Natick, MA) to perform automated classification of the medication entries. The MATLAB "Text Analytics Toolbox" and the "Statistics and Machine Learning Toolbox" contain widely used and validated implementations.[12] Functions from these toolboxes were used in custom written MATLAB code for this study. Processing steps included the following: randomly selecting and partitioning training and test data, with 60% used for training and 40% used for testing; obtaining textual descriptions for each medication entry through an automated API web search; parsing and processing the text using natural language processing techniques; training a supervised machine learning classification algorithm with the processed text; applying the trained algorithm to the held-out test data; and, evaluating the accuracy of the results compared to the human classifications. A schematic outline of the approach is illustrated in Figure 1, with the training and testing process specific to the machine learning algorithm outlined in Figure 2.

Natural Language Processing

Custom MATLAB code was used to perform a Bing (Microsoft Corporation, Redmond, WA) web-search with the contents of each medication entry as the search string. Searches were automatically performed using the Microsoft Azure "Bing API v7" (Microsoft Corporation, Redmond, WA). The medication name, the webpage title and the summary

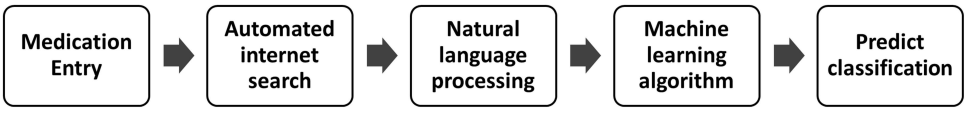

Figure 1 Overview of classification approach.
Notes: An overview of the proposed classification approach begins with the input medication entry from the EHR, which is used to perform an automated internet search. The textual results are used as input for the natural language processing algorithm, and subsequently as input to the machine learning algorithm which then predicts the classification result.

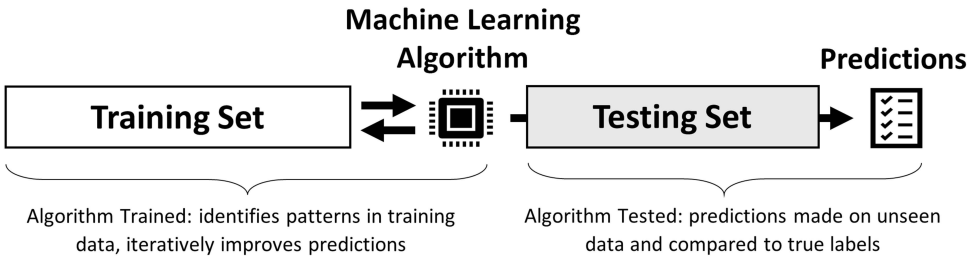

Figure 2 Overview of machine learning algorithm training, testing, and prediction.
Notes: The machine learning algorithm is presented with a random subset of the labeled input data, and iteratively identifies patterns that predict the correct classification (opioid or not) from the input features (in this case, the processed text from the natural language processing algorithm). Once trained, the algorithm is tested on the remaining unseen testing set data, used to create predictions, and compared to the actual classification according to the human reviewers.

descriptions, "snippets", of the first 20 English results were used to create a descriptive textual representation of the medication entry.

The search result text for each entry was then pre-processed using several common methods that have been described previously.[13] These included the following: removal of all non-alphanumeric characters; splitting of text by white space to create word tokens; English lemmatization to word root forms; removal of English stop words (such as "a", "and"); removal of words less than 4 or greater than 15 characters to improve data quality; removal of infrequent words that appeared twice or less in the entire dataset. A count-based bag-of-words approach was used to represent the pre-processed text for each training example.[14,15] This method stores textual information in a sparse matrix format with values representing the number of times each word appears in the training example text.

Training and Testing Sets

The processed NLP output and human labeled classifications were partitioned into randomly selected 60% training and 40% testing sets. Stratification for opioid medication frequency was utilized during partitioning to ensure the training and testing sets had approximately the same proportion of opioids as the entire dataset. To prevent contamination of the test set, the training set was the only data used to create the NLP processing steps and words included in the bag-of-words vocabulary. To account for imbalanced outcomes (fewer opioid compared to non-opioid entries), weights were given to all opioids and all non-opioids in training sample. The weights were set to be inversely proportional to classification frequency in the training set. This weighted each opioid training example more than a given non-opioid example, increasing the importance of accurate identification during training with the goal of improving performance during testing.

Machine Learning Algorithm

A linear error-correcting output code (ECOC) algorithm "fitcecoc" was employed as part of the MATLAB "Statistics and Machine Learning Toolbox" add-on.[12,15] This approach can be extended to multi-class classification problems and can utilize data from the bag-of-words approach. Training data used for the algorithm included the bag-of-words matrix representation of the processed search results, the ground-truth classification, and the weight of the example. Automatic hyperparameter optimization in the MATLAB toolbox ("optimizehyperparameter" set to "auto") was used to obtain the best regularization and learning parameters for the ECOC algorithm, with separate internal cross-validation and training datasets that were randomly selected from the 60% training data. Either a logistic regression or support vector machine-based learner was automatically selected based on performance on the cross-validation set. After training, the finalized model was then used to predict the true classification of the previously unseen test dataset.

Input Variations

The training and testing process was repeated several times for various training set sizes to identify practical sizes for actual applications. Input data was selected to ensure a 1:5 proportion of opioids to non-opioids, though the actual input

was selected randomly and independently each time. The remaining several thousand data entries in each case were used as the test set.

Classification Performance

The classification results were evaluated with several metrics including overall accuracy, defined as number of correct predictions divided by total number of predictions. Other metrics included the sensitivity to classify a medication as an opioid, the positive predictive value of a predicted opioid, the F1 score (a combined metric representing both the positive predictive value and the sensitivity of a classifier), and the area under the curve (AUC) of a receiver operating characteristic (ROC) curve.[16]

Results

The 4216 medication entries were classified by the human reviewers as 225 opioids (5.3%) and 3991 (94.7%) non-opioid medications (examples shown in Table 1). Input data was pre-processed successfully using the steps described, and the ML algorithm was trained and converged appropriately, with logistic regression chosen automatically as the optimal learner. The resulting predictions were 99.8% accurate when applied to the testing set. The algorithm achieved a sensitivity of 97.8% in

Table 1 Predictive Words Identified by the Machine Learning Algorithm

Opioid Associated Words	Non-Opioid Associated Words
Codeine	Vit
Bitartrate	Water
Morphine	Plus
Tramadol	Oval
Tapentadol	Topical
Oxycodone	Headache
Hydrocodone	Sideeffects
Zohydro	Interaction
Analgesic	Silenor
Fentanyl	Information
Oxymorphone	Acid
Xtampza	Sodium
Opana	Safety
Abuse	Disease
Bisect	Learn
Narcotic	Shape
Duragesic	Brand
Opioid	Ingredient

Notes: Shown are the words that were automatically identified by the machine learning algorithm as being highly predictive of either an opioid (left) or non-opioid classification (right). The words are sorted from top to bottom as most predictive to less predictive words.

Table 2 Example Medication Entries and Classification Results

Medication Entry	Human Label*	Predicted*	Posterior Probability†
Bayer aspirin oral	0	0	0
BD insulin syringe ultra-fine 0.3 ML 31 GAUGE X 5/16"	0	0	0
Butalbital 50 MG-Acetaminophen 300 MG-Caffeine 40 MG-Codeine 30 MG Cap	1	0	0.499
Ciclopirox 0.77% topical suspension	0	0	0
Codeine sulfate 30 MG tablet	1	1	1
Diazepam 5 MG/5 ML (1 MG/ML) oral solution	0	0	0
Fioricet oral	0	0	0
Hydrocodone-homatropine 5 MG-1.5 MG/5 ML (5 ML) Syrup	1	1	1
Morphine ER 30 MG-Naltrexone 1.2 MG capsule, extend release, oral only	1	1	1
Multivitamin-iron 9 MG-folic acid 400 MCG-calcium and minerals tablet	0	0	0
Oxymorphone ER 7.5 MG tablet, crush resistant, extended release 12 HR	1	1	1
Oyster shell calcium oral	0	0	0
Pentazocine 50 MG-Naloxone 0.5 MG tablet	1	0	0.007
Percocet 7.5 MG-325 MG tablet	1	1	1
Pregabalin 20 MG/ML oral solution	0	0	0
Testosterone ER 30 MG buccal system, extended release 12HR	0	0	0
Vitamin D3 1000 unit chewable tablet	0	0	0
Vivitrol 380 MG intramuscular suspension, extended release	0	0	0
Xtampza ER 9 MG capsule sprinkle	1	1	1
Zubsolv 2.9 MG-0.71 MG sublingual tablet	1	1	1

Notes: Selected example classification results from the automated approach are shown, including the medication entry text, the human reviewer label, the predicted label from the automated approach, and the posterior probability. †The posterior probability is the likelihood that a medication is an opioid, as determined by the machine learning algorithm, ranging from 0 (unlikely opioid) to 1 (likely opioid). *Label of 1 represents an opioid, label of 0 represents a non-opioid.

detecting an opioid labeled entry, a positive predictive value (PPV) of 94.6%, and an F1 score of 0.9617. The AUC for the derived ROC curve was 0.998. Word tokens identified as most strongly predictive of the classification are shown in Table 2.

When evaluating varying input sizes of training data (opioids and non-opioid examples in a 1:5 ratio), the learning algorithm first reached 90% and 95% accuracy after 7 and 15 opioid examples; an AUC of 0.9, 0.95 after 10, 17 examples; and sensitivity of 90%, 95% after 30, 35 examples, respectively. Increasing performance with increasing training set opioid sample sizes is demonstrated in Figure 3. The relatively lower PPVs represent falsely including non-opioid examples in the opioid classified set.

Discussion

Overall Approach Performance

The automated classification algorithm was able to achieve a high level of performance in classifying the test data. Even with a limited quantity of human-labeled input examples shown in Figure 3, accuracy exceeded 95% at 15 opioid examples and sensitivity exceeded 95% at 25 opioid examples. For practical applications, these results are encouraging, as this approach could allow an investigator to provide a smaller set of opioid examples with non-opioid examples and utilize this approach to automatically classify the remainder of the dataset. The high accuracy and sensitivity could allow

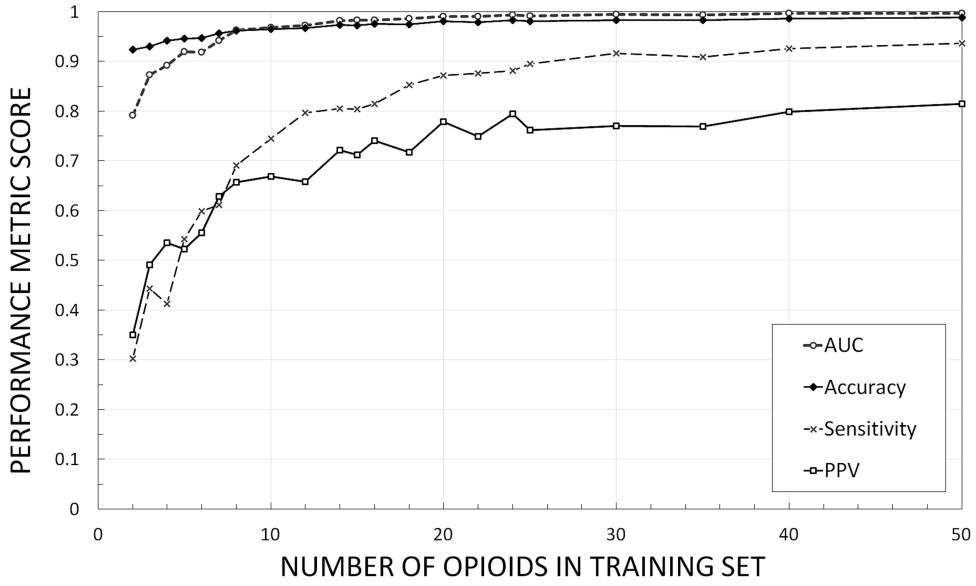

Figure 3 Performance of algorithm compared to training set sizes.
Notes: Practical evaluation of the machine learning algorithm with small training sets shows increasing performance of accuracy, sensitivity, positive predictive value (PPV), and area under the receiver operating characteristic curve (AUC) with increasing training set size.

real-world use with minimal correction of incorrect predictions. Our findings suggest that combining NLP and ML is a promising approach to improving the methods of analyzing EHR data. We present these findings to demonstrate the proof of concept of such an approach applied to pain medicine research.

Practicality of Approach

Previous investigations utilizing ML and NLP have evaluated necessary sample sizes to achieve highly accurate fitting of training example data for various models, input data and problem type. One such investigation suggests that between 80 and 560 samples are needed, varying by problem type, and reaching relatively decreasing returns at around 80 to 200 samples.[17] These numbers are relatively consistent among other studies as well as the current investigation which identified that approximately 15 to 25 positive samples (at a 1:5 ratio) for approximately 90 to 150 total samples are needed to achieve high accuracy, sensitivity, and AUC.[18] Given that the nature of this investigation was primarily a proof of concept, it is feasible that further optimization of the input data, NLP process, and ML algorithm could achieve similar results with less input data. For future applications, there are also several described methods to calculate the sample size needed to achieve a given accuracy, with some suggesting that performance increases logarithmically with increasing sample sizes.[17]

Another practical aspect of employing this method is the ease of obtaining the prediction confidence or "posterior probability". This information reflects the probability of an input belonging to a given class given the training data and model generated. Examples are shown in Table 1. This can be practically used to set a cutoff whereby adjusting the tradeoff between the corresponding sensitivity/specificity of the algorithm. For example, lowering the cutoff would reduce the specificity while increasing the sensitivity. Though this would create additional work for a human operator reviewing identified positive results, it could ensure that classification is verified to be more accurate, and allow for less intensive review of the negative examples.

Algorithm Choice and Limitations

The approach described is reproducible, reliable, extendable, and allows for evaluation of the general methodology, although there are many more sophisticated machine learning algorithms used today. Recent work utilizing deep learning shows promise for complex applications and could theoretically be incorporated in a future version of this approach.[9,16,18] The error-correcting output code algorithm used in this study allows highly accurate multi-class classification training to occur and can be scaled to multiclass large input/output sets.[19] Even relatively small training sets could be used, though a minimal number of input

examples are needed to avoid underfitting or overconfidence in an inaccurate classifier.[14] A limitation of the bag-of-words technique is that contextual information and word ordering are lost. For this investigation, high performance was achieved even without preservation of contextual information, but this may become important in more complex classification problems, necessitating more complex NLP approaches.

Other extensions of this technique could incorporate active learning techniques. These approaches iteratively identify examples which are presented sequentially to the human operator who classifies them, and the algorithm sequentially refines the predictions of the unlabeled examples after each new training example(s).[20] This modification could address and improve upon a limitation of the original approach and maximize the impact of each human-labeled example and thereby minimize the number required to achieve optimal performance.

Applications

The approach presented in this investigation shows promise for providing additional ways to manage and process the large amount of data that is generated from the EHR. The proposed methodology is shown as a proof of concept classifying opioid medications as an example, but this approach could be expanded to utilize other EHR data of interest. For example, identifying other groups of medications (ie, NSAIDs), groups of related diagnosis codes (ie, lumbar pain), or procedures (ie, interventional pain procedures) would allow more standardized organization of the EHR data and ease of use in research. If successful, the paradigm of manual chart review could be supplanted by artificial intelligence methods not prone to human error or fatigue. This would reduce labor-intensive, costly, and sometimes inaccurate, human chart review that is often required for modern research.

Beyond data organization, more complex applications can also be envisioned similar to those described in prior publications.[5,7,8] These could include identification of complex patterns to predict adverse events, identify clusters of related patients and patterns, and a multitude of other interesting investigations in pain medicine and beyond.

Conclusion

The automated approach achieved a highly accurate classification of opioid vs non-opioid medications by combining machine learning and natural language processing techniques. This could significantly reduce the need for manual chart review and improve the structuring of data in the EHR. Future applications could involve more complex analyses and pattern recognition with a multitude of potential applications in pain medicine and beyond.

Abbreviations

AUC, area under the curve; CHOIR, Collaborative Health Outcomes Information Registry; ECOC, error-correcting output code; EHR, electronic health record; ML, machine learning; NLP, natural language processing; PPV, positive predictive value; ROC, receiver operating characteristic curve.

Acknowledgments

The authors would like to acknowledge the contribution of Andrea G. Gillman Ph.D. (UPMC Pain Medicine, Pittsburgh, PA, USA), for her contributions to data collection and study design, including medication classification and data acquisition.

Disclosure

The authors report no conflicts of interest in this work. Internal departmental funding was utilized in support of this work.

References

1. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol*. 2017;106(1):1–9. doi:10.1007/s00392-016-1025-6
2. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Bias associated with mining electronic health records. *J Biomed Discov Collab*. 2011;6:48–52. doi:10.5210/disco.v6i0.3581
3. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395–405. doi:10.1038/nrg3208

4. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(803):S30–S37. doi:10.1097/MLR.0b013e31829b1dbd
5. Caelen O, Bontempi G, Barvais L. Machine learning techniques for decision support in anesthesia. In: Bellazzi R, Abu-Hanna A, Hunter J, editors. *Artificial Intelligence in Medicine. Lecture Notes in Computer Science*. Berlin Heidelberg: Springer; 2007:165–169.
6. Gambus P, Shafer SL. Artificial Intelligence for Everyone. *Anesthesiology*. 2018;128(3):431–433. doi:10.1097/ALN.0000000000001984
7. Hatib F, Jian Z, Buddi S, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology*. 2018;129(4):663–674. doi:10.1097/ALN.0000000000002300
8. Kendale S, Kulkarni P, Rosenberg AD, Wang J. Supervised machine-learning predictive analytics for prediction of postinduction hypotension. *Anesthesiology*. 2018;129(4):675–688. doi:10.1097/ALN.0000000000002374
9. Lee HC, Ryu HG, Chung EJ, Jung CW. Prediction of bispectral index during target-controlled infusion of propofol and remifentanil: a deep learning approach. *Anesthesiology*. 2018;128(3):492–501. doi:10.1097/ALN.0000000000001892
10. Dressler AM, Gillman AG, Wasan AD. A narrative review of data collection and analysis guidelines for comparative effectiveness research in chronic pain using patient-reported outcomes and electronic health records. *J Pain Res*. 2019;12:491–500. doi:10.2147/JPR.S184023
11. Harbor N. AAPM 2020 award-winning scientific poster abstracts - poster abstract 211. *Pain Med*. 2020;21(6):1307–1310. doi:10.1093/pm/pnaa106
12. Statistics M, Release MLT, The MathWorks Inc. Natick Mass US; 2016.
13. Scott S, Matwin S. Feature engineering for text classification; 1999.
14. Ruch P, Baud R, Geissbühler A. Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. *Int J Med Inf*. 2002;67(1–3):75–83. doi:10.1016/S1386-5056(02)00057-6
15. MATLAB & Simulink. Create simple text model for classification; 2019. Available from: https://www.mathworks.com/help/textanalytics/ug/create-simple-text-model-for-classification.html. Accessed August 28, 2019.
16. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Npj Digit Med*. 2018;1(1):18. doi:10.1038/s41746-018-0029-1
17. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12(1):8. doi:10.1186/1472-6947-12-8
18. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? ArXiv151106348 Cs; 2016. Available from: http://arxiv.org/abs/1511.06348. Accessed January 7, 2020.
19. Dietterich TG, Bakiri G. Solving multiclass learning problems via error-correcting output codes. *J Artif Intell Res*. 1995;2:263–286. doi:10.1613/jair.105
20. Olsson F. A literature survey of active machine learning in the context of natural language processing; 2009.