Caries Research

Systematic Review

Caries Res 2022;56:161–170
DOI: 10.1159/000524167

Received: October 14, 2021
Accepted: March 13, 2022
Published online: May 30, 2022

Machine Learning in the Diagnosis and Prognostic Prediction of Dental Caries: A Systematic Review

Lilian Toledo Reyes Jessica Klöckner Knorst Fernanda Ruffo Ortiz
Thiago Machado Ardenghi

Department of Stomatology, School of Dentistry, Federal University of Santa Maria, Santa Maria, Brazil

Keywords
Artificial intelligence · Dental caries · Diagnosis · Machine learning · Prognosis

Abstract
We performed a systematic review to evaluate the success of machine learning algorithms in the diagnosis and prognostic prediction of dental caries. The review protocol was a priori registered in the PROSPERO, CRD42020183447. The search involved electronic bibliographic databases: PubMed/Medline, Scopus, EMBASE, Web of Science, and grey literature until December 2020. We excluded review articles, case series, case reports, editorials, letters, comments, educational methodologies, assessments of robotic devices, and articles with less than 10 participants or specimens. Two independent reviewers selected the studies and performed the assessment of the methodological quality based on standardized scales. We summarize data on the machine learning algorithms used; software; performance outcomes such as accuracy/precision, sensitivity/recall, specificity, area under the receiver operating characteristic curve (AUC), and positive/negative predictive values related to dental caries. Meta-analyses were not performed due to methodological differences. Our review included 15 studies (10 diagnostic studies and 5 prognostic prediction studies). Cross-sectional design studies were predominant (12). The most frequently used statistical measure of performance reported in diagnostic studies was AUC value, which ranged from 0.745 to 0.987. For most diagnostic studies, data from contingency tables were not available. Reported sensitivities were higher in low risk of bias prognostic prediction studies (median [IQR] of 0.996 [0.971–1.000] vs. unclear/high risk of bias studies 0.189 [0–0.340]; *p* value 0.025). While there were no significant differences in the specificity between these subgroups, we concluded that the use of these technologies for the diagnosis and prognostic prediction of dental caries, although promising, is at an early stage. The general applicability of the evidence was limited given that most models were developed outside the real clinical setting with a prevalence of unclear/high risk of bias. Researchers must increase the overall quality of their research protocols by providing a comprehensive report on the methods implemented.

Karger@karger.com
www.karger.com/cre

Correspondence to:
Thiago Machado Ardenghi, thiardenghi@hotmail.com

Karger

Introduction

Recent epidemiological reports rank untreated caries in permanent teeth as the most prevalent health condition worldwide [Peres et al., 2019], generating a high economic impact [Listl et al., 2015]. In addition to material impacts, oral diseases also have numerous consequences on the well-being and quality of life of individuals [Feldens et al., 2016]. For this reason, new strategies must be developed focused on reducing the disease, decreasing its social impacts, and lowering the health costs generated by dental caries.

Currently, evidence regarding the treatment of dental caries supports the preventive and conservative approach, moving away from the surgical paradigm of removal and replacement of the affected tissue [Pitts and Stamm, 2004; Selwitz et al., 2007]. In accordance with the foregoing statement, efforts should be directed to improve the alternatives of risk assessment and prediction of dental caries. Also, attention should be paid to the refinement of methods for early detection of disease progression, which can lead to less invasive treatments [Schwendicke et al., 2015].

In recent years, advances have been documented with the introduction of artificial intelligence (AI) algorithms within the field of health care in support of prognostic, diagnostic, and decision-making, contributing to the reduction of medical errors [Rajkomar et al., 2019; Shan et al., 2021]. AI could be described as the non-biological capacity of a machine that tries to mimic human intelligence in the development of complex tasks, such as problem solving, recognition of objects, words, and decision-making. As a major arm of AI, we find machine learning (ML) where models learn from examples rather than preprogrammed tables of rules [Shan et al., 2021].

Recent studies have illustrated that these technologies are being introduced in the diagnosis and prognostic prediction of different stages of dental caries [Lee et al., 2018; Casalegno et al., 2019; Hung et al., 2019; Liu et al., 2020]. The expansion of these studies to daily clinical practice may lead to the improvement of health care services in the near future. However, the use of these new methods in support of dental caries must be validated for reliability. Currently, there is not a complete body of evidence available on the scope of ML algorithms to aid in these tasks. As soon as some studies begin to report performance analysis of ML algorithms in general dentistry [Park and Park, 2018; Shan et al., 2021], a more detailed synthesis is needed, specifically on their success rate in the diagnosis and prognostic prediction of dental caries. Thus, the reproduction and generalization of research that implements these techniques would be more objective. In this context, the aim of the present study was to systematically evaluate the performance of ML algorithms in the diagnosis and prognostic prediction of dental caries, highlighting the specific algorithms that have been proposed to date.

Methods

The present systematic review was reported in accordance with the checklist of preferred reporting items for systematic reviews and meta-analyses (PRISMA) (online suppl. Material 1; for all online suppl. material, see www.karger.com/doi/10.1159/000524167). The protocol study was registered at "PROSPERO: International prospective register of systematic reviews," (registration number: PROSPERO 2020 CRD42020183447).

Review Question

What is the current impact of ML algorithms on the diagnosis and prognostic prediction of dental caries?

Eligibility Criteria

We considered eligible those studies with a cross-sectional or longitudinal design that used AI technologies in the diagnosis or prognostic prediction of dental caries and reported accuracy measures: sensitivity, specificity, or area under the receiver operating characteristic curve (AUC). The analysis was guided by the following PICO elements:

Population (P)

Data set obtained from human subjects (radiographic, photographic, or near-infrared light transillumination [NILT] images, and medical records).

Intervention (I)

Diagnostic or prognostic prediction of dental caries assisted by non-logistic regression (non-LR) ML algorithms.

Comparator (C)

Expert's judgment, clinical/histological examination, classifiers reference as logistic regression (LR).

Outcome (O)

Analysis of ML performance in detection, diagnosis, or prognostic prediction of dental caries (outcomes such as accuracy/precision, sensitivity/recall, specificity, receiver operating characteristic curve, area under the curve, or positive/negative predictive values). The exclusion criteria were review articles, case series, case reports, editorials, letters, comments, educational methodologies, assessments of robotic devices, and articles with fewer than 10 participants/specimens.

Information Sources and Search Strategy

The search involved the following electronic bibliographic databases: PubMed/Medline, Scopus, EMBASE, and Web of Science, including publications until December 28, 2020. Unpublished literature was tracked through OpenGrey and International and American Dental Associations Research congresses. In addition,

references in retrieved papers were checked manually. There were no restrictions on the publication data and language.

The search strategy was performed for each electronic database. The keywords were combinations of medical subject headings terms and free terms. The vocabulary and syntax were adjusted for each database (online suppl. Material 2). A reference management system (Mendeley Desktop 1.17.13, Elsevier, Atlanta, GA, USA) was used to upload all the potentially eligible studies and remove duplicates.

Study Selection

The selection of studies was carried out in two phases. First, two trained and calibrated reviewers (L.T.R and J.K.K.) screened the titles and abstracts of all search results independently for potential relevance. In the second phase, the full texts of potentially relevant reports were retrieved and independently evaluated by the same two reviewers. Disagreements were resolved by consensus. If disagreement persisted, the judgment of a third reviewer (F.R.O.) was considered to be decisive. Papers that fulfilled all the selection criteria were processed for data extraction. The reliability of the reviewers was tested in 10% of the papers initially selected. The Kappa statistic was calculated and demonstrated excellent interexaminer agreement ($K = 1.00$).

Data Extraction

Extracted information included: author, publication year, country, study purpose, ML task**,** study design, data source, target condition, type of teeth, data set, training data set, validation/test data set, validation technique, reference standard/comparator, ML algorithm, software, metric outcome and its value (accuracy/precision, sensitivity/recall, specificity, receiver operating characteristic curve, area under the curve, positive/negative predictive values) (online suppl. Tables 1–4). When more information was required, we tried to contact the authors by email. One reviewer (L.T.R.) extracted the data from the included studies. A second reviewer (F.R.O) independently verified the extracted data, and consensus was achieved through discussion (with a third reviewer when necessary).

Risk of Bias (Quality) Assessment

The quality of the studies included was assessed by two reviewers, independently (L.T.R. and J.K.K.). The diagnostic studies were based on the quality assessment of diagnostic accuracy studies (QUADAS-2) [Whiting et al., 2011]. The quality and applicability of prediction studies were measured on the prediction model risk of bias assessment tool [Wolff et al., 2019]. For the consideration of applicability, in addition to the previous guidelines, criteria such as the use of an independent set to reduce model overfitting and/or verification of model performance through external validation were added to these analyses [Park and Han, 2018]. The studies were rated as "low quality (LRB)," "high quality (HRB)," or "unclear (?)" according to concerns regarding the risk of bias and applicability. Data on study quality were summarized by applying the Cochrane risk of bias tool Review Manager (RevMan) (software). Version 5.4.1, the Cochrane Collaboration, 2020.

Strategy for Data Synthesis

We provided a narrative synthesis from the included studies. For analyses, diagnostic and prognostic prediction studies were treated separately. Accuracy measures (sensitivity, specificity, and AUC) were reported when available. Contingency tables consisted of true-positive, false-positive, true-negative, and false-negative results and were used to calculate sensitivity and specificity. For the studies that reported sensitivities and specificities, if 2×2 tables were not available, we back calculated counts based on reported accuracy measures. If a study provided multiple contingency tables for different algorithms, we assumed that these were independent of each other.

A meta-analysis was not undertaken due to the great clinical and methodological heterogeneity between these studies. Most of the diagnostic studies lack the raw data required for the diagnostic accuracy measures of the meta-analysis. The forest plots were used to visually assess heterogeneity in all prognostic prediction studies, reporting sensitivity, specificity, and true positives. All statistics were done considering a 95% confidence interval. Study-level factors such as: risk of bias (low or unclear/high) and algorithm type (non-LR ML or LR) were evaluated as potential determinants of the reported accuracy measures through the boxplot graph, using the Mann-Whitney U test. Level of significance was considered at $p < 0.05$. The data were processed using RStudio software version 1.4.1717.

Results

Study Selection

In the present review, a total of 491 unique studies were retrieved, 8 additional records were identified through other sources, and 22 duplicates were removed. Based on the title and abstract screening phase, 441 papers were excluded. Subsequently, the full-text analysis was performed on 28 potentially relevant studies. From this list, 13 studies were excluded for the following reasons: no data about the algorithm used (3), related research, but not properly a caries diagnostic or prognostic prediction study (10). Finally, this systematic review compiled a sample of 15 studies (Fig. 1).

Study Characteristics

The basic characteristics of the studies were summarized in online supplementary Tables 1–4. All studies were published in a frame of time between 2008 and 2020 (6 corresponded to the last year), and several of these studies used secondary data. Among the 15 articles included, the studies represented by China and Japan were the most frequent (3 from each country) [Tamaki et al., 2009; Araki et al., 2010; Ito et al., 2011; Liu et al., 2020; Wang et al., 2020; Zhang et al., 2020]. The other 9 studies come from Brazil, Greece, the USA, Korea, Switzerland, and Mexico (online suppl. Table 1). There were 10 reports evaluating the usefulness of ML in the development of diagnostic models, while 5 focused on the development of dental caries prognostic prediction models. The diag-

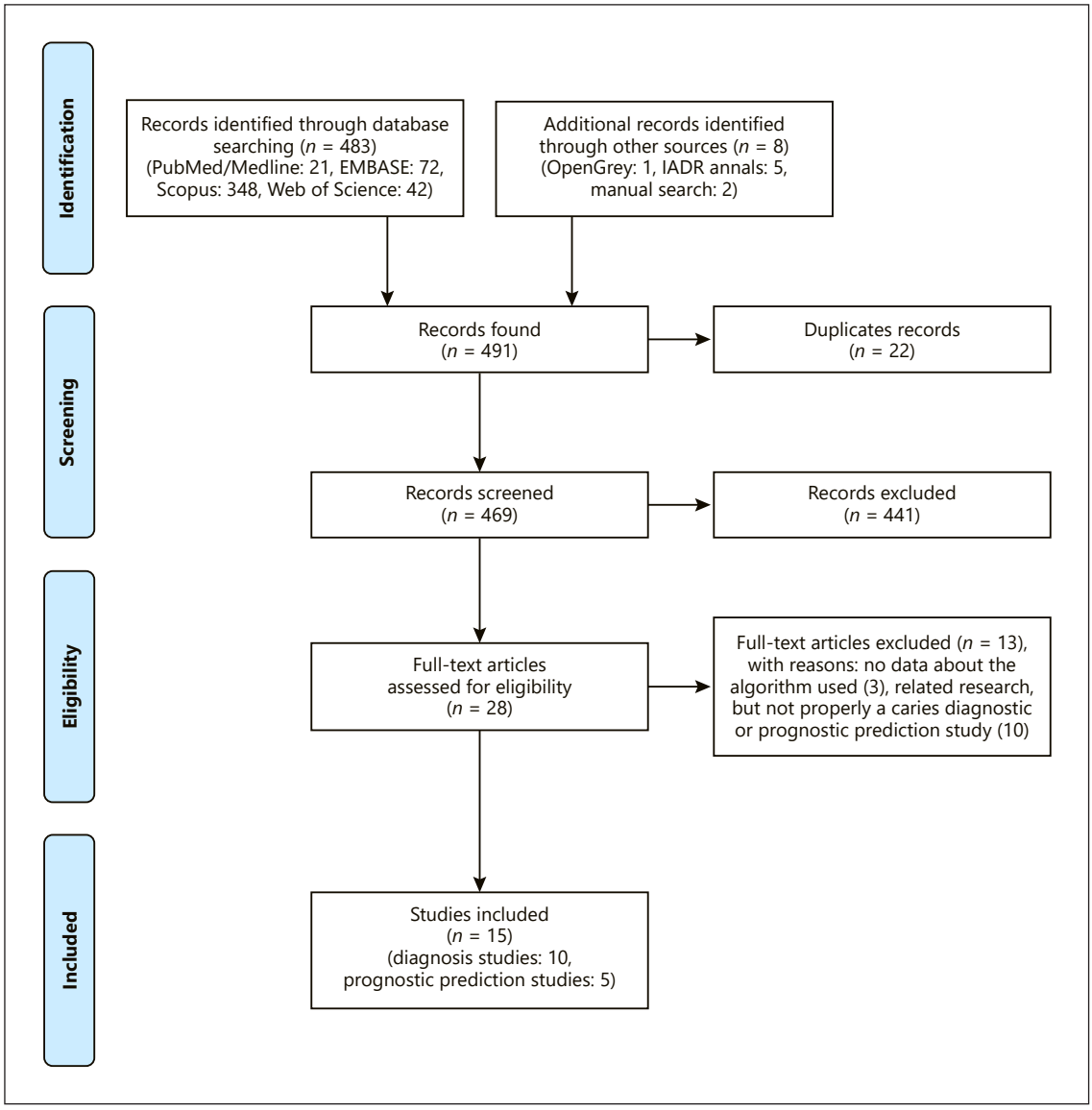

Fig. 1. PRISMA flow chart displaying the selection process.

nostic studies were those where the target condition measured (dental caries) was actually present when contrasted with the standard reference. Our review included the identification of the outcome via examination of different types of images such as radiographic images, photographic images, or images obtained through the NILT. The so-called prognostic prediction studies were those that used ML methods to model the probability; risk of the appearance; progression of dental caries based on the evaluation of a broader set of potential predictors; called prognostic factors; or prognostic indicators. Typically, decisions about which prognostic factors would be relevant and should be considered in advance were based on the primary studies, including clinical experience and knowledge of the biology of the disease. Most of these articles identified prognostic factors that are subsequently used in a classification scheme [Laupacis et al., 1994; Park and Han, 2018; Wolff et al., 2019].

Diagnostic Studies

Cross-sectional design studies were predominant in these studies (9) [Devito et al., 2008; Berdouses et al., 2015; Lee et al., 2018; Casalegno et al., 2019; Cantu et al., 2020; Geetha et al., 2020; Schwendicke et al., 2020; Wang

et al., 2020; Zhang et al., 2020]. Included diagnostic studies mostly evaluated proximal dental caries [Devito et al., 2008; Araki et al., 2010; Lee et al., 2018; Cantu et al., 2020; Geetha et al., 2020; Schwendicke et al., 2020]. One diagnostic study was conducted on proximal and occlusal surfaces [Casalegno et al., 2019] and another only on occlusal surfaces [Berdouses et al., 2015]. Two other studies used occlusal and smooth caries as target conditions [Wang et al., 2020; Zhang et al., 2020]. In one study, the interest in detecting white spot lesions in combination with dental plaque through fluorescence variations in dental images using a dual-channel imaging system was revealed [Wang et al., 2020].

Several image types and formats comprised data sets for the diagnostic studies, such as conventional or digital radiographic images (5 studies) [Devito et al., 2008; Araki et al., 2010; Lee et al., 2018; Cantu et al., 2020; Geetha et al., 2020], photographic images were collected in 3 studies [Berdouses et al., 2015; Wang et al., 2020; Zhang et al., 2020], and images obtained through the NILT were used in 2 reports [Casalegno et al., 2019; Schwendicke et al., 2020]. The mean set size of these reports was 1,873 images (range 100–7,200).

Expert's judgment was commonly used as standard reference. Additionally, some studies chose to use other methods in combination, such as micro-CT and histology examination [Devito et al., 2008; Araki et al., 2010].

The cross-validation technique was mainly applied as part of the internal validation process (5 studies) [Devito et al., 2008; Berdouses et al., 2015; Casalegno et al., 2019; Geetha et al., 2020; Schwendicke et al., 2020]. The split validation was used in 4 studies [Lee et al., 2018; Cantu et al., 2020; Wang et al., 2020; Zhang et al., 2020]. One study quantified the performance of these resources in external data [Araki et al., 2010].

The most frequently reported statistical measure of performance in diagnostic studies was the AUC value (online suppl. Table 2) evaluated in 8 studies (72.7% of the total diagnostic studies surveyed) [Devito et al., 2008; Araki et al., 2010; Berdouses et al., 2015; Lee et al., 2018; Casalegno et al., 2019; Geetha et al., 2020; Schwendicke et al., 2020; Zhang et al., 2020]. This metric varied from 0.740 to 0.987 for the models that showed the highest performance in each diagnostic study. Among the 5 studies focused on proximal caries lesion detection [Devito et al., 2008; Araki et al., 2010; Lee et al., 2018; Cantu et al., 2020; Schwendicke et al., 2020], 4 reported variations of AUC values from 0.74 to 0.917 (mean 0.835, SD 0.112). Regarding caries classification tasks, 3 of the 4 studies identified for this purpose reported AUC values ranging from 0.857 to 0.987 [Berdouses et al., 2015; Geetha et al., 2020; Zhang et al., 2020]. The highest value corresponded to the caries classification on the proximal surfaces of teeth. AUC values for the classification of lesions on occlusal surfaces and for the classification of caries on occlusal/smooth surfaces together were 0.98 and 0.856, respectively. The task of segmentation was the objective of one study, with an AUC value of 0.836 for occlusal lesions and 0.856 for proximal lesions [Casalegno et al., 2019].

Within the diagnostic reports, the most popular algorithms were variants of artificial neural networks (ANNs), including 9/10 studies [Devito et al., 2008; Araki et al., 2010; Lee et al., 2018; Casalegno et al., 2019; Cantu et al., 2020; Geetha et al., 2020; Schwendicke et al., 2020; Wang et al., 2020; Zhang et al., 2020]. We found different software and programming languages supporting the implementation of these algorithms, with Python being the most commonly used (3 studies) [Lee et al., 2018; Cantu et al., 2020; Schwendicke et al., 2020].

Prognostic Prediction Studies

Cross-sectionally collected data were used in 3 studies [Zanella-Calzada et al., 2018; Hung et al., 2019; Liu et al., 2020]. Two studies used prospective data [Tamaki et al., 2009; Ito et al., 2011].

These studies mostly interpreted dental caries as binary values in permanent teeth [Tamaki et al., 2009; Ito et al., 2011; Liu et al., 2020]. Furthermore, one study also considered this outcome in deciduous teeth [Zanella-Calzada et al., 2018]. Another study evaluated the target condition specifically on the root surfaces of permanent teeth [Hung et al., 2019].

The data sets for prognostic prediction studies were collected from medical and dental records of different populations: pre-elementary schoolchildren (5–6 years) [Tamaki et al., 2009], adults [Ito et al., 2011; Hung et al., 2019], the geriatric population (65–74 years) [Liu et al., 2020], and a population containing people aged 0–80 years [Zanella-Calzada et al., 2018]. The mean set size in these studies was 4,061 units (range 560–9,812). Most of these studies applied LR comparatively (4/5) [Tamaki et al., 2009; Ito et al., 2011; Hung et al., 2019; Liu et al., 2020].

Independent data sets (hold-out) were used in three of these reports to validate the models [Zanella-Calzada et al., 2018; Hung et al., 2019; Liu et al., 2020]. In one study, cross-validation was performed [Tamaki et al., 2009]. Another article did not have information about validation data [Ito et al., 2011].

Among the prognostic prediction studies, 4/5 provided sufficient information to enable the calculation of con-

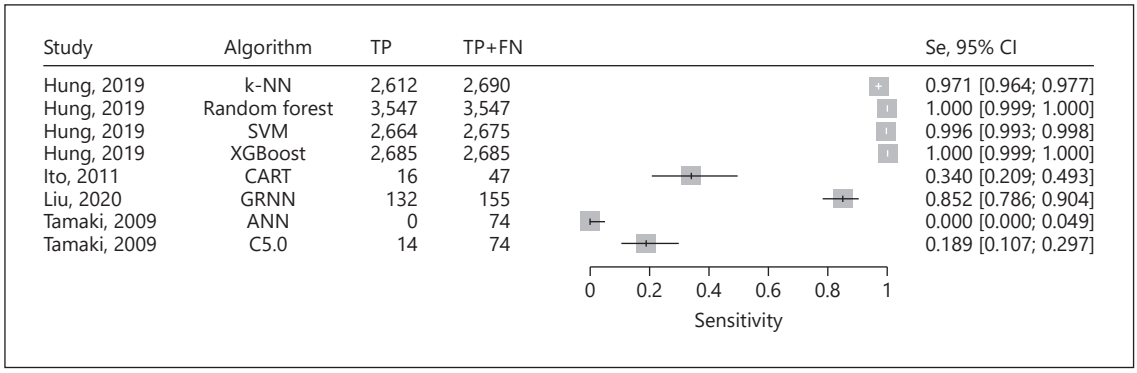

Fig. 2. Sensitivity forest plot for predictive studies.

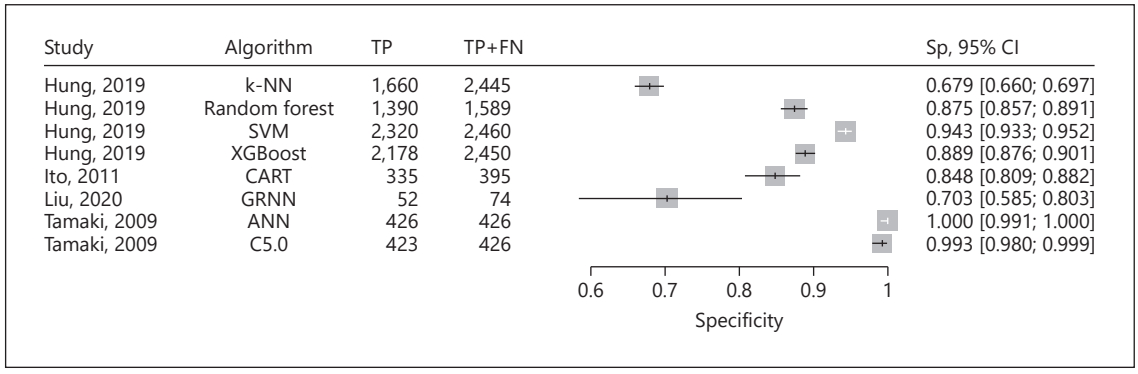

Fig. 3. Specificity forest plot for predictive studies.

tingency tables and the calculation of test performance parameters for a total of 11 tables across these studies (8 using non-LR ML algorithms and 3 using LR) (online suppl. Table 4). Within this group, the sensitivity of the ML models ranged from 0.000 to 0.971 (mean 0.669, SD 0.420) and the specificity ranged between 0.679 and 0.993 (mean 0.866, SD 0.121) derived from internally validated test samples (Fig. 2, 3). The report of the sensitivity value of 0.000 when using the ANN was derived from one of the studies ranked as high risk of bias [Tamaki et al., 2009].

Low risk of bias studies had higher reported sensitivities (median [IQR] of 0.996 [0.971–1.000] vs. unclear/high risk of bias studies of 0.189 [0–0.340]; *p* value 0.025). While there were no significant differences in the specificity of these subgroups (low risk of bias of 0.875 [0.703–0.889] vs. unclear/high risk of bias of 0.993 [0.848–1.000]; *p* value 0.179 (online suppl. Fig. 1), these metrics were also higher in low risk of bias studies using non-LR ML algorithms: sensitivity 0.996 [0.971–1,000], specificity 0.875 [0.703–0.889] versus studies using LR: sensitivity 0.808 [0.771–0.845], specificity 0.511 [0.311–0.711]. We found no significant difference in the comparison sensitivity (*p* value 0.051) and specificity (*p* value 0.245) (online suppl. Fig. 2).

Overall, three of these studies used more than 1 ML algorithm (online suppl. Table 3). The use of RNA classifiers was reported in 3 articles [Tamaki et al., 2009; Zanella-Calzada et al., 2018; Liu et al., 2020]. Support-vector machine (SVM), k-nearest neighbors, classification trees (e.g., CART, C5.0), random forest, and gradient boosting were also implemented, sometimes concurrently. The Python language was the most frequently used in the studies to implement these algorithms.

Risk of Bias in Individual Studies

Risk of Bias in Diagnostic Studies

Our findings indicated the high quality of one study associated with dental caries diagnostic in the comprehensive evaluation of all domains [Cantu et al., 2020]. On the other hand, concerns regarding the applicability were identified in the patients' selection domain, where 8 reports were rated as "high" or "unclear." This ranking was

due to the poor representativeness of the data set for various reasons: small data set size [Devito et al., 2008; Araki et al., 2010; Berdouses et al., 2015; Casalegno et al., 2019; Geetha et al., 2020; Schwendicke et al., 2020], data type (ex vivo), concern regarding the use of inappropriate exclusion criteria [Devito et al., 2008; Araki et al., 2010; Berdouses et al., 2015], and failures in enrollment strategy [Wang et al., 2020; Zhang et al., 2020]. In 2 diagnostic studies, this domain was rated as "low" [Lee et al., 2018; Cantu et al., 2020].

The index test domain evaluation showed risk bias "high" or "unclear" for 3 studies [Araki et al., 2010; Wang et al., 2020; Zhang et al., 2020] and maintained concerns regarding the applicability for 8 studies rated as "unclear" [Devito et al., 2008; Araki et al., 2010; Berdouses et al., 2015; Lee et al., 2018; Casalegno et al., 2019; Geetha et al., 2020; Wang et al., 2020; Zhang et al., 2020]. This classification was based on the limitations of some studies in describing the test in detail, lack of external validation, or lack of tests in real-world settings. Differences in the versions of the system for the implementation of the algorithm and difficulties in its use by professionals were reported in one of the studies [Araki et al., 2010].

The majority of these studies had "unclear" for risk bias in the reference standard domain. The lack of transparency in some reports on the selection and training of experts led to this judgment [Araki et al., 2010; Berdouses et al., 2015; Lee et al., 2018; Casalegno et al., 2019; Geetha et al., 2020; Wang et al., 2020; Schwendicke et al., 2020]. Flow and timing were not affected (online suppl. Fig. 3, 4).

Risk of Bias in Prognostic Studies

Concerning prognostic predictive studies (online suppl. Fig. 5, 6), two papers were rated as "high quality" in the applicability overall consensus [Hung et al., 2019; Liu et al., 2020]. Participant domain had a "high risk" of bias in two studies where details about the enrollment and sampling strategy showed flaws [Tamaki et al., 2009; Ito et al., 2011]. These studies were also rated as "high risk" in the statistical analysis domain. Specifically, the sensitivity value of 0.000, which renders the model completely useless, was reported only in one of these studies when implementing ANNs algorithms [Tamaki et al., 2009]. According to the authors, aspects such as the lack of representativeness of the sample, incidence of caries, and class imbalance could contribute to this result and are pointed out as limitations of the study.

Another study had specific problems of predictor measurement and concerns related to the definition and determination of the outcome. Consequently, it was rated as "high risk" of bias in the predictor and outcome domains [Zanella-Calzada et al., 2018].

Discussion

This systematic review focused on assessing the performance of ML algorithms in the diagnosis and prediction of prognosis of dental caries, highlighting the specific algorithms that have been implemented to date for these purposes. According to our results, the studies suggested a positive effect when non-LR ML algorithms are used. However, the majority of the diagnosis studies surveyed had a high or unclear bias risk across several domains of assessment, raising concerns about their applicability. In this sense, the use of a recently proposed checklist can be useful for both researchers and reviewers to improve the quality of dental research in this field and contribute to comparability between studies [Schwendicke et al., 2021]. Efforts were applied more frequently to the analysis of images in support of caries diagnosis, using different types of ANNs, such as convolutional neural networks. Previous reports suggest the good performance of these classifiers for the analysis of medical and dental images, which helps reduce medical errors and refine the clinical decision-making process [Bera et al., 2019; Jiang et al., 2020; Lee et al., 2020; Sultan et al., 2020; Welch et al., 2020; Kuwana et al., 2021; Shan et al., 2021].

Nevertheless, it should be noted that there are differences between the ML tasks, the level of analysis, the metrics reported, the study design, and the specificities of the algorithms implemented in these studies. These aspects affected the comparisons. The AUC was the most commonly reported metric, which has limited utility in decision-making where over- and under detection are not equally important. On the other hand, accuracy values can be affected by unbalanced classes, which is often a reality during caries detection [Reyes et al., 2021]. A more extensive understanding of the performance of these algorithms is achieved with the comprehensive reporting of metrics such as sensitivity, specificity, and contingency value tables. This may be of greater interest to the practitioner and offer greater reproducibility and confidence in the analyses [Schwendicke et al., 2019; Leite et al., 2020; Pethani, 2021; Schwendicke et al., 2021].

Regarding the prognostic prediction of dental caries, only three studies made available data on the performance of LR and non-LR ML algorithms, allowing com-

parisons. These analyses showed at the study level a better performance in terms of sensitivity of the models that implemented non-LR ML algorithms versus LR, with modest levels of significance (p = 0.05). The prediction achieved the best performance using SVM [Hung et al., 2019] and generalized regression neural network (GRNN) [Liu et al., 2020] from low-risk bias reports. A previous study in dentistry also showed satisfactory performance (AUC of 0.83) when applying SVM algorithms in Periodontology [Feres et al., 2018]. As for the GRNN approach, the regression method produces an estimated value for which it minimizes the root mean square error for the dependent variable. This classifier has good nonlinear mapping ability and learning speed. Consequently, the use of GRNN to solve classification problems showed beneficial prediction effects with only a few training samples available [Zhao et al., 2020]. However, it is worth noting the limitations in the interpretation of these models [Schwendicke et al., 2021]. In order to extend these studies to clinical dental practice, interpretable models may be preferred. Ensemble methods such as random forest and gradient boosting also revealed high performance.

Strengths and Limitations

We tried to conduct a comprehensive search, assess rigorously the quality of the studies, and adequately synthesize the data in an area that, according to our knowledge, has gained greater interest and has been little explored. In the present review, the meta-analysis could not be conducted due to the heterogeneity of the reports. In addition, the methodology and results reported were very often incomplete and unclear.

The studies were mainly based on the development of models in isolation, outside the real dental clinical environment. Diagnostic studies based on clinical data primarily collected secondary data with no primary focus on the outcome of caries. Only one study was focused on this outcome during initial target collection [Casalegno et al., 2019]. In the case of prognosis prediction studies, two studies obtained data from medical public repositories without specific dental purposes [Zanella-Calzada et al., 2018; Hung et al., 2019]. The remaining three studies used data collected in real clinical settings, which makes it more feasible to objectively test the accuracy of the models [Tamaki et al., 2009; Ito et al., 2011; Liu et al., 2020]. The availability of public data repositories in the dental field that considers the ethical requirements demanded will be a priority task in the massive implementation of these technologies by researchers and practitioners in the years to come. Consensus must also be established for the preprocessing and labeling of the data [Rajkomar et al., 2019; Park and Han, 2018].

On the other hand, the representativeness of the data was affected in most of the studies. We understand that, for experimental designs, this aspect can be a challenge. However, to guarantee the external validity of these technologies, larger studies considering representative samples should be conducted [Park and Han, 2018].

The sample included in this review was characterized by the paucity of studies carried out in real clinical settings and the lack of prospective data in the analysis. The prospective design was used in three studies. The first one focused on the diagnosis through a computer-assisted diagnostic system [Araki et al., 2010]. The other two studies were in support of caries prediction [Tamaki et al., 2009; Ito et al., 2011]. However, these articles provided weak evidence as they were rated "low quality" in the general consensus of applicability. Faced with this scenario, it is advisable to moderate excess optimism in the interpretation of findings from the available studies. Future efforts using prospective data and using randomized clinical trials, when possible, should improve levels of evidence [Park and Han, 2018; Rajkomar et al., 2019].

In the present review, a wide variability of criteria was established to determine the target condition. Currently, a wide spectrum of indices and criteria has been identified for assisted caries assessment. This variability in the definitions could lead to inconclusive results when we compare studies aimed at evaluating the performance of ML in the diagnosis and prediction of the disease. We strongly recommend multiple tests such as clinical and radiographic examinations with the aim of reducing misinterpretations and properly defining the target condition, which will be labeled as output in the training set. In this context, indicators such as the International Caries Detection and Assessment System provide valuable information on the progression of lesions and should be assumed in these analyses. In this way, initial lesions can be identified and more objective strategies prioritizing noninvasive treatments may be conducted [Reyes et al., 2021].

Some included studies also reported conflicting data regarding the choice of the reference standard and predictors used. These aspects could affect the validity of the results and future comparisons. The surveyed studies mainly use expert judgment as a reference standard. However, the experience of the examiners, methods to measure interexaminer and intraexaminer variability, or measures to reduce this variability were often not described.

The present systematic review has intrinsic limitations. Currently, there are still few studies on the subject and generally, available research has been conducted outside of the real clinical environment. Therefore, our findings were generally based on development models. Perhaps some studies could be lost, although we tried to conduct a comprehensive search strategy previously defined. Due to some inconsistencies with the terminology used in the studies, it is possible that some flaws appeared during data extraction. Nevertheless, a second reviewer independently verified the extracted data, and a consensus was achieved.

Implications for Research and Clinical Practice

This review may be useful for researchers and clinicians seeking to extend their studies to verify the performance of ML algorithms in the diagnosis and prognostic prediction of dental caries. The absence of prospective studies and protocols that evaluate the effectiveness of these technologies makes it clear that this type of research is required, before extending the results of these practices. In this context, new research on cost-effectiveness analysis, improvements in the quality of medical care, and acceptance by professionals and patients should be conducted [Pethani, 2021; Schwendicke and Krois; 2021]. With the development of new studies, systematic reviews with meta-analyses may be conducted, contributing to a more solid evidence base.

To achieve these goals, researchers should also increase the overall quality of their research protocols, providing a comprehensive report on the methods implemented and outlining the risk of bias. New efforts to ensure the development, transparency, replicability, and ethics of AI's research environment should be embodied in new guidelines for evaluating and presenting this type of research [Wiens and Shenoy, 2018; Schwendicke and Krois, 2021]. Training will be required for the use of resources and automated systems to carry out these tasks incorporating ML, which should be guided by multidisciplinary teams, where practitioners, radiologists, epidemiologists, statisticians, and public policymakers, among others, must collaborate.

We concluded that the use of these technologies for the diagnosis and prognostic prediction of dental caries is promising. However, the general applicability of the evidence was limited given that most models were developed outside of the real clinical setting with the prevalence of unclear/high risk of bias. Studies focused on predicting prognosis contributed with the best evidence. Moreover, it is essential to expand the research on the subject, carrying out validation in independent samples and contributing to developing cost-effectiveness analysis, which supports the introduction of these technologies into clinical practice.

Statement of Ethics

An ethics statement is not applicable because this study is based exclusively on published literature.

Conflict of Interest Statement

The authors have no potential conflicts of interest to declare.

Funding Sources

This study had in part financial support from the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil* (CAPES) – Finance Code 001 and the National Council for Scientific and Technological Development (CNPq) – Process 160258/2020-0.

Author Contributions

Lilian Toledo Reyes, Jessica Klöckner Knorst, and Fernanda Ruffo Ortiz contributed to data collection, analysis, and interpretation; drafted the paper; and critically reviewed the manuscript. Thiago Machado Ardenghi contributed to conception, design, and critically revised the manuscript. All authors gave their final approval and agree to be responsible for all aspects of the manuscript.

Data Availability Statement

All data generated or analyzed during this study are included in this article and its online supplementary information files. Further inquiries can be directed to the corresponding author.

References

Araki K, Matsuda Y, Seki K, Okano T. Effect of computer assistance on observer performance of approximal caries diagnosis using intraoral digital radiography. Clin Oral Investig. 2010 Jun;14(3):319–25.

Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology: new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019 Nov;16(11):703–15.

Berdouses ED, Koutsouri GD, Tripoliti EE, Matsopoulos GK, Oulis CJ, Fotiadis DI. A computer-aided automated methodology for the detection and classification of occlusal caries from photographic color images. Comput Biol Med. 2015 Jul 1;62:119–35.

Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, et al. Detecting caries lesions of different radiographic extension on bitewings using deep learning. J Dent. 2020 Sep 1; 100:103425.
Casalegno F, Newton T, Daher R, Abdelaziz M, Lodi-Rizzini A, Schürmann F, et al. Caries detection with near-infrared transillumination using deep learning. J Dent Res. 2019 Oct; 98(11):1227–33.
Devito KL, de Souza Barbosa F, Felippe Filho WN. An artificial multilayer perceptron neural network for diagnosis of proximal dental caries. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2008 Dec 1;106(6):879–84.
Feldens CA, Ardenghi TM, Dullius AI, Vargas-Ferreira F, Hernandez PA, Kramer PF. Clarifying the impact of untreated and treated dental caries on oral health-related quality of life among adolescents. Caries Res. 2016;50(4): 414–21.
Feres M, Louzoun Y, Haber S, Faveri M, Figueiredo LC, Levin L. Support vector machine-based differentiation between aggressive and chronic periodontitis using microbial profiles. Int Dent J. 2018 Feb 1;68(1):39–46.
Geetha V, Aprameya KS, Hinduja DM. Dental caries diagnosis in digital radiographs using back-propagation neural network. Health Inf Sci Syst. 2020 Dec;8(1):8–4.
Hung M, Voss MW, Rosales MN, Li W, Su W, Xu J, et al. Application of machine learning for diagnostic prediction of root caries. Gerodontology. 2019 Dec;36(4):395–404.
Ito A, Hayashi M, Hamasaki T, Ebisu S. Risk assessment of dental caries by using classification and regression trees. J Dent. 2011 Jun 1; 39(6):457–63.
Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun. 2020 Apr;40(4):154–66.
Kuwana R, Ariji Y, Fukuda M, Kise Y, Nozawa M, Kuwada C, et al. Performance of deep learning object detection technology in the detection and diagnosis of maxillary sinus lesions on panoramic radiographs. Dentomaxillofac Radiol. 2021 Jan 1;50(1):20200171.
Laupacis A, Wells G, Richardson WS, Tugwell P. How to use an article about prognosis: evidence-based medicine working group. JAMA. 1994;272(234):1994–237.
Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent. 2018 Oct 1;77:106–11.
Lee JH, Kim DH, Jeong SN. Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network. Oral Dis. 2020 Jan; 26(1):152–8.
Leite AF, Vasconcelos KF, Willems H, Jacobs R. Radiomics and machine learning in oral healthcare. Proteomics Clin Appl. 2020 May; 14(3):e1900040.
Listl S, Galloway J, Mossey PA, Marcenes W. Global economic impact of dental diseases. J Dent Res. 2015 Oct;94(10):1355–61.
Liu L, Wu W, Zhang SY, Zhang KQ, Li J, Liu Y, et al. Dental caries prediction based on a survey of the oral health epidemiology among the geriatric residents of Liaoning, China. Biomed Res Int. 2020 Dec;7:2020.
Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2018 Mar;286(3): 800–9.
Park WJ, Park JB. History and application of artificial neural networks in dentistry. Eur J Dent. 2018 Oct;12(4):594.
Peres MA, Macpherson LMD, Weyant RJ, Daly B, Venturelli R, Mathur MR, et al. Oral diseases: a global public health challenge. Lancet. 2019 Jul 20;394(10194):249–60.
Pethani F. Promises and perils of artificial intelligence in dentistry. Aust Dent J. 2021 Jun; 66(2):124–35.
Pitts NB, Stamm JW. International Consensus Workshop on Caries Clinical Trials (ICW-CCT) – final consensus statements: agreeing where the evidence leads. J Dent Res. 2004 Jul; 83(1 Suppl):125–8.
Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019 Apr 4; 380(14):1347–58.
Reyes LT, Knorst JK, Ortiz FR, Ardenghi TM. Scope and challenges of machine learning-based diagnosis and prognosis in clinical dentistry: a literature review. J Clin Transl Res. 2021 Aug 26;7(4):523.
Shan T, Tay FR, Gu L. Application of artificial intelligence in dentistry. J Dent Res. 2021 Mar; 100(3):232–44.
Selwitz RH, Ismail AI, Pitts NB. Dental caries. Lancet. 2007 Jan 6;369(9555):51–9.
Schwendicke F, Tzschoppe M, Paris S. Radiographic caries detection: a systematic review and meta-analysis. J Dent. 2015 Aug 1;43(8): 924–33.
Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: a scoping review. J Dent. 2019 Dec 1;91:103226.
Schwendicke F, Elhennawy K, Paris S, Friebertshäuser P, Krois J. Deep learning for caries lesion detection in near-infrared light transillumination images: a pilot study. J Dent. 2020 Jan 1;92:103260.
Schwendicke F, Krois J. Better reporting of studies on artificial intelligence: CONSORT-AI and beyond. J Dent Res. 2021 Mar;3: 0022034521998337.
Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent. 2021 Apr 1; 107:103610.
Sultan AS, Elgharib MA, Tavares T, Jessri M, Basile JR. The use of artificial intelligence and deep machine learning in oncologic histopathology. J Oral Pathol Med. 2020 Oct;49(9): 849–56.
Tamaki Y, Nomura Y, Katsumura S, Okada A, Yamada H, Tsuge S, et al. Construction of a dental caries prediction model by data mining. J Oral Sci. 2009;51(1):61–8.
Wang C, Qin H, Lai G, Zheng G, Xiang H, Wang J, et al. Automated classification of dual channel dental imaging of auto-fluorescence and white lightby convolutional neural networks. J Innov Opt Health Sci. 2020 Jul 7;13(04): 2050014.
Welch ML, McIntosh C, Traverso A, Wee L, Purdie TG, Dekker A, et al. External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification. Phys Med Biol. 2020 Feb 5;65(3):035017.
Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529–36.
Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018 Jan 1;66(1):149–53.
Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019 Jan 1;170(1):51–8.
Zanella-Calzada LA, Galván-Tejada CE, Chávez-Lamas NM, Rivas-Gutierrez J, Magallanes-Quintanar R, Celaya-Padilla JM, et al. Deep artificial neural networks for the diagnostic of caries using socioeconomic and nutritional features as determinants: data from NHANES 2013–2014. Bioengineering. 2018 Jun;5(2): 47.
Zhang X, Liang Y, Li W, Liu C, Gu D, Sun W, et al. Development and evaluation of deep learning for screening dental caries from oral photos. Oral Dis. 2020 Dec;19.
Zhao M, Ji S, Wei Z. Risk prediction and risk factor analysis of urban logistics to public security based on PSO-GRNN algorithm. PLoS One. 2020 Oct 5;15(10):e0238443.