HHS Public Access

Author manuscript

Pediatr Dent. Author manuscript; available in PMC 2021 November 15.

Published in final edited form as:

Pediatr Dent. 2021 May 15; 43(3): 191–197.

An Automated Machine Learning Classifier for Early Childhood Caries

D.S. Karhade, DMD[1], **J. Roach, PhD**[2], **P. Shrestha, BDS, MSc**[3], **M.A. Simancas-Pallares, DDS, MSc**[4], **J. Ginnis, DDS**[5], **Z.J. Burk, BA**[6], **A.A. Ribeiro, DDS, MSD, PhD**[7], **H. Cho, MS**[8], **D. Wu, PhD**[9], **K. Divaris, DDS, PhD**[10]

[1]Dr. Karhade is pediatric dentistry resident, Division of Pediatric and Public Health, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Dr. Roach is senior scientific research associate, Research Computing, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3]Dr. Shrestha is pediatric dentistry resident, Division of Pediatric and Public Health, Adams School of Dentistry, and PhD candidate, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4]Dr. Simancas-Pallares is a pediatric dentistry resident, Division of Pediatric and Public Health, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[5]Dr. Ginnis is assistant professor, Division of Pediatric and Public Health, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[6]Mr. Burk is DDS candidate, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[7]Dr. Ribeiro is associate professor, Division of Diagnostic Sciences, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[8]Mr. Cho is PhD candidate, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[9]Dr. Wu is associate professor, Department of Biostatistics, Gillings School of Global Public Health, and Division of Oral and Craniofacial Health Research, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[10]Dr. Divaris is professor, Division of Pediatric and Public Health, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Abstract

Purpose: The purpose of the study was to develop and evaluate an automated machine learning algorithm (AutoML) for children's classification according to early childhood caries (ECC) status.

Methods: Clinical, demographic, behavioral, and parent-reported oral health status information for a sample of 6,404 three- to five-year-old children (mean age equals 54 months) participating in

Correspond with Dr. Karhade at deepti_karhade@unc.edu.

an epidemiologic study of early childhood oral health in North Carolina was used. ECC prevalence (decayed, missing, and filled primary teeth surfaces [dmfs] score greater than zero, using an International Caries Detection and Assessment System score greater than or equal to three caries lesion detection threshold) was 54 percent. Ten sets of ECC predictors were evaluated for ECC classification accuracy (i.e., area under the ROC curve [AUC], sensitivity [Se], and positive predictive value [PPV]) using an AutoML deployment on Google Cloud, followed by internal validation and external replication.

Results: A parsimonious model including two terms (i.e., children's age and parent-reported child oral health status: excellent/very good/good/fair/poor) had the highest AUC (0.74), Se (0.67), and PPV (0.64) scores and similar performance using an external National Health and Nutrition Examination Survey (NHANES) dataset (AUC equals 0.80, Se equals 0.73, PPV equals 0.49). Contrarily, a comprehensive model with 12 variables covering demographics (e.g., race/ethnicity, parental education), oral health behaviors, fluoride exposure, and dental home had worse performance (AUC equals 0.66, Se equals 0.54, PPV equals 0.61).

Conclusions: Parsimonious automated machine learning early childhood caries classifiers, including single-item self-reports, can be valuable for ECC screening. The classifier can accommodate biological information that can help improve its performance in the future.

Keywords

EARLY CHILDHOOD CARIES; MACHINE LEARNING; CLASSIFICATION; PREDICTION; RISK ASSESSMENT

Early childhood caries (**ECC**) is a common complex disease that persists as a clinical and public health problem, despite major advances in the science and practice of dentistry.[1] It is prevalent, affecting one in three children in the United States, and its effects often extend into adulthood.[2,3] ECC is caused by a dysbiotic shift in the supragingival biofilm due to an imbalance between protective and disease-promoting factors (e.g., consumption of fermentable carbohydrates). It has a complex etiology that can be considered at several levels, and includes molecular, microbiological, behavioral, social, environmental, health care system, and policy factors.[4] In terms of associated risk indicators, consumption of sugar-sweetened beverages and snacks, frequency of toothbrushing, parental education level, and optimal fluoridation have all been linked to ECC.[5–7] Knowledge of these factors that influence caries incidence on the population level is certainly desirable and useful. However, it is conceptually and practically different from determining individuals' likelihood of developing ECC (i.e., risk assessment) or presenting with ECC (i.e., screening), given their personal protective and disease-promoting factors.[4,8]

Caries risk assessment tools (**CAT**) are important aids for patient stratification, education, and optimization of care. These tools combine information from several protective and risk factors and generally assign subjective risk levels (e.g., low, moderate, high) to individual patients. Ideally, CAT should be evidence-based, valid, precise, replicated in more than one population, and contribute to improvements in oral health care delivery and, ultimately, oral health.[9] However, it is commonly understood that current risk assessment or disease prediction models do not perform well and are characterized by substantial variability in

their outputs.[9,10] There is a need for increased consistency in the development and evaluation of CAT, because, despite the presence of numerous caries risk prediction models, most are not based on high-quality evidence and are not routinely employed in practice.[11] Moreover, there is limited evidence regarding their validity and most utilize caries experience (i.e., a clinical manifestation of disease) as the predominant predictor of ECC incidence.[4] In other words, there is room for improvement in this important domain of caring for young children's oral health.

The use of machine learning (**ML**) for the development of screening or predictive models has been gaining popularity in medicine. The premise of ML is based on the ability of algorithms and classifiers to detect and leverage latent (i.e., "hidden") data structures and interrelationships that may not be evident in casual analyses or may be too complex to describe. Once an ML model is constructed, its utility can be determined using conventional metrics (i.e., sensitivity and positive predictive value) and using real data. For example, by constructing a decision tree, heart disease prediction models have been able to predict the heart disease status of patients by analyzing select clinical features.[12] Several other ML applications exist, including models that predict patients' quality of life from the text in physicians' notes or use home telemonitoring data to personalize predictions of asthma exacerbations.[13,14] An excellent overview and summary of ML applications in health care was recently published by Rajkomar et al.[15]

A wealth of information exists on proximal (e.g., clinical or biological) and distal (e.g., social and demographic) factors that are plausibly associated with dental caries risk (i.e., incident disease) or ECC status (i.e., prevalent disease).[16] ML can aid in the utilization of such information efficiently to inform clinical decision-making or public health applications. However, despite their promise, ML applications are scant in the oral health domain and have not been applied to the context of ECC and early childhood oral health, despite being considered likely elements of the future disruptive innovation in dentistry.[16,17]Automated machine learning (**AutoML**) is the process of automating the steps involved in "traditional" ML, including the model architecture search, feature engineering, and model deployment.

The purpose of the present study was to develop and evaluate an automated machine learning algorithm for children's classification according to early childhood caries status.

Methods

Study population and variables.

The present study used clinical, behavioral, demographic, and laboratory data (i.e., domestic home water fluoride concentration, as a measure of an environmental protective exposure) from a large sample of preschool-age children enrolled in a community-based epidemiologic study of early childhood oral health in North Carolina (ZOE 2.0 study).[18] A comprehensive description of the study overview, design, and cohort's demographic profile has been reported.[19] An overview of the specific data domains used for the present analysis in the context of the parent study is presented in Figure 1. Participating children were ages 36 to 71 months (mean equals 54) and attended public preschool centers (Head Start) in North Carolina. Written informed consent was obtained from the children's legal guardians to

perform clinical examinations for caries assessment, collection of biospecimens (saliva and plaque), collection of a home water sample, and completion of a questionnaire.[20]

Clinical examinations were performed by 10 trained and calibrated clinical examiners using artificial light and magnification. The clinical examination was done after toothbrush prophylaxis using portable dental equipment at the preschool centers that children attended. Questionnaires in English and Spanish were administered to guardians to collect proxy-reported oral health-related information, oral health behaviors, and demographic data. These included questions regarding toothbrushing patterns and frequency, daily frequency of sugary snack and beverage consumption, primary source of domestic water, dental home, history of the child being placed in bed with a bottle containing anything other than water, etc. Information collected from the guardians included their highest level of education completed and the perception of their children's health and their own oral health; these were collected using the five-level (i.e., poor, fair, good, very good, or excellent) National Health and Nutrition Examination Survey (**NHANES**) item. Additionally, fluoride concentration was measured from home water samples using the EPA 300.0 method in the North Carolina State Laboratory of Public Health and was categorized as optimal (greater than or equal to 0.60 ppm F) or suboptimal (less than 0.60 ppm F). A comprehensive description of the study's clinical protocol and data collection procedures has been previously reported.[20] The present study received ethics approval by the Institutional Review Board of the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA (#14–1992) and has been supported by grants from the National Institutes of Health (NIDCR U01DE025046 and R03DE028983).

Analytical strategy.

Parent-reported child oral health status and clinical, demographic, and behavioral data were available for 6,404 children ages 36 to 71 months. ECC cases were considered children with one or more decayed, missing, and filled primary surfaces (**dmfs**) greater than zero, wherein the "d" component was defined at the International Caries Detection and Assessment System threshold of greater than or equal to three (i.e., established caries lesions). Using this criterion, ECC prevalence was 54 percent in the study's analytical sample. To identify the best performing model, several sets of plausible ECC predictors were entered in the AutoML algorithm. Fourteen variables grouped in eight domains well-established for their association with ECC (i.e., children's age, parental education, diet/nutrition, dental home, oral hygiene, fluoride exposure, race/ethnicity, parent and child's reported oral health status), were considered iteratively until a maximum classification performance was reached using an AutoML tables deployment on the Google Cloud platform. The ECC classification accuracy of these models was assessed using conventional classification metrics (e.g., area under the ROC curve [**AUC**], sensitivity [**Se**], and positive predictive value [**PPV**]). Higher values for these metrics indicate better classification performance. The best performing combinations of predictors were carried forward to internal validation and external replication.

Internal validation.

The model underwent a customary internal validation step: 80 percent of the data were used to train the machine learning algorithm, and the remaining 20 percent were used for validation and testing. The best performing model's predicted probabilities were compared

to actual probabilities of children having ECC, across children's ages (i.e., three-, four-, and five-year-olds) and parents' reports of their children's oral health status (i.e., five levels, including poor, fair, good, very good, and excellent). Additionally, visual methods were used to illustrate the distribution of the automated ML classifier's predicted probabilities (i.e., risk scores) and their relationship with children's clinically determined caries experience, as measured by the dmfs index. Spearman's rho and corresponding 95 percent confidence intervals were obtained with bootstrapping to quantify this latter association.

External replication.

Nationally representative data from the latest four cycles of the NHANES, including the years 2011 through 2018 (N equals 2,311), were utilized to externally replicate the performance of the best-performing AutoML ECC classifier. NHANES is carried out by the Centers for Disease Control and Prevention, and clinical examination data are obtained from trained and calibrated dentists operating in mobile units. Children ages three to five years with clinical and questionnaire data (N equals 2,308), matching with the parent study's age distribution (i.e., range 37 to 71, mean equals 53 months) were included. Identical variable specifications (e.g., age in months and subjective measures of parent-reported child oral health status) were used in NHANES and the parent study. ECC prevalence in the NHANES dataset was 27 percent. Similar classification metrics (e.g., AUC, Se, PPV, etc.) were obtained for the best-performing model in ZOE 2.0 for the NHANES data set. All analyses were performed using Stata 16.1 software (StataCorp LP, College Station, Texas) and Google Cloud AutoML (Google, Mountain View, Calif., USA).

Results

Descriptive information of the analytical sample is presented in Table 1. Performance metrics for the AutoML models using 10 different sets of input variables are shown in Table 2, in order of increasing classification performance. Overall, increasing the number of input variables did not correlate with better ECC classification performance of the machine learning algorithm. A parsimonious model containing only two inputs, children's age and parent-reported child oral health status, had the highest AUC (0.74), Se (0.67), and PPV (0.64). The ECC probabilities predicted by this best-performing model were associated with monotonic increases in the actual clinically determined caries experience in this community-based cohort of preschool-age children, as illustrated in Figure 2. Most predicted probabilities were in the range of 0.4 to 0.7, which is expected due to the ECC prevalence of 54 percent in this sample, with few participants having very high (greater than 0.8) predicted probabilities and none having less than 0.2. The association between these predicted ECC probabilities and dmfs was of moderate magnitude and was statistically confirmed (rho equals 0.47; 95 percent confidence interval [**95% CI**] equals 0.45 to 0.49; $P<0.001$).

The best-performing model was then utilized to make predictions for specific strata of age and parentally-reported child oral health—these results are presented in Table 3. In general, the model-predicted and empirical distribution of ECC cases were concordant. For example, a five-year-old child whose parent reported "poor" oral health status had a model-predicted ECC probability of 91 percent versus 90 percent actual ECC probability. The largest

discordance between prediction and observation was found for the youngest children with excellent parent-reported oral health, where the predicted ECC probability (31 percent) was somewhat larger than the observed (24 percent). Finally, it was confirmed that the best-performing model (including terms for children's age in months and parent-reported child oral health status) performed equally well in the NHANES dataset (AUC equals 0.80, Se equals 0.73, PPV equals 0.49). The model-predicted ECC probability score, similar to the ZOE 2.0 study sample, was monotonically and robustly correlated with caries experience (rho equals 0.44; 95% CI equals 0.41 to 0.48; $P<0.001$) (Figure 3). The predicted probability distribution was different in the NHANES sample; the distribution was shifted to the left due to the lower predicted probabilities in NHANES compared to the ZOE 2.0 sample, owing to the lower baseline ECC prevalence (27 percent versus 54 percent in ZOE 2.0). A publicly available deployment of the model can be found at https://eccclassifier.childrensoralhealth.org/.

Discussion

This study developed, evaluated, and deployed an AutoML classifier for ECC. The study's results suggest that, in the absence of information from a clinical examination, ECC status can be reasonably inferred via a screening question obtaining a parent's perception of their child's oral health on a five-level scale combined with knowledge of the child's age. These findings are demonstrative of the potential of machine learning for augmenting and occasionally replacing traditionally collected data elements and making efficient use of other existing information. It must be stressed that the ML model is not ready for clinical deployment and cannot and will not replace a clinical examination; however, it can immediately augment administrative, hospital, and public health datasets that do not have clinical dental information with estimates of young children's ECC propensity. Machine learning ECC classification and risk models will be continuously refined and improved with the addition of new information and will be further replicated in diverse populations with different ECC prevalence, demographics, and risk profiles.

This novel report of an AutoML application to classify ECC is important for two main reasons. First, these results demonstrate that the approach is feasible and can accommodate large numbers of several types of input variables, including future additions of perhaps more informative, biological predictors (e.g., genomics, microbiome, metabolome, salivary properties), social and area-level influences (e.g., proxied by zip code), intraoral images, etc. Second, it enables the imputation of ECC status in large, administrative, and health care datasets where clinical examination information may be unavailable or unattainable but other potentially useful proxy information exists or can be easily obtained (e.g., via an online questionnaire). This would allow the examination of ECC associations with numerous other systemic health conditions using electronic health records and further enhance the ability to conduct large-scale surveillance.

The performance of the best-performing classification algorithm was modest; with a sensitivity of 0.67 one would expect to detect two-thirds of true cases, and with a specificity of 0.67 one would expect one-third of noncases to be misidentified as ECC cases. Nevertheless, these sensitivity estimates are on the upper end of the sensitivity range (0.41 to

0.75) that was recently reported in a systematic review of Cariogram applications for caries risk assessment in children.[21] Gao et al. reported high sensitivity (0.99) and low specificity (0.05) for the American Academy of Pediatric Dentistry (**AAPD**) tool, while caries management by risk assessment (CAMBRA) had a more balanced performance, with 0.94 sensitivity and 0.44 specificity.[22] In another study comparing four different ECC screening approaches, the investigators also reported a very high (100 percent) sensitivity for the AAPD tool; however, that was again combined with a three percent specificity, resulting in virtually equal proportions of true and false positives (positive predictive value less than 0.50; Yoon et al.).[23] Of note, consideration of a *Streptococcus mutans* salivary culture alone outperformed the AAPD caries risk assessment tool in the Yoon study, illustrating the utility of considering biological information in ECC screening. [23] This finding is consistent both with the current understanding of caries pathogenesis[1] and an earlier report by Saxena et al., who achieved a 92 percent accurate classification of *S mutans* species as severe-ECC or caries-free associated using a machine learning approach.[24]

The wide variation in tools and approaches available for ECC screening or risk assessment is well-documented.[10] Moreover, the accuracy and generalizability of most existing models are limited because they have not been validated in independent populations and prospective cohort studies.[9] In general, it appears that most current models tend to overestimate caries risk or prevalence (i.e., very high sensitivities and very low specificities) and result in substantial proportions of false positives. Despite the obvious drawbacks of model imprecision, identification of those at highest risk is vital from a clinical standpoint. For example, 12-year-old children who were classified as very high risk using the full Cariogram model were found to develop approximately 30 times more caries lesions two years later versus those who were classified as very low risk. [25]

The results of this investigation, although promising, should be considered while acknowledging its limitations. First and foremost, the developed machine learning model was tested and optimized for inference as a screening tool (i.e., the probability of a child having ECC) and not caries risk (i.e., the likelihood of disease incidence). In other words, the present study should be interpreted as a demonstration of machine learning for the development of a screening tool for prevalent ECC.

Second, the data used for its development are from one cohort of high-risk children in low-income families in one state, which is not representative of all three- to five-year-old children in North Carolina or the United States. Despite this, the authors verified that the obtained model performed at least equally well when applied to a large nationally representative sample of similarly aged preschool children.

Third, the algorithm's reliance on parents' perceptions of their children's oral health may initially appear as unwarranted. Although perhaps of limited value on an individual case basis, previous reports have shown that parents' perceptions of preschool-age children's oral health can be used as reliable proxy information.[26–28] This association may be due to parent's direct observation or awareness of oral or dental conditions, associated symptoms, information from health care providers, or other mechanisms. More importantly, it is well documented that measures of self-rated health are informative and highly predictive of

health outcomes, including mortality in population-based studies.[29] Factors such as culture and ethnicity, spoken language, and health literacy are undoubtedly influential for parents' perceptions and reports of their children's oral health. Despite this, here the authors demonstrate that parental perceptions remain the best possible predictors, to date, among a large set of demographic, behavioral, and environmental variables. On the other hand, age is a *de facto* considered factor in ECC risk or classification models. This is because time is integral to the concept of risk, and ECC experience is irreversible, thus monotonically increasing by age.

Despite its limitations, the authors support that the application of machine learning for ECC classification is of value. Not only did the ML algorithm perform reasonably well within this study (including in training and validation subsamples), but it also performed well in an external, nationally representative dataset of similarly aged children. Upon development and replication, the algorithm was used to compute actual ECC probabilities for subgroups of children that were found to be accurate when compared to actual, observed probabilities. Furthermore, the algorithm was deployed in a publicly accessible domain where it can be accessed and further evaluated by clinicians, researchers, and the public. Importantly, the algorithm will be continuously refined and updated to accommodate additional biological information (e.g., microbiome, metabolome, human genomics, etc.) that is known to increase performance. ML classifiers, among other uses, can serve as valuable aids for health care professionals to identify children who may have ECC and direct them to a dental care provider. Furthermore, deployed classifiers can be useful for the imputation of ECC status in administrative or other health care datasets without clinical examination information and enable analyses that are otherwise impossible to accomplish.

Conclusions

Based on the results of this study, the following conclusions can be made:

1. A parsimonious model containing only two inputs had the highest early childhood caries classification performance. The inclusion of a screening question regarding children's oral health was more informative and led to a better performing model compared to data on several traditional caries risk factors.
2. A relatively naïve machine learning model based on children's age and parental perception of their oral health can be used to estimate the probability of being an ECC case, independent of a clinical encounter, with reasonable accuracy.
3. Machine learning can be utilized to make high-quality classifiers capable of imputing ECC status based on proxy-reported and demographic data. This capacity can augment oral health information of large administrative or public health datasets and will likely be improved upon the inclusion of more biological (e.g., microbiome, metabolome, human genome) information.

Acknowledgements

The study was supported by grants from the National Institutes of Health/National Institute of Dental and Craniofacial Research (NIH/NIDCR) U01DE025046 and R03DE028983.

References

1. Selwitz RH, Ismail AI, Pitts NB. Dental caries. Lancet 2007;369(9555):51–9. [PubMed: 17208642]
2. Broadbent J, Thomson W, Poulton R. Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life. J Dent Res 2008;87(1):69–72. [PubMed: 18096897]
3. Casamassimo PS, Thikkurissy S, Edelstein BL, Maiorini E. Beyond the dmft: the human and economic cost of early childhood caries. J Am Dent Assoc 2009;140(6):650–7. [PubMed: 19491160]
4. Divaris K Predicting dental caries outcomes in children: a “risky” concept. J Dent Res 2016;95(3):248–54. [PubMed: 26647391]
5. Peltzer K, Mongkolchati A, Satchaiyan G, Rajchagool S, Pimpak T. Sociobehavioral factors associated with caries increment: a longitudinal study from 24 to 36 months old children in Thailand. Int J Environ Res Public Health 2014;11(10):10838–50. [PubMed: 25329535]
6. Park S, Lin M, Onufrak S, Li R. Association of sugar-sweetened beverage intake during infancy with dental caries in 6-year-olds. Clin Nutr Res 2015;4(1):9–17. [PubMed: 25713788]
7. Chaffee BW, Feldens CA, Rodrigues PH, Vítolo MR. Feeding practices in infancy associated with caries incidence in early childhood. Community Dent Oral Epidemiol 2015;43(4):338–48. [PubMed: 25753518]
8. Rose G Sick individuals and sick populations. Int J Epidemiol 1985;14(1):32–8. [PubMed: 3872850]
9. Mejàre I, Axelsson S, Dahlën Ga, et al. Caries risk assessment: a systematic review. Acta Odontol Scand 2014;72(2):81–91. [PubMed: 23998481]
10. Halasa-Rappel YA, Ng MW, Gaumer G, Banks DA. How useful are current caries risk assessment tools in informing the oral health care decision-making process? J Am Dent Assoc 2019;150(2):91–102. [PubMed: 30691581]
11. Fontana M, Carrasco-Labra A, Spallek H, Eckert G, Katz B. Improving caries risk prediction modeling: a call for action. J Dent Res 2020;99(11):1215–20. [PubMed: 32600174]
12. Pandey AK, Pandey P, Jaiswal K, Sen AK. A heart disease prediction model using decision tree. IOSR-JCE 2013;12(6):83–6.
13. Pakhomov S, Shah N, Hanson P, Balasubramaniam S, Smith SA. Automatic quality of life prediction using electronic medical records. AMIA Annu Symp Proc 2008;2008:545–9.
14. Finkelstein J, Jeong IC. Machine learning approaches to personalize early prediction of asthma exacerbations. Ann N Y Acad Sci 2017;1387(1):153–65. [PubMed: 27627195]
15. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347–58. [PubMed: 30943338]
16. Fisher-Owens SA, Gansky SA, Platt LJ, et al. Influences on children’s oral health: a conceptual model. Pediatrics 2007;120(3):e510–e520. [PubMed: 17766495]
17. Joda T, Yeung A, Hung K, Zitzmann N, Bornstein M. Disruptive innovation in dentistry: what it is and what could be next. J Dent Res 2021;100(5):448–53. [PubMed: 33322997]
18. Divaris K, Joshi A. The building blocks of precision oral health in early childhood: the ZOE 2.0 study. J Public Health Dent 2020;80 (suppl 1):S31–S36. [PubMed: 30566750]
19. Divaris K, Slade GD, Ferreira Zandona AG, et al. Cohort profile: ZOE 2.0: a community-based genetic epidemiologic study of early childhood oral health. Int J Environ Res Public Health 2020;17:8056.
20. Ginnis J, Ferreira Zandona AG, Slade GD, et al. Measurement of early childhood oral health for research purposes: dental caries experience and developmental defects of the enamel in the primary dentition. Methods Mol Biol 2019;1922:511–23. [PubMed: 30838597]
21. Cagetti MG, Bonta G, Cocco F, Lingstrom P, Strohmenger L, Campus G. Are standardized caries risk assessment models effective in assessing actual caries status and future caries increment? A systematic review. BMC Oral Health 2018;18(1):123. [PubMed: 30012136]
22. Gao X, Di Wu I, Lo EC, Chu CH, Hsu CY, Wong MC. Validity of caries risk assessment programmes in preschool children. J Dent 2013;41(9):787–95. [PubMed: 23791698]

23. Yoon RK, Smaldone AM, Edelstein BL. Early childhood caries screening tools: a comparison of four approaches. J Am Dent Assoc 2012;143(7):756–63. [PubMed: 22751977]

24. Saxena D, Caufield PW, Li Y, Brown S, Song J, Norman R. Genetic classification of severe early childhood caries by use of subtracted DNA fragments from *Streptococcus mutans*. J Clin Microbiol 2008;46(9):2868–73. [PubMed: 18596144]

25. Hebbal M, Ankola A, Metgud S. Caries risk profile of 12-year-old school children in an Indian city using Cariogram. Med Oral Patol Oral Cir Bucal 2012;17(6):e1054–e1061. [PubMed: 22926464]

26. Filstrup SL, Briskie D, Da Fonseca M, Lawrence L, Wandera A, Inglehart MR. Early childhood caries and quality of life: child and parent perspectives. Pediatr Dent 2003;25(5):431–40. [PubMed: 14649606]

27. Divaris K, Vann WF Jr, Baker AD, Lee JY. Examining the accuracy of caregivers' assessments of young children's oral health status. J Am Dent Assoc 2012;143(11):1237–47. [PubMed: 23115154]

28. Talekar BS, Rozier RG, Slade GD, Ennett ST. Parental perceptions of their preschool-aged children's oral health. J Am Dent Assoc 2005;136(3):364–72. [PubMed: 15819352]

29. Idler EL, Benyamini Y. Self-rated health and mortality: a review of 27 community studies. J Health Soc Behav 1997;38(1):21–37. [PubMed: 9097506]

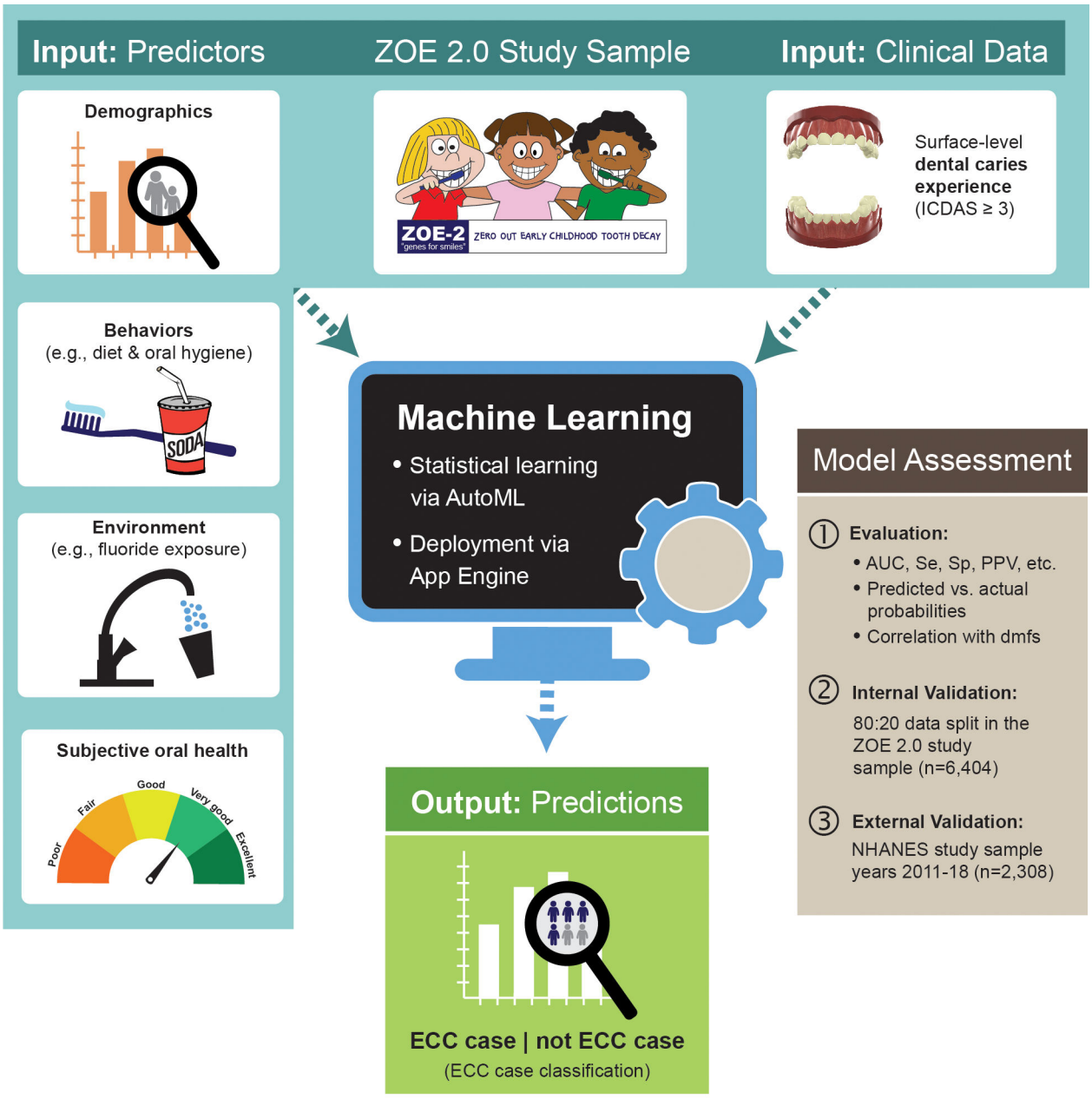

Figure 1.

Overview of study data, model development, and assessment procedures. First, clinical ECC data and information on postulated ECC predictors from 6,404 three- to five-year-old participants of the ZOE 2.0 study were used to develop an ECC classification model. An automated machine learning (AutoML) deployment on Google Cloud was used to identify the best-performing model, as determined by conventional classification metrics such as the area under the receiver operator curve (AUC), sensitivity (Se), specificity (Sp), and positive predictive value (PPV). The model was internally validated using a customary 80:20 data split, wherein the model is developed in a random 80 percent of the sample and then evaluated in the remaining 20 percent. External replication of the best-performing model

developed in the ZOE 2.0 study was done using nationally representative data from the National Health and Nutrition Examination Survey (NHANES) study years 2011 to 2018 (N equals 2,308, similarly aged children).

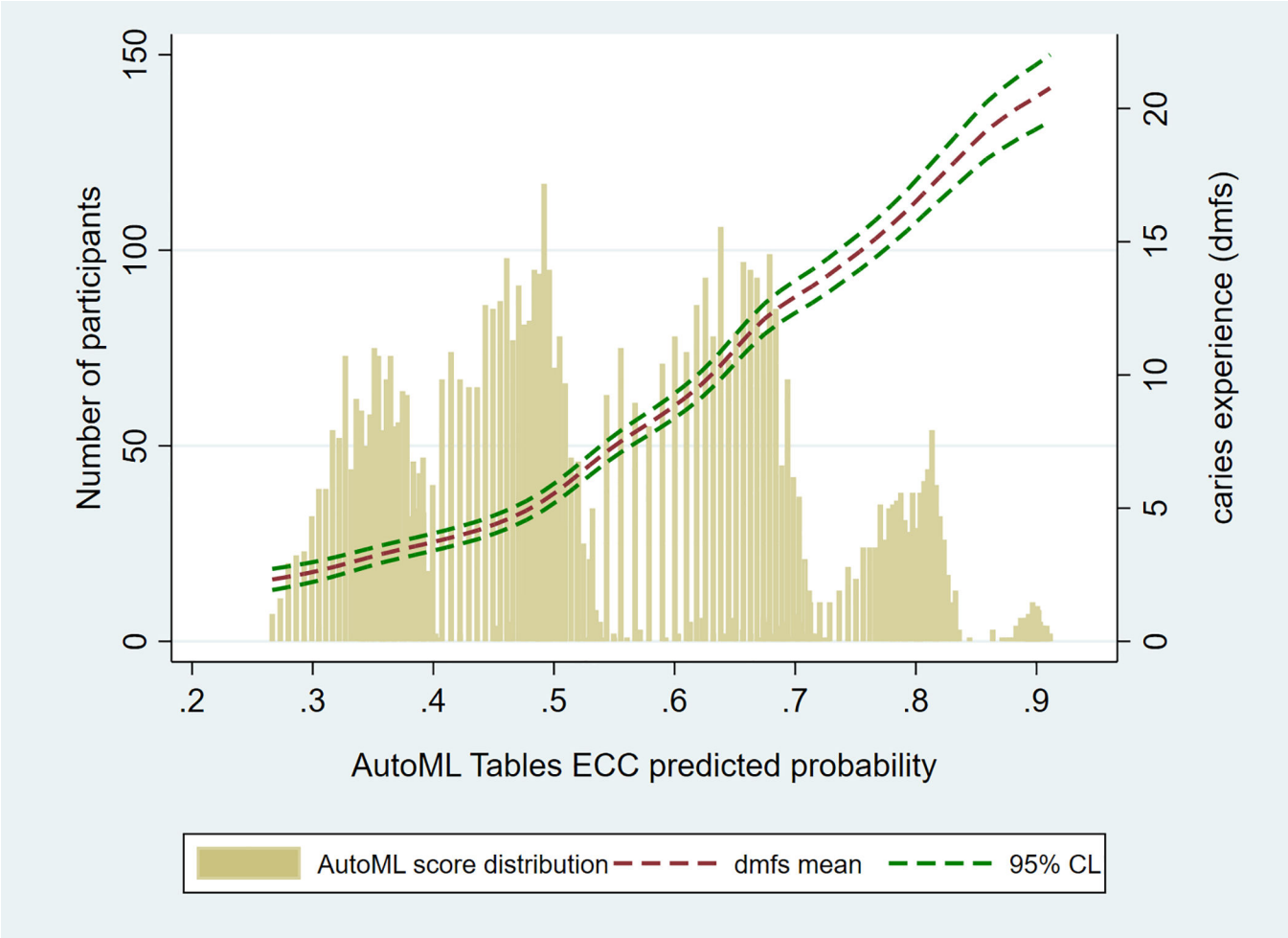

Figure 2.

Joint distribution of the best-performing automated machine learning (AutoML) model predicted early childhood caries (ECC) probabilities (horizontal axis) and caries experience (as quantified by the decayed, missing, and filled primary tooth surface (dmfs) index and 95% confidence limit (CL), right vertical axis) in the ZOE 2.0 study sample (N equals 6,404).

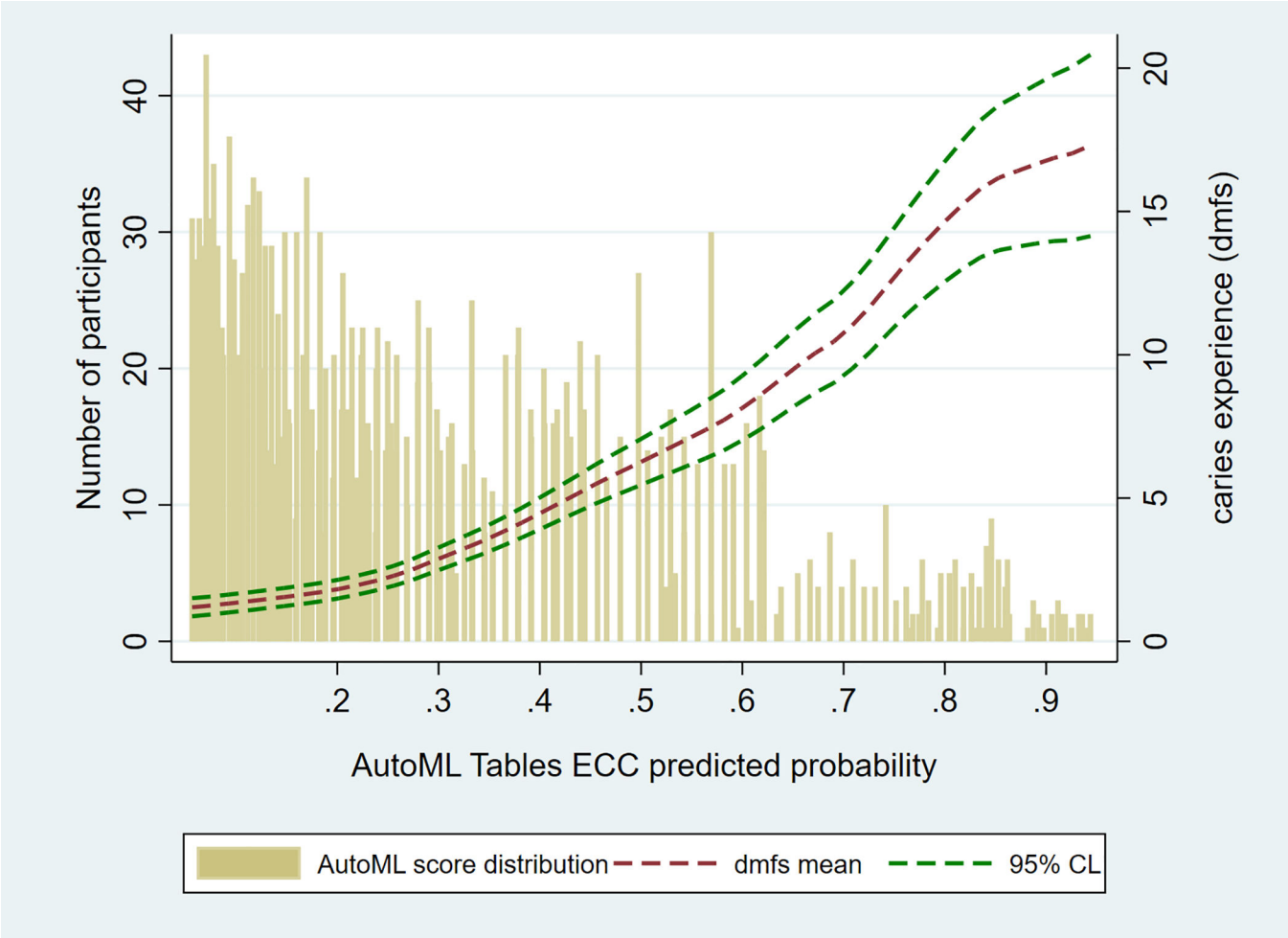

Figure 3.

Joint distribution of the best performing automated machine learning (AutoML) model predicted early childhood caries (ECC) probabilities (horizontal axis) and caries experience (as quantified by the decayed, missing, and filled primary tooth surface (dmfs) index and 95% confidence limits (CL), right vertical axis) in the National Health and Nutrition Examination Survey years 2011 to 2018 (N equals 2,308).

Table 1.

PARTICIPANTS' EARLY CHILDHOOD CARIES (ECC) DIAGNOSIS AND DEMOGRAPHIC CHARACTERISTICS IN THE ZOE 2.0 STUDY

		ECC*	
	N (column %)	**Cases N (row %)**	**dmfs†, mean (median) among cases**
Entire sample	6,404 (100)	3,465 (54)	15 (8)
Sex			
Boy	3,188 (50)	1,747 (55)	15 (8)
Girl	3,216 (50)	1,718 (53)	14 (8)
Age (years)			
3	1,502 (23)	673 (45)	12 (6)
4	3,341 (52)	1,833 (55)	15 (8)
5	1,561 (24)	959 (61)	18 (10)
Months, mean (±SD)	54 (7)	54 (7)	
Race			
African American	3,094 (48)	1,622 (52)	14 (7)
American Indian or Alaskan Native	186 (3)	127 (68)	22 (14)
Asian	32 (1)	24 (75)	21 (17)
Native Hawaiian, Pacific Islander	4 (0.1)	4 (100)	21 (21)
White	1,385 (22)	725 (52)	16 (8)
>1 race	835 (13)	426 (51)	13 (7)
Other	864 (14)	534 (62)	18 (11)
Hispanic ethnicity			
Yes	1,291 (20)	788 (61)	17 (11)
No	5,042 (80)	2,643 (52)	14 (8)

*ECC was defined as one or more decayed, restored, or extracted primary tooth surfaces due to caries (i.e., decayed, missing, and filled primary tooth surface [dmfs] score more than zero), where caries lesions were considered at the International Caries Detection and Assessment System (ICDAS) threshold of greater than or equal to three.

†dmfs is the sum of decayed, missing, or filled (i.e., restored) primary tooth surfaces due to dental caries.

Table 2.

OVERVIEW OF EARLY CHILDHOOD CARIES (ECC) PREDICTORS INCLUDED IN 10 DIFFERENT AUTOMATED MACHINE LEARNING (AUTOML) MODELS AND CORRESPONDING MODEL CLASSIFICATION PERFORMANCE METRICS IN THE ZOE 2.0 STUDY SAMPLE (n EQUALS 6,404)*

Input variables										Model classification performance				
Child age†	Parental education‡	Nutrition		Dental home	Oral hygiene	Fluoride exposure	Race/ ethnicity	Reported oral health status§						
		ssbs	ntbottle		tbfr; brhelp; fltpaste	wfl, wsource	race, ethnicity	parent's	child's	AUC	Accuracy	Sp	PPV	Se
X		X		X	X					0.60	0.58	0.72	0.58	0.44
X	X									0.60	0.58	0.77	0.59	0.35
X								X		0.62	0.58	0.80	0.59	0.33
X		X	X	X	X					0.62	0.60	0.77	0.61	0.41
X		X	X	X	X	X				0.64	0.61	0.78	0.63	0.42
X	X	X	X	X	X	X				0.66	0.61	0.71	0.61	0.50
X	X	X	X	X	X	X	X			0.66	0.62	0.69	0.61	0.54
X	X	X	X	X	X	X	X	X		0.67	0.62	0.46	0.63	0.49
X	X	X	X	X	X	X	X	X	X	0.71	0.65	0.57	0.61	0.73
X									X	0.74	0.67	0.66	0.64	0.67

*The models are ordered according to increasing performance as indicated by the area under the receiver operator curve (AUC) statistic. The final model, containing terms for children's age and parent-reported child oral health status, demonstrated the highest classification performance and was subsequently replicated in the National Health and Nutrition Examination Survey (NHANES) sample (n equals 2,308); ssbs=frequency of daily consumption of sugar-containing snacks and beverages; ntbottle=history of the child ever been put to bed with a bottle containing anything but water; tbfr=daily frequency of toothbrushing; brhelp=adult involvement in child's toothbrushing; fltpaste=fluoride toothpaste is used every time child's teeth are brushed; wfl=child lives in an area with optimally fluoridated community water; wsource=source of water consumed by the child at home (i.e., bottled, city, well, or combinations); AUC=area under the receiver operator curve; Sp=specificity; PPV=positive predictive value; Se=sensitivity

†Measured in months

‡Highest level of education completed

§Measured using the NHANES item, as excellent, very good, good, fair, or poor

Table 3.

AUTOMATED MACHINE LEARNING (AUTOML) MODEL-PREDICTED VERSUS ACTUAL EARLY CHILDHOOD CARIES (ECC) PROBABILITIES ACCORDING TO PARENTS' REPORTS AND CHILDREN'S AGES

Age group		Parent-reported child oral health status				
	ECC* probability	Poor	Fair	Good	Very good	Excellent
3-year-old	Predicted	0.88	0.74	0.54	0.40	0.31
	Actual	1.00	0.83	0.55	0.36	0.24
4-year-old	Predicted	0.90	0.79	0.63	0.47	0.35
	Actual	0.96	0.86	0.66	0.46	0.29
5-year-old	Predicted	0.91	0.82	0.69	0.51	0.38
	Actual	0.90	0.87	0.73	0.50	0.36

*ECC was defined as one or more decayed, restored, or extracted primary tooth surfaces due to caries (i.e., dmfs greater than zero) where caries lesions were considered at the International Caries Detection and Assessment System (ICDAS) threshold of greater than or equal to three.