Japanese Dental Science Review 57 (2021) 242–249

Contents lists available at ScienceDirect

Japanese Dental Science Review

journal homepage: www.elsevier.com/locate/jdsr

Review Article

An artificially intelligent (or algorithm-enhanced) electronic medical record in orofacial pain

Anette Paulina Vistoso Monreal[a], Nicolas Veas[b], Glenn Clark[a,*]

[a] Herman Ostrow School of Dentistry, University of Southern California, Los Angeles, CA, USA
[b] McCombs School of Business, The University of Texas, Austin, TX, USA

ARTICLE INFO

Article history:
Received 24 August 2021
Received in revised form 2 November 2021
Accepted 3 November 2021

Keywords:
Orofacial pain
Prediction
Algorithms
Modeling
Machine learning

ABSTRACT

This review examines how a highly structured data collection system could be used to create data-driven diagnostic classification algorithms. Some preliminary data using this process is provided. The data collection system described is applicable to any clinical domain where the diagnoses being explored are based predominately on clinical history (subjective) and physical examination (objective) information. The system has been piloted and refined using patient encounters collected in a clinic specializing in Orofacial Pain treatment. In summary, whether you believe a branching hybrid check-box based data collection system with built-in algorithms is needed, depends on your individual agenda. If you have no plans for data analysis or publishing about the various phenotypes discovered and you do not need pop-up suggestions for best diagnosis and treatment options, it is easier to use a semi-structured narrative note for your patient encounters. If, however, you want data-driven diagnostic and disease risk algorithms and pop-up best-treatment options, then you need a highly structured data collection system that is compatible with machine learning analysis. Automating the journey from data collection to diagnoses has the potential to improve standards of care by providing faster and reliable predictions.

1. Introduction

If a clinician fails to collect the needed data during a patient encounter or if they fail to interpret the data correctly, misdiagnosis is probable. Correct diagnosis requires high domain knowledge as the clinician must learn and retain a relatively large list of variables that define a wide variety of diagnoses. This review examines how a highly structured data collection system could be used to help solve the above problems and create data-driven diagnostic classification algorithms. Some preliminary data using this process is provided and the data collection system described is applicable to any clinical domain where the diagnoses being explored are based predominately on clinical history (subjective) and physical examination (objective) information. The system has been piloted and refined using patient encounters collected in a clinic specializing in Orofacial Pain treatment.

2. What are the problems that an algorithm enhanced electronic medical record might solve?

Novice clinicians, e.g., clinicians in training and non-specialists, are at greater risk of making inefficient and/or incorrect decisions and these risks increase with uncommon diseases [1]. Many reasons have been identified that lead to such errors [1,2]. The Institute of Medicine in a 2015 report, highlighted the fact that physician diagnostic error is common and suggested that computer algorithms may eventually be able to make accurate clinical diagnoses [2,3]. The essential problem is that clinical medicine and disease is increasing complex and potentially overwhelming the practitioner's ability to diagnose complex cases correctly.

One potential innovation that would help, is to collect and enter the patient's signs and symptoms into a highly coded electronic medical record (EMR) which has built-in predictive diagnostic and maybe even predictive best-treatment algorithms. In this scenario, before the clinician finalizes their diagnosis and treatment plan and presents it to the patient, they could check with the algorithms in the EMR to see which diagnoses and which associated treatment

Abbreviations: EMR, electronic medical record; OFP, orofacial pain; ML, machine learning; D1, internal derangement with reduction (DDWR); D3, masticatory myalgia and/or cervical myalgia; D4, arthralgia/capsulitis; D5, TMJ arthritis.

* Corresponding author at: School of Dentistry, 925 W. 34th St, Los Angeles, CA, 90089, USA.
E-mail address: gtc@usc.edu (G. Clark).

https://doi.org/10.1016/j.jdsr.2021.11.001

approach the data is suggestive of. This idea is our primary driving motivation for the work described in this paper. Essentially, we propose to use algorithms based on coded clinical features (signs and symptoms) to suggest possible diagnoses to clinicians after they enter but before they make their final diagnosis. The system described in this review paper will be specific for patients attending an orofacial pain (OFP) clinic with a variety of pain, headache and temporomandibular disorders and dysfunctions.

Before we can create and employ predictive diagnostic or even predictive best-treatment algorithms, we must first code the collected patient encounter information (signs and symptoms) so that the output is readily compatible with machine learning (ML) analysis. With ML analysis we can then select the most relevant clinical features, build algorithms and use them to make diagnostic and treatment predictions. A 2021 paper recently described the need for machine learning modeling in the diagnosis of temporomandibular disorders. Specifically, the authors of this paper, focused on the use of image analysis to diagnose TMJ Osteoarthritis [4]. The authors suggested that in addition to clinical data, imaging data could one day be used to achieve an automatic disease classification. They recommended that ML based statistical analytic methods will play a crucial role in this process.

3. What ongoing and prior work has been done on creating medical algorithms?

A medical algorithm is a very narrow form of artificial intelligence and is any computation, formula, statistical survey or look-up table, useful in healthcare. Medical algorithms include risk/survival predictions as well as diagnostic and treatment choice predictions (e.g., if symptoms A, B, and C are evident, then diagnosis X is probable and use treatment Y). Attempts to build a comprehensive algorithm enhanced medical record is ongoing in the EMR industry and may yet be achieved. One notable failure, at an estimated cost of $39,000,000 USD, that is worth discussing and involved an attempt to build an algorithm enhanced EMR system designed specifically for use in a Cancer center. This effort was a 2012 partnership between M.D. Anderson Partners and IBM Watson in Houston, Texas [5]. The early promotional news stories that described the project, stated that the plan was to combine genetic data, pathology reports with clinicians' notes and relevant journal articles to help doctors come up with diagnoses and treatments. However, five years later, in February 2017 M.D. Anderson announced it had shuttered the project since after several years of trying, it had not produced a tool for use with patients that was ready to go beyond pilot tests. In theory, machine learning compatible systems are supposed to continually readjust their data algorithms to produce the highest possible percentage of correct answers.

While the M.D. Anderson-IBM Watson shuttering story is daunting, there are researchers who continue to strive for an algorithm enhanced EMR. Over the past few decades, several medical diagnostic support systems have been described in the literature (e.g. DXplain, Isabel, VisualDx, Iliad, QMR, Ada DX) [6–10]. Unfortunately, these earlier systems generally consisted of stand-alone software into which an operator can enter various clinical observations and then receive a list of potential diagnoses. The algorithms behind these systems have mostly been based on hand-crafted expert opinion [11–13]. One recent and very innovating article (2021), described a new Bayesian Inference/Exploration model based EMR system [14]. The author lamented that currently, most electronic medical records are problematic because they use free text based narrative notes, which are prone to ambiguities, omissions and errors and this makes text-mining very difficult. The article went on to describe how the author created an EMR system with built in algorithms. The author claimed their algorithms would help clinicians by suggesting multiple possible diagnoses for a given a set of clinical findings. Moreover, the article described their system as one that would be able to advise the clinician, given a specific diagnosis, on which would be the most useful diagnostic tests and best treatment approach to select.

4. Are algorithms really needed in medicine and dentistry

The answer to this question, "Are algorithms really needed?", depends entirely on what type of clinical service is being provided and whether there are definitive diagnostic tests available for these diagnoses. If the service deals with a narrow scope of diseases, dysfunction and disorders, then most clinicians quickly learn and can recognize the limited number of diagnostic conditions that present for help. Moreover, if there are available, definitive diagnostic tests (e.g. images, serology, histopathology) that define the diagnosis then algorithms are less helpful. However, if the clinical service has a large number of potential diagnoses that do not have a definitive diagnostic test then data-driven diagnostic algorithms could be very helpful. In the clinical domain of Orofacial Pain, we estimate there are over 100 different problems that must be differentiated and diagnosed. To illustrate this last point, in 2020, a new International Classification of Orofacial Pain (ICOP) was published that has 40 specific diagnoses [15]. As the name implies, the ICOP categorization system is pain focused and it is a great advance in orofacial disease classification. Unfortunately, patients who attend an Orofacial Pain clinic do not always have pain and the ICOP system does not include non-painful dysfunctions, or deformities of the masticatory system. It also does not classify many of the abnormal jaw mobility, abnormal oral motor or sleep-breathing disorders of the oropharyngeal region that vex patients.

Assuming you want to create or compile a larger set of "data-based" algorithms for these multiple orofacial pain conditions, and some are quite rare, you first will need to collect many thousands (maybe 5000–10,000) of highly structured patient records. This is because you must have a sufficient number or case examples for each diagnosis to be able to identify the defining features of the diagnosis. Existing patient records can be "text mined" if they are in a narrative notes style. As mentioned previously, narrative notes are usually often inconsistent because too often critical variables are not recorded. This inconsistency makes it hard to text mine and during the early phase of the work described in this paper, we examined narrative notes collected in the orofacial pain clinic and found that even though our clinic used a standardized SOAP note system, crucial diagnosis data was too often omitted.

5. What algorithms are needed in the orofacial pain domain?

If your goal is to achieve data-driven algorithms, beyond simple predicting a potential set of diagnoses from the clinical data, there are many additional types of algorithms that would be very helpful. For example, disease risk algorithms and outcome data-based treatment algorithms. In addition, it would be very useful to know how age, medical status, psychological status, race, economic status or other categorizing factors influence these predictions. Both risk/survival algorithms and treatment outcome algorithms will require long term data collection. Every diagnosis should have an associated data-based best treatment protocol but the only way they can be data-based, is to systematically collect structured outcomes data on these various diagnoses, where a specific treatment approach was provided. With such data you might then be able to characterize the specific patients that respond versus those that do not respond. This is called personalized medicine.

6. How good are existing OFP algorithms?

With the increased use and availability of electronic medical record (EMR) data, machine learning (ML) approaches have been used extensively for making data-driven clinical predictions [16,17]. EMR data may include, patient interview notes, medical history, physical examination findings, imaging and laboratory test and patient facing questionnaires. Several studies have used ML for clinical predictions, e.g., for symptom severity in mental care [18], to diagnose common headaches [19] and predict fertility [20]. This work relies on traditional ML approaches (e.g., logistic regression, decision trees, support vector machine) that are known to perform well on smaller datasets.

In the domain of orofacial pain, McCartney et al. developed and used a Neural Network (NN) based system to diagnose facial pain syndromes from patient's self-assessment responses [21]. Limonadi et al. also used this NN model on another set of 143 patients to diagnose facial pain syndromes using data collected from a questionnaire with 18 binomial (yes/no) questions. They reported good results on some of the 7 diagnoses that were considered [22]. In other clinical applications, where larger datasets are available, NN approaches have been shown to be successful, e.g., to predict optimal treatment strategies [23]. While these works realize the potential for automatic diagnosis, they do not explore the process of selecting discriminative variables and how to scale up beyond a limited set of diagnoses.

More recently in the domain of headache disorders there have been two articles that describe algorithms for diagnostic purposes. One 2019 paper created algorithms using a hybrid expert opinion-based method [19]. The authors analyzed narrative notes from clinical encounter and with their algorithm were able to diagnose four types of headaches from a set of 190 patients achieving accuracy levels greater than 90%. In 2020, another article reported on how machine learning modeling created algorithms that produced automated classification of headache disorders using only patient-reported questionnaire data [24]. Specifically, this paper merged the diagnoses into five major headache entities and the created the algorithms using training data (n = 1286). Once the algorithms were created, the authors then they examined their accuracy using an additional test set (n = 876) patient. The methods used to create the algorithms was a stacked classifier model with four layers of XGBoost classifiers. Each layer selected different features from the self-reports by using least absolute shrinkage and selection operator. In the test cohort, the stacked classifier obtained an overall accuracy of 81% with a sensitivity of 88%, 69%, 65%, 53%, and 51%, and specificity of 95%, 55%, 46%, 48%, and 51% for migraine, Tension-type headache, Trigeminal Autonomic Cephalgias and Epicranial/Thunderclap Headaches, respectively. This article concluded that a machine-learning based approach was applicable in analyzing patient-reported questionnaire data.

7. Why create a ML compatible note taking system?

It goes without saying that any and all diagnostic predictive algorithms created, will need to be based on good data. To collect such data, you can text-mine narrative notes from an EMR or you can start fresh, by collecting highly coded new data as you see patients. As mentioned, the approach we did not employ was text-mine existing data by using of a Natural Language Processing (NLP) code stack applied to our narrative patient notes. Depending on the questions being asked NLP is a powerful technique that is very useful. NLP based analytic text-mining usually involves manually constructing keyword lists and rules that allow the narrative notes to be analyzed. Unfortunately, this process is very difficult, and the findings are usually not replicable if you try to examine data from another institution [25]. The paper by Zeng et al., suggested that even well-designed keyword search and rule-based systems, which will show high accuracy for the data it was developed on it, will generally have low generalizability and scalability when applied to other data sets, suggesting a new keyword set and new rules will need to be created.

Instead of using NLP text-mining we elected to build a new data collection system. This system would be best described as a hybrid narrative note with a branching, checkbox-based architecture that is machine learning compatible. The system is hybrid in that it has supplemental note fields (because some elements of the patient's story will always be outside of available check boxes). All follow-up, after treatment has been initiated, patient encounters have a set of standardized outcome assessment questions. Finally, each diagnosis has a suggested set of treatment plan choices which can be customized by the clinician. We have collected in 9 months, over 721 new patient encounters. By being both highly structured and compatible with Machine Learning analytic methods we will be able to phenotype different diagnostic conditions by suggesting a set of key variables that define the diagnosis. Once we have several years of outcome data, we will examine it to see if we can predict the probability of treatment success or failure in different demographic groups. Currently each patient record has 922 fields in the history and examination sections of the record. The system allows for parameter tuning and some limited feature engineering and all data recorded are periodically synchronized for consistency. The SQL data is exported as a CSV file and it is moved to a machine learning compatible notebook. The machine learning tools we use are essentially classification modeling.

8. Where are we with our predictive diagnostic algorithms

With the data being collected now, we are actively creating multiple OFP diagnostic algorithms. One large problem is that in our patient pool, we can readily identify 100 plus orofacial pain and dysfunction related diagnoses, each with its own ICD code. To make our task more achievable, we elected to target our algorithms around a set of "preliminary" diagnoses. This was done by merging those diagnoses that had a similar clinical presentation, into a single composite preliminary diagnosis category. For example, two of the composite or preliminary diagnoses we use are facial asymmetry and limited opening. There are multiple reasons a patient presents with a facial asymmetry, usually involving lower jaw, (e.g. mandibular trauma with growth interference, hyperplasia, hypoplasia, osteochondroma, etc.). There are multiple reasons a patient can develops a limited mandibular opening also. These reasons include jaw closer muscle trismus, TMJ disc displacement without reduction, extracapsular fibrosis or contracture, and Scleroderma to name only a few. Using this process, we reduced the number or needed algorithms down to thirty. Unfortunately, even after creating 30 composite preliminary diagnoses, the incidence of some of our diagnoses is still so rare (<1 diagnosis per 500–1000 new patient cases) that to that get 100 cases of a diagnoses so we can proceed with analysis with take 5–10 years of data collection, assuming we collect 1000 cases are year.

Having a set of algorithms that can predict a probable preliminary diagnosis, with the final diagnosis ultimately being made by the clinician, is a reasonable and achievable starting point for our project. We estimate that within 1–2 years we will have algorithms for more than 15 of our 30 preliminary diagnoses. Once we have enough data to create algorithms for most if not all of the preliminary diagnoses, we will incorporate these algorithms into our note-taking system. Early elements of this work are described in a preliminary paper presented at the 2020 American Medical Informatics Association meeting [26]. In this prior research, we created

Table 1
Performance of D1 algorithms.

Model	Train accuracy	Test accuracy	Train recall	Test recall	Train precision	Test precision	Train F1-score	Test F1-score
Tuned decision tree	0.9385	0.9355	0.9868	0.9846	0.8371	0.8312	0.9058	0.9014
Tuned random forest	0.9782	0.9631	0.9934	0.9846	0.9375	0.9014	0.9646	0.9412
Tuned gradient boosting classifier	0.9960	0.9677	1.0000	0.9692	0.9869	0.9265	0.9934	0.9474
Stacking classifier	0.9921	0.9724	1.0000	0.9692	0.9742	0.9403	0.9869	0.9545
Bagging classifier	0.9960	0.9539	0.9868	0.9538	1.0000	0.8986	0.9933	0.9254
Tuned bagging classifier	0.9841	0.9493	0.9934	0.9385	0.9554	0.8971	0.9740	0.9173
Gradient boosting classifier	1.0000	0.9585	1.0000	0.9385	1.0000	0.9242	1.0000	0.9313
KNN classifier	1.0000	0.9585	1.0000	0.9385	1.0000	0.9242	1.0000	0.9313
Tuned logistic regression	0.9821	0.9493	1.0000	0.9077	0.9438	0.9219	0.9711	0.9147
Random forest	1.0000	0.9585	1.0000	0.9077	1.0000	0.9516	1.0000	0.9291
Logistic regression	0.9901	0.9539	0.9801	0.8769	0.9867	0.9661	0.9834	0.9194
Decision tree	1.0000	0.9355	1.0000	0.8769	1.0000	0.9048	1.0000	0.8906
AdaBoost classifier	1.0000	0.9493	1.0000	0.8769	1.0000	0.9500	1.0000	0.9120
Tuned AdaBoost classifier	0.9980	0.9447	1.0000	0.8769	0.9934	0.9344	0.9967	0.9048
Tuned KNN classifier	0.9008	0.8710	0.7881	0.6769	0.8686	0.8627	0.8264	0.7586

a database of 451 cases that we text-mined manually from narrative notes taken from our clinic. The cases had a variety of Orofacial Pain disorders, but for analysis we analyzed a subset of 5 diagnoses, those that were most frequent. The Machine Learning models we created for these five diagnoses showed a variable accuracy using Recall scores (0.98–0.59). The differences between our prior ML analysis and the current one described below is that we increased the number of cases, expanded and refined both the subjective and objective variables, we eliminated the need for NLP text-mining as we coded the data during the collection phase. Like the prior paper we only conducted ML analysis on the most frequent diagnoses, but we raised the inclusion criteria for ML modeling to require at least 100 case examples of the diagnosis. As a result of these changes our Recall scores improved substantially (see Tables 1,3,5 and 7).

Table 2
Variable Importance for tuned random forest on D1.

Variable	Importance
EXAM: TM joint clicking positive	0.8217
HPI: Question #02-clicking	0.07586
Exam: TMJ-right click	0.01545
OFPD02 (DDNR)	0.013853
HPI-Duration-lasts sec-mins	0.012156
HPI-Question #04-lock open/hypermobility	0.006128
Age-	0.006043
HPI-Severity	0.005819
HPI-Onset-present for months	0.003112
HPI-Frequency-only a few times a week	0.002899

9. Preliminary data preview

Using our new prospective dataset of 721 consecutive cases collected using our data collection system we first performed a careful review of this dataset for omissions and typographical errors. These data were stored in the form of a dataframe [27] as a table of rows, where each row corresponds to a case with features and labeled diagnoses. In all the cases where a given diagnosis occurs and when a sufficient number of cases were available, we applied ML modeling. As of now we have analyzed only the 4 most frequent diagnoses, those that had a sufficient number of cases (n = >100). This process involved classification analysis using a K-fold cross validation with K = 5 and we report performance averaged over the folds.

Our dataset of 721 new patient cases, had an age range from 2 to 87, mean age was 46 years. There were 526 females (72.95%) and 196 males (27.04%). Of the 922 possible history and examination variables, we had 649 variables with information. Many of the variables with data have so little data they are not useful for algorithmic classification and 273 of the variables were never selected so could potentially be removed for the system. Among the 649 variables with information, 612 were dichotomous, and 37 were continuous. The continuous variables were age, pain severity on a discrete scale [0–10], tenderness scores on palpation on a discrete scale [0–3], and various jaw motion measurements in millimeters. The 4 diagnoses we analyzed are presented in Fig. 1 below.

In the remainder of the paper, we will refer to these 4 preliminary diagnoses as d1, d3, d4, d5 as show in Fig. 1 above. Because patients often have multiple diagnoses, each case in the dataset can have any or all of the diagnoses considered. Fig. 2 shows the case counts for all possible combinations of the include diagnoses. Our new data set was transferred into python notebook. We used Scikit-learn [28] to implement ML algorithms. ROC curve [29] was used to determine the optimal algorithm threshold score in the logistic

Table 3
Performance of D3 algorithms.

Model	Train accuracy	Test accuracy	Train recall	Test recall	Train precision	Test precision	Train F1-score	Test F1-score
Tuned bagging classifier	0.5893	0.5714	1.0000	1.0000	0.5893	0.5714	0.7416	0.7273
Tuned logistic regression	0.9266	0.9217	0.9865	0.9758	0.8988	0.8963	0.9406	0.9344
Gradient boosting classifier	0.9940	0.9493	1.0000	0.9758	0.9900	0.9380	0.9950	0.9565
Tuned gradient boosting classifier	1.0000	0.9447	1.0000	0.9758	1.0000	0.9308	1.0000	0.9528
KNN classifier	0.9940	0.9493	1.0000	0.9758	0.9900	0.9380	0.9950	0.9565
Tuned decision tree	0.9345	0.9401	0.9663	0.9677	0.9258	0.9302	0.9456	0.9486
Random forest	1.0000	0.9217	1.0000	0.9677	1.0000	0.9023	1.0000	0.9339
Tuned random forest	0.9544	0.9401	0.9731	0.9677	0.9507	0.9302	0.9617	0.9486
Tuned AdaBoost classifier	0.9345	0.9401	0.9663	0.9677	0.9258	0.9302	0.9456	0.9486
Stacking classifier	0.9325	0.9309	0.9630	0.9516	0.9256	0.9291	0.9439	0.9402
Logistic regression	0.9444	0.9032	0.9697	0.9355	0.9381	0.8992	0.9536	0.9170
Decision tree	1.0000	0.9171	1.0000	0.9274	1.0000	0.9274	1.0000	0.9274
AdaBoost classifier	0.9901	0.9217	0.9933	0.9274	0.9899	0.9350	0.9916	0.9312
Bagging classifier	0.9960	0.9171	0.9933	0.9194	1.0000	0.9344	0.9966	0.9268
Tuned KNN classifier	0.8770	0.8111	0.9024	0.8871	0.8904	0.8029	0.8963	0.8429

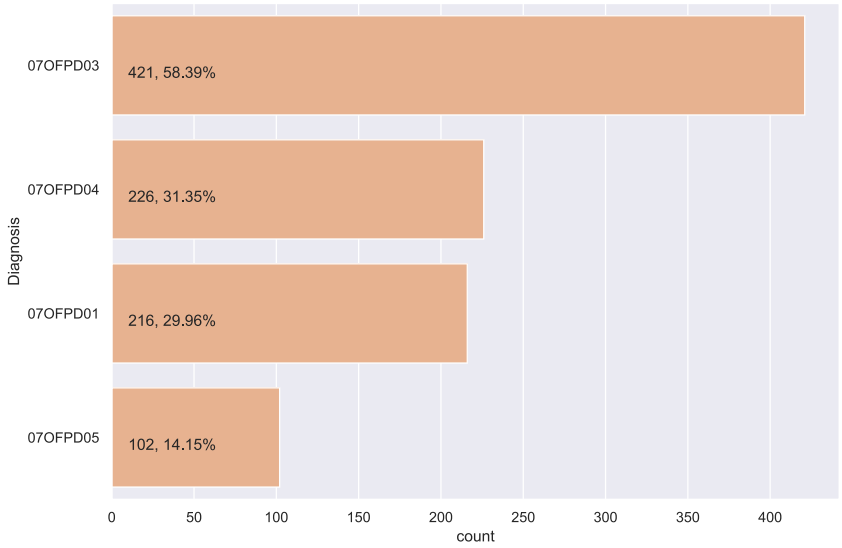

Fig. 1. Analyzed preliminary diagnoses. The selected and analyzed diagnoses were the most frequent diagnoses and each had over 100 case examples in our dataset. D1: internal derangement with reduction; D3: masticatory and cervical myalgia/myofascial pain; D4: arthralgia/capsulitis and D5: osteoarthritis of the TMJ.

Table 4
Odds for logistic regression for D3.

Variable	Odds
CC – Location: jaw_muscles	0.957712
Exam: Masseter tenderness right (0–3)	0.675065
HPI: Question #02-very sore jaw muscles	0.672429
Exam: Masseter tenderness left (0–3)	0.420552
Exam: Masticatory muscle tenderness score	0.371771
CC-Pain	0.283078
Exam: TM joint tenderness score	0.23122
HPI-Severity	0.227297
HPI-Pattern-continuous	0.196904
HPI-Cause-stress	0.189058

Table 6
Variable importance for D4.

Variable	Importance
Exam-TM joint tenderness score	0.925483
Exam-TMJ-left tend (0–3)	0.074517
HPI-Cause was trauma	0
HPI-Side-bilaterally	0
HPI-Pattern-continuous	0
HPI-Pattern-intermittent	0
HPI-Frequency-one or more times a day	0
HPI-Frequency-only a few times a week	0
HPI-Duration-lasts sec-mins	0
HPI-Duration-lasts hours	0

regression models and parameters adjustment was performed to improve the results of the selected algorithms.

10. Model building

From the total of 721 cases, 504 (70%) were used as training data and 217 (30%) as the test data, split in a stratified fashion. For the modeling, we selected the following algorithms: Logistic Regression, Decision Tree, Bagging Classifier, Random Forest, AdaBoost Classifier, Gradient Boosting, K-NN-classifier and Stacking, using K-fold cross validation. In this case we evaluate the performance of the algorithms based on the test recall parameter (see 5th column in tables). The recall parameter indicates the presence of false negative. We chose this metric since the goal for this paper was to determine the diagnosis and decrease the chances of misdiagnosis or false negatives. For D1, the best performance achieved were with the Tuned Decision Tree and the Tuned Random Forest Model (Table 1 – top two data rows). Both obtained an estimated test recall of 98%, and the chosen algorithm model was Random Forest, even though when the overfitting was lower in Decision tree model. For this diagnosis we prioritized the better performance of Random Forest using secondary criteria (Test Precision and F1 scores). In addition, Random Forest modeling is a collection of decision trees trained in different parts of the data using a random subset of

Table 5
Performance of D4 algorithms.

Model	Train accuracy	Test accuracy	Train recall	Test recall	Train precision	Test precision	Train F1-score	Test F1-score
Tuned decision tree	0.8214	0.8341	0.9620	0.9559	0.6441	0.6633	0.7716	0.7831
Tuned random forest	0.8413	0.8249	0.9620	0.9265	0.6726	0.6563	0.7917	0.7683
Tuned AdaBoost classifier	0.8155	0.8341	0.9241	0.9265	0.6432	0.6702	0.7584	0.7778
Tuned bagging classifier	0.9921	0.8433	0.9810	0.8529	0.9936	0.7073	0.9873	0.7733
Gradient boosting classifier	0.9821	0.8479	0.9873	0.8235	0.9571	0.7273	0.9720	0.7724
KNN classifier	0.9821	0.8479	0.9873	0.8235	0.9571	0.7273	0.9720	0.7724
Tuned logistic regression	0.9484	0.8618	0.9684	0.7941	0.8793	0.7714	0.9217	0.7826
Tuned gradient boosting classifier	0.9702	0.8525	0.9684	0.7647	0.9387	0.7647	0.9533	0.7647
Random forest	1.0000	0.8433	1.0000	0.7500	1.0000	0.7500	1.0000	0.7500
Decision tree	1.0000	0.7834	1.0000	0.7059	1.0000	0.6400	1.0000	0.6713
AdaBoost classifier	0.9147	0.8387	0.8608	0.6912	0.8662	0.7705	0.8635	0.7287
Logistic regression	0.9583	0.8295	0.9367	0.6765	0.9308	0.7541	0.9338	0.7132
Bagging classifier	0.9881	0.8065	0.9620	0.6765	1.0000	0.6970	0.9806	0.6866
Stacking classifier	0.8194	0.8111	0.7405	0.6471	0.7006	0.7213	0.7200	0.6822
Tuned KNN classifier	0.8690	0.7604	0.7278	0.5294	0.8333	0.6429	0.7770	0.5806

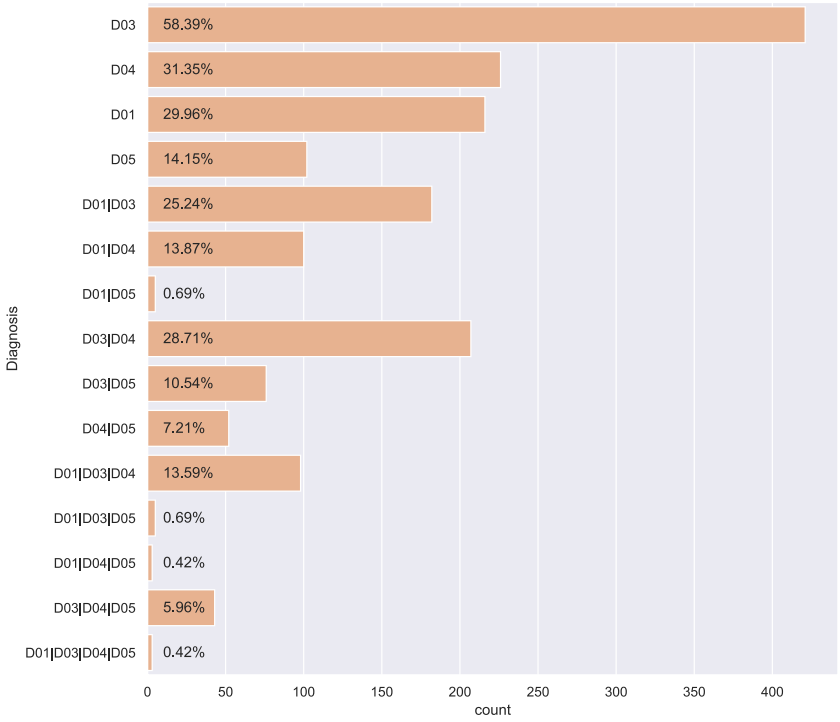

Fig. 2. Case counts for D1, D3, D4 and D5.

Table 7
Performance of D5 algorithms.

Model	Train accuracy	Test accuracy	Train recall	Test recall	Train precision	Test precision	Train F1-score	Test F1-score
Tuned decision tree	0.9782	0.9677	0.8904	0.8966	0.9559	0.8667	0.9220	0.8814
Tuned random forest	0.9782	0.9677	0.8904	0.8966	0.9559	0.8667	0.9220	0.8814
Bagging classifier	0.9940	0.9631	0.9589	0.8966	1.0000	0.8387	0.9790	0.8667
Tuned bagging classifier	0.9921	0.9631	0.9589	0.8966	0.9859	0.8387	0.9722	0.8667
Gradient boosting classifier	1.0000	0.9631	1.0000	0.8966	1.0000	0.8387	1.0000	0.8667
Tuned gradient boosting classifier	0.9960	0.9631	0.9726	0.8966	1.0000	0.8387	0.9861	0.8667
KNN classifier	1.0000	0.9631	1.0000	0.8966	1.0000	0.8387	1.0000	0.8667
Tuned logistic regression	1.0000	0.9309	1.0000	0.8621	1.0000	0.6944	1.0000	0.7692
Stacking classifier	0.9821	0.9539	0.8904	0.8621	0.9848	0.8065	0.9353	0.8333
Logistic regression	1.0000	0.9447	1.0000	0.8276	1.0000	0.7742	1.0000	0.8000
Random forest	1.0000	0.9677	1.0000	0.8276	1.0000	0.9231	1.0000	0.8727
AdaBoost classifier	1.0000	0.9447	1.0000	0.8276	1.0000	0.7742	1.0000	0.8000
Tuned AdaBoost classifier	1.0000	0.9447	1.0000	0.8276	1.0000	0.7742	1.0000	0.8000
Decision tree	1.0000	0.9355	1.0000	0.7586	1.0000	0.7586	1.0000	0.7586
Tuned KNN classifier	0.9345	0.8802	0.6986	0.5172	0.8226	0.5556	0.7556	0.5357

features to reduce variance. Each Decision tree makes its own prediction and to decide the final prediction of the model, it requires a majority vote.

The variables of importance for the tuned Random Forest model for D1 are demonstrated in Table 2. Three variables that were critical to this diagnosis include a combination of subjective and objective data. By far the most important variable was the examination-based variable called Joint Clicking Positive. This variable was considered Yes if either the right or left side TMJ exhibited clicking sounds on exam. In the subjective realm, a Yes answer to the question “Does your jaw make clicking sounds?” was found to be of secondary importance and the examination variable of right-side clicking was also included in the model. The remaining variables in Table 2 had far less importance but are listed.

For D3 the Bagging Classifier Tuned gave us a recall of 1 (Table 3), but it is just classifying everyone as D3 positive, so the performance is not appropriate. Tuned Logistic Regression, Gradient Boosting Classifier, Tuned Gradient Boosting Classifier and K-NN have the same value in terms of recall. We selected Tuned Logistic Regression as our best model as it is less overfitted than the other models (the difference between the train recall and test recall is lower).

For the diagnosis D3 (Masticatory and/or Cervical Myalgia), there were multiple important variables drawn from the history of present illness, location of the chief complaint and from the examination needed to distinguish this diagnosis for all other diagnoses. The top six variables with the highest logistic regression odds ratios (greater than 0.25) in the case of D3 are seen in Table 4. The variable that was most important was the location variable (Jaw Muscles). The second variable was a tenderness examination variable based on palpation of the right masseter. The third variable of importance was a YES answer to the question “Do you have sore jaw muscles?”. The fourth variable was a tenderness examination variable left masseter. The fifth variable in the table was an examination variable called Masticatory Muscle Tenderness Score. This variable was a continuous numerical variable produced by adding all tenderness scores from the right and left masseter and temporalis muscles. The sixth variable in Table 4 was a chief complaint

Table 8
Variable importance for D5.

Variable	Importance
Exam: TMJ crepitation positive	0.736588
HPI: Question #03-crunching	0.092535
Exam: TMJ-left crepitation	0.077404
HPI: Question #02-clicking	0.047895
Exam: Osteoarthritis on X-ray (left-TMJ)	0.013063
Dx: Trigeminal neuropathy	0.006261
Exam: Pain free opening (mm)	0.004961
Exam: TMJ-left click	0.00419
Exam: Joint hypermobility (Beighton's score >5)	0.003579
Exam: TM joint tenderness score	0.002056

of pain. The remaining variables in Table 4 had far less importance but are listed.

For D4 (arthralgia/capsulitis) the algorithm chosen was Tuned decision tree since without a doubt it provided this models performance gave the best test recall (Table 5).

The variables of importance discovered using the tuned decision tree model for TMJ arthralgia/capsulitis (D4), are seen in Table 6. The variable that was most important was and examination variable called TM Joint Tenderness Score (considered positive if either side the right or left TMJ exhibited severe level tenderness or both sides exhibited moderate tenderness). The individual exam variable, TMJ tenderness on the left side was an important variable, but not the right side. All of the subjective data variables collected from the history of present illness (HPI) seen in Table 6 were found to be irrelevant to the final algorithms.

Finally, for D5 (TMJ arthritis) Tuned Decision tree and random forest gave us the same metrics. We again chose the Tuned Random Forest as our best Model to create our algorithm, because it implements many decision trees to make the final estimation. Other ML models that are overfitted as they have bigger values on the training data than on the test data (Table 7).

The variables of importance for the tuned Random Forest in the case of TMJ osteoarthritis (D5) are shown in Table 8. The most important variable was an examination variable called joint crepitation positive exhibited the highest importance. This variable was considered Yes if either the right or left side TMJ exhibited crepitation sounds on examination. In the subjective data realm, a Yes answer to the question "Does your jaw make crunching sounds?" was also found to be important. Interesting extra oral exam of left side TMJ crepitation was an important variable but really should be discounted since it was a component of the Joint Crepitation Positive score. All other variable in the Table 8 since in Table 8 had an importance level of less than 0.05 were not considered.

10.1. Algorithm tuning

With any machine learning analysis, the selected classification parameters used in an algorithm must be adjusted to achieve the best results. For example, in the tuned Random Forest, when we analyzed D1, the adjusted parameters included a maximum depth = 4, maximum features = none, n_estimators = 80. The recall obtained was 98.4% on the test data and 99.3% in training data. For the second diagnosis D3 the Logistic Regression Tuned Parameters: solver: liblinear, Threshold: 0.3089786703216764. For D4 Decision Tree Tuned Parameters: maximum depth = 2, maximum leaf nodes = 3, minimum_impurity_decrease = 0.0001 and lastly for D5 the tuned Random Forest parameters were maximum depth = 2, maximum features = none, n_estimators = 80.

The four diagnoses examined in this review are the most common ones given to patient who attends an Orofacial Pain specialty clinic. As it turns out, three of these diagnoses (D1, D4 and D5) have a relatively simple one, two or at most three-node decision tree that can be used to define these diagnoses. The fourth diagnosis (D3) is more complex and has 4 nodes in a decision tree. The variables of greatest importance for defining myalgia are primarily examination-based variables. As we have discovered, when a ML classification model is not as robust as hoped for, even after tuning the model, then additional features may be needed in the dataset to better make an accurate classification. Alternatively, if two diagnoses cannot be separated, they might need to be combined. Only by ongoing careful examination of feature data, combined with expertise in the domain, will you identify which variables are missing, which are redundant, and which can be removed for the data collection process. With this knowledge, decisions to amend the note-taking protocol can be made and new algorithms created.

11. Conclusions

In summary, whether you believe a branching hybrid check-box based data collection system with built-in algorithms is needed, depends on your individual agenda. If you have no plans for data analysis or publishing about the various diagnostic phenotypes discovered and you do not need pop-up suggestions for best diagnosis and treatment options, it is easier to use a semi-structured narrative note for your patient encounters. If you want data-driven diagnostic and disease risk algorithms and pop-up best-treatment options, then you need a highly structured data collection system that is compatible with machine learning analysis. Our selected data collection system was hybrid because patient stories vary greatly, you must have the ability to write explanatory narrative notes. Any data collection system must allow branching since it should be easy to operate and have a low click burden. Many healthcare givers are already resentful of increased time spent with EMR software instead of the patient. The EMR is an essential component of healthcare, but it needs to be logical, save time, prevent diagnostic oversights and provide useful treatment suggestions.

Automating the journey from data collection to diagnoses has the potential to improve standards of care by providing faster and reliable predictions. In addition, predictions can inform clinicians-in-training by relating important combinations of variables to potential diagnoses. The data presented in this paper is only preliminary and we are actively creating and validating a larger set of OFP diagnostic prediction algorithms. In our future work, we will examine how ML approaches and classifier metrics can be used to support both differential and combinatorial diagnoses and by extending the number of cases this will allow us to consider a wider set of OFP diagnostic algorithms. We will seek to expand and improve our algorithms once we upscale the dataset to prove feasibility with all of our additional preliminary OFP diagnoses.

Role of the funding source

No funding.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported internally by Orofacial Pain and Oral Medicine Center at the Herman Ostrow School of Dentistry of USC. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Hardeep S, Schiff GD, Graber ML, Onakpoya I, Thompson MJ. The global burden of diagnostic errors in primary care. BMJ Qual Saf 2017;26(6):484–94.
[2] Ball JR, Balogh E. Improving diagnosis in health care: highlights of a report from the national academies of sciences, engineering, and medicine. Ann Intern Med 2016;164(January (1)):59–61.
[3] Topol EJ. The future of medicine is in your smartphone. Wall Street J 2015;(January).
[4] Bianchi JRA, Prieto J, Li T, Soroushmehr R, Najarian K, Gryak J, et al. Decision support systems in temporomandibular joint ostteoarthritis: a review of data science and artificial ingelligence applications. Semin Orthod 2021;27(June (2)):78–86.
[5] Herper M. MD Anderson benches IBM Watson in setback for artificial intelligence in medicine. Forbes. 2017 February 19.
[6] Verdell E, Bou-Crick C. VisualDx: a visual diagnostic decision support tool. Med Ref Serv Q 2012;31(4):414–24.
[7] London S. DXplainTM: a web-based diagnostic decision support system for medical students. Med Ref Serv Q 1998;17(2):17–28.
[8] Vardell E, Moore M. Isabel, a clinical decision support system. Med Ref Serv Q 2011;30(2):158–66.
[9] Lemaire JB, Schaefer JP, Martin LA, Faris P, Ainslie MD, Hull RD. Effectiveness of the Quick Medical Reference as a diagnostic tool. CMAJ 1999;161(September (6)):725–8.
[10] Warner HR, Haug P, Bouhaddou O, Lincoln M, Warner Jr H, Sorenson D, et al. ILIAD as an expert consultant to teach differential diagnosis. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. 1988. p. 371–6.
[11] Riches N, Panagiot M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. PLoS One 2016;11(March (3)):e0148991.
[12] Elkin PL, Schlegel DR, Anderson M, Komm J, Ficheur G, Bisson L. Artificial intelligence: Bayesian versus Heuristic method for diagnostic decision support. Appl Clin Inform 2018;9(April (2)):432–9.
[13] Elkin PL, Barnett GO, Famiglietti KT, Kim RJ. Closing the loop on diagnostic decision support systems. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. 1990. p. 589–93.
[14] Loeb GE. A new approach to medical diagnostic decision support. J Biomed Inform 2021;116(March):103723.
[15] Benoliel R, May A, Svensson P, Pigg M, Alstergren P, Baad-Hansen L, et al. International classification of orofacial pain, 1st edition (ICOP). Cephalalgia 2020;40(February (2)):129–221.
[16] Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. N Engl J Med 2017;376(June (26)):2507–9.
[17] Weng WH. Machine learning for clinical predictive analytics. In: Celi L, Majumder M, Ordóñez P, Osorio J, Paik K, Somai M, editors. Leveraging data science for global health. Cham: Springer; 2020. p. 199–217.
[18] Karystianis G, Nevado AJ, Kim CH, Dehghan A, Keane JA, Nenadic G. Automatic mining of symptom severity from psychiatric evaluation notes. Int J Methods Psychiatr Res 2018;27(March (1)):e1602.
[19] Khayamnia M, Yazdchi M, Heidari A, Foroughipour M. Diagnosis of common headaches using hybrid expert-based systems. J Med Signals Sens 2019;9(September (3)):174–80.
[20] Sahoo AJ, Kumar Y. Seminal quality prediction using data mining methods. Technol Health Care 2014;22(August (4)):531–45.
[21] McCartney S, Weltin M, Burchiel KJ. Use of an artificial neural network for diagnosis of facial pain syndromes: an update. Stereotact Funct Neurosurg 2014;92(January (1)):44–52.
[22] Limonadi FM, McCartney S, Burchiel KJ. Design of an artificial neural network for diagnosis of facial pain syndromes. Stereotact Funct Neurosurg 2006;84(November (5–6)):212–20.
[23] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018;24(October (11)):1716–20.
[24] Kwon J, Lee H, Cho S, Chung CS, Lee MJ, Park H. Machine learning-based automated classification of headache disorders using patient-reported questionnaires. Sci Rep 2020;10(August (1)):14062.
[25] Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-Based computational phenotyping. EEE/ACM Trans Comput Biol Bioinf 2019;16(January–February (1)):139–53.
[26] Nocera L, Vistoso A, Yoshida Y, Abe Y, Nwoji C, Clark GT. Building an automated orofacial pain, headache and temporomandibular disorder diagnosis system. In: AMIA Annual Symposium Proceedings. 2020. p. 943–52.
[27] McKinney W. In: Beaugureau M, editor. Python for data analysis: data wrangling with pandas, NumPy, and IPython. 1st ed. Sebastopol, CA: O'Reilly Media, Inc.; 2012.
[28] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(October (85)):2825–30.
[29] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 1st ed. Chapman and Hall/CRC; 1984.