Machine Learning Diagnosis of Peritonsillar Abscess

Michael B. Wilson, MD¹, S. Ahmed Ali, MD¹, Kevin J. Kovatch, MD¹, Josh D. Smith, MD¹, and Paul T. Hoff, MD¹

Abstract
Peritonsillar abscess (PTA) is a difficult diagnosis to make clinically, with clinical examination of even otolaryngologists showing poor sensitivity and specificity. Machine learning is a form of artificial intelligence that “learns” from data to make predictions. We developed a machine learning classifier to predict the diagnosis of PTA based on patient symptoms. We retrospectively collected clinical data and symptomatology from 916 patients who underwent attempted needle aspiration for PTA. Machine learning classifiers were trained on a subset of the data to predict the presence or absence of purulence on attempted aspiration. The performance of the model was evaluated on a holdout set. The accuracy of the top-performing algorithm, the artificial neural network, was 72.3%. Artificial neural networks can use patient symptoms to exceed human ability to predict PTA in patients with clinical suspicion for PTA. Similar models can assist medical decision making for clinicians who have suspicion of PTA.

Keywords
machine learning, artificial intelligence, artificial neural networks, peritonsillar abscess

Received March 7, 2019; accepted July 17, 2019.

Peritonsillar abscess (PTA) is the most common head and neck abscess. Management involves needle aspiration and sometimes formal incision and drainage. The decision to drain a PTA has historically been a clinical one, with concerning signs and symptoms including sore throat, trismus, otalgia, unilateral palatal fullness, and uvular deviation. These signs are imprecise, as evidenced by a previous study suggesting that even expert diagnosis of PTA based on examination findings alone has a specificity of 50% and a sensitivity of 78%.

Machine learning (ML), a form of artificial intelligence, offers promise for aiding clinical decision making. ML uses algorithms that “learn” to make predictions directly from data without explicit programming. ML algorithms have been used successfully for clinical triaging and decision making. In the emergency department setting, ML algorithms surpassed human ability in predicting clinical outcomes and pathologic findings. Recent articles have reviewed applications of ML in otolaryngology and highlighted the potential impact of artificial intelligence in enhancing the delivery of clinical care.

Supervised learning algorithms that “learn” from labeled data have featured prominently in medical applications of ML. Random forest, an ensemble technique that combines results from a collection of decision trees, is one of the most powerful approaches in ML and performs well in most problem domains, including medical diagnosis. Artificial neural networks (ANNs) are a class of algorithm that loosely mimic the structure of the human brain. They are composed of individual activation functions that, like neurons, are interconnected and layered. Each interconnection has a weight that is adjusted incrementally during training. ANNs can model complex nonlinear relationships between variables, and they have been used successfully in medical diagnosis. Logistic regression is a widely used algorithm for binary classification, often used for baseline comparison in ML studies.

Methods
Data Collection
We performed a University of Michigan Institutional Review Board–approved single-institution retrospective analysis of patients of all ages who presented with concern for PTA. Demographics, symptomatology, and result of needle aspiration were obtained from the electronic medical record. Aspiration was considered positive if purulent fluid was aspirated. Imaging was not obtained on all patients and was not included in modeling. Data from these patients were used to train ML models.

¹Department of Otolaryngology–Head and Neck Surgery, University of Michigan, Ann Arbor, Michigan, USA

Corresponding Author:
Michael B. Wilson, MD, Department of Otolaryngology–Head and Neck Surgery, Michigan Medicine, 1500 East Medical Center Drive, Ann Arbor, MI 48109, USA.
Email: mbwils@med.umich.edu
Model Training
Data were randomly divided into a training set and a test set: 70% was used to train the classifiers and 30% was reserved for testing model accuracy (Figure 2). Classifiers were trained to predict the presence of purulence on attempted aspiration. We used Scikit-learn, an open source ML library, to implement the classifiers.

Results
A total of 916 patients underwent attempted needle aspiration for PTA. This cohort had a mean age of 24.7 years (range, 1-88) and was 50.1% male. The symptoms that most strongly correlated with the results of needle aspiration were included in the algorithm: otalgia, trismus, duration of symptoms, neck pain, worsening of symptoms, and no previous treatment (for group demographics and symptomatology, see Supplemental Table S1, available in the online version of the article). The accuracy observed for the top-performing algorithm, the ANN, was 72.3%, and the sensitivity and specificity were 86% and 50%, respectively (Table 1).

Discussion
PTA is difficult to diagnose clinically. One study found clinical diagnosis of otolaryngologists to be 64% accurate, with a sensitivity of 78% and a specificity of 50%. Another study found that only 40% of pediatric patients diagnosed clinically with PTA had positive imaging findings. The ML algorithms described herein surpass reported accuracies.
of expert clinical diagnosis; the ANN demonstrated an accuracy of 72.3% and a sensitivity of 86.5%.

Notably, the variables used as input for this algorithm for prediction consist entirely of patient-reported symptoms. This supports the utility of these algorithms to increase diagnostic accuracy independent of physical examination findings, which are examiner dependent. Indeed, the accuracy demonstrated without physical examination findings suggests even greater utility in decision support for clinicians who are less familiar with the pertinent clinical examination findings of PTA.

Limitations of this study include retrospective data collection and reliance of patient-reported data and electronic medical record documentation. Fortunately, given the well-defined symptomatology of the PTA, there was great uniformity within the clinical reports examined. We are mindful not to substitute the outcome predicted here—purulence on attempted aspiration—as a direct proxy for presence of PTA; falsely negative attempted aspirations are well documented. When attempted aspiration is considered, however, the accurate prediction of a successful aspiration is arguably the most clinically pertinent end point.

Importantly, the accuracy attained in this study would be expected only when making predictions on populations similar to the one on which the models were trained. The patients enrolled in this study had been previously evaluated by clinicians, and suspicion for PTA was high enough to warrant needle aspiration attempt. This alters the prevalence of PTA in the study group as compared with similar groups, such as all-comers presenting with odynophagia. While the features used in these predictions do not include physical examination findings, examination findings had likely been selected for prior to inclusion. This limits the applicability of the algorithms developed here beyond patients who have had a physician evaluation.

**Conclusion**

The algorithms developed here demonstrate the potential of ML to make clinically meaningful predictions at accuracies matching or exceeding those of human experts. Such algorithms are easily made available for general use; as an example of a potential implementation, we have made the ANN described here available in a web app calculator (PTA Calculator, http://www.PTAcalculator.org). Similar algorithms could help clinicians of all experience levels enhance their diagnostic accuracies and improve the delivery of clinical care in this and other diagnostic challenges.

**Table 1. Performance Measures of Machine Learning Algorithms.**

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial neural network</td>
<td>0.72 (0.68-0.76)</td>
<td>0.86 (0.77-0.95)</td>
<td>0.50 (0.35-0.65)</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.71 (0.66-0.75)</td>
<td>0.74 (0.65-0.82)</td>
<td>0.66 (0.55-0.76)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.71 (0.66-0.76)</td>
<td>0.77 (0.69-0.88)</td>
<td>0.62 (0.51-0.73)</td>
</tr>
</tbody>
</table>

**Author Contributions**

Michael B. Wilson, drafted manuscript, programmed machine learning algorithms, literature review; S. Ahmed Ali, drafted manuscript, data collection, literature review; Kevin J. Kovatch, data collection, manuscript editing, literature review; Josh D. Smith, data collection, manuscript editing, literature review; Paul T. Hoff, manuscript editing, substantial contributions to design of the work.

**Disclosures**

**Competing interests:** None.

**Sponsorships:** None.

**Funding source:** None.

**Supplemental Material**

Additional supporting information is available in the online version of the article.

**References**


8. Heckerling PS, Canaris GJ, Flach SD, Tape TG, Wigton RS, Gerber BS. Predictors of urinary tract infection based on...


