Journal of Dental Research
1–10


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00220345251398027
journals.sagepub.com/home/jdr

Uncovering Dental Caries Heterogeneity in NHANES Using Machine Learning

A. Orlenko[1], J.D. Mure[2], J.I. Gluch[3], J. Gregg[4], C.W. Compher[5,6], Z. Ren[7], H. Koo[7*], and J.H. Moore[1*]

Abstract

National Health and Nutrition Examination Survey (NHANES), one of the largest curated repositories of population-level health indicators including physical examinations, blood/urine biochemistry, self-reported surveys, and dietary intake, offers rich resources for oral health research but presents challenges for machine learning analysis due to heterogeneity, missing data, and complexity. Dental caries, the most prevalent chronic disease worldwide, is a multifactorial disease and exhibits variability in clinical manifestation, calling for advanced analytical approaches for deeper understanding. Here, we develop an integrated data-cleaning and subtype discovery pipeline using unsupervised machine learning for comprehensive analysis and visualization of data patterns in the NHANES database. Our multidimensional pipeline declutters and optimizes the NHANES dataset by addressing missingness and outliers to streamline data integration and create a machine learning–ready version. Applying this pipeline reveals data patterns that led to the discovery of previously unrecognized subtypes and variables associated with the clinical heterogeneity of dental caries. We observed diverging patterns of similarity across age groups and variable subsets, identifying distinct clusters particularly in children (<5 y) and senior adults (>65 y). We also discovered unexpected associations involving lead exposure and specific laboratory markers and, importantly, identified novel dietary signatures by linking food type and co-occurring consumption patterns to caries. Altogether, we report a comprehensive data-processing and data-analysis approach that reveals significant dental caries heterogeneity in NHANES data and can support the development of more precise and robust machine learning models for dental caries and other health conditions.

Keywords: epidemiology, machine learning, data mining, risk factors, oral health, cluster analysis

Introduction

The National Health and Nutrition Examination Survey (NHANES) is one of the largest curated repositories of nationally representative population-level health-related indicators, including physical examination, blood biochemistry, self-reported surveys, and dietary intake data. Overseen by the Centers for Disease Control, NHANES data are acquired through personal interviews combined with routine clinical assessments (dental and medical examinations), dietary surveys, and laboratory tests and have been used continuously to monitor US oral and overall health since 1999. This resource provides comprehensive data for studying oral diseases such as dental caries.

Despite its breadth, NHANES presents challenges for machine learning due to data heterogeneity, varied sample sizes, missing values, and inconsistencies in data collection methods (Pfeiffer et al 2017; Dye et al 2019), all of which can affect predictive modeling. These limitations are further compounded by outliers and mixed data types, requiring dedicated pipelines for data correction (Willemink et al 2020). Moreover, complex health outcomes such as dental caries often exhibit high clinical heterogeneity, highlighting the need for more refined phenotypic subtyping as well as effective visualization and interpretation methods for high-dimensional data patterns.

In this study, we addressed these limitations by establishing a data-cleaning pipeline with a novel outlier detection algorithm and unsupervised machine learning to identify phenotype subtypes within NHANES dental caries data. The NHANES includes highly detailed oral health information, and when combined with age, environmental, laboratory, and dietary data, this dataset can be leveraged for new insights into caries

[1]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA, USA
[2]School of Dental Medicine, University of Pennsylvania, Philadelphia, PA USA
[3]Department of Preventive and Restorative Sciences, University of Pennsylvania School of Dental Medicine, Philadelphia, PA, USA
[4]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA
[5]Biobehavioral Health Science Department, School of Nursing, University of Pennsylvania, Philadelphia, PA, USA
[6]Clinical Nutrition Support Services, Hospital of the University of Pennsylvania, Philadelphia, PA, USA
[7]Center for Innovation & Precision Dentistry, School of Dental Medicine and School of Engineering & Applied Sciences, University of Pennsylvania, Philadelphia, PA, USA
*Authors sharing senior authorship.

A supplemental appendix to this article is available online.

Corresponding Author:
H. Koo, Center for Innovation & Precision Dentistry, School of Dental Medicine and School of Engineering & Applied Sciences, University of Pennsylvania, 240 South 40th Street, Levy Bldg., Philadelphia, PA 19104-6030, USA.
Email: koohy@upenn.edu

determinants. A nutrition taxonomy based on the US Department of Agriculture's dietary guidelines (US Department of Agriculture and US Department of Health and Human Services 2020) further enhances precision in identifying associations between diet and disease.

We focus on dental caries, the most prevalent oral disease affecting more than 3.5 billion people globally (Marcenes et al 2013; Richards 2013). We developed an integrated data-cleaning–subtype discovery pipeline using unsupervised machine learning with a graph-based algorithm for visualization of data-clustering patterns and outcome analysis. The pipeline declutters, optimizes, and streamlines the integration of NHANES data, producing a "machine learning–ready" dataset, while revealing previously unrecognized subtypes and variable associations as well as diverging similarity patterns underlying the clinical heterogeneity of dental caries.

Methods

Data Acquisition

We analyzed publicly available data for 8,099 samples from the 2017–2018 NHANES survey (Centers for Disease Control and Prevention 2018), which includes demographic, dietary, examination, laboratory, and questionnaire data. Full details and ethical considerations are in the Appendix.

Data Preprocessing

NHANES datasets were aggregated and processed by removing variables with >50% missing data (Appendix Fig 1), selecting domain-relevant features, and applying one-hot encoding to nonordinal categorical variables. This resulted in a final dataset of 438 variables (Appendix Table 1). The full workflow is detailed in the Appendix.

Outlier Detection and Data Imputation

We applied the Skew and Tail-heaviness Adjusted Removal of Outliers (STAR) (Verardi and Vermandele 2018; Gregg and Moore 2023) for outlier removal, followed by multivariate imputation of missing values using *Scikit-learn*'s iterative imputer. Detailed rationale and parameters are provided in the Appendix.

Clinical Outcome Definitions

We derived binary (any caries) and categorical (mild vs severe) outcomes from oral examination data, considering both active-only and combined active/past disease statuses. The threshold for severe caries (≥4 affected surfaces/teeth) was based on the established definition of severe early childhood caries (S-ECC) (Tinanoff et al 2019). Full details and rationale are provided in the Appendix.

Unsupervised Learning Analysis

We used bootstrap-resampled hierarchical clustering (*pvclust* R package; Suzuki and Shimodaira 2006) to identify significant variable clusters within subgroups stratified by age, children (≤5 y), youth (>5 to ≤18 y), adult (>18 to <65 y), and senior (≥65 y), and by caries status (severity, activity). Analysis was also performed on nutrition-only subsets stratified by sugar content. Clustering parameters were chosen based on a quantitative validation (Appendix Table 3), with full details in the Appendix.

Data Visualization

Key findings were visualized using cluster maps. Spearman correlation networks were generated for age- and lead-stratified subgroups and *TooManyCells* spectral clustering (Schwartz et al 2020) for sample relationships. The rationale for the lead categorization and other specific visualization techniques is detailed in the Appendix.

Results

NHANES Data-Cleaning Pipeline

We developed an integrated data-cleaning–subtype discovery pipeline for observational studies (Fig 1). Specifically, NHANES 2017–2018 datasets for caries, including demographic, disease severity/activity, laboratory, examination, dietary (total nutrient intake and dietary supplement), and questionnaire indicators, were assembled into a single dataset that, upon removal of variables with a missing data fraction greater than 50%, domain-relevant variable selection, redundancy correction, and nonordinal categorical variable encodings, contained 438 variables.

First, we quantified the proportion of missing values. Variables with a missingness fraction <50% were retained, which reduced the dataset size from 1,486 to below 500 variables (Appendix Fig 1). Next, oral health–relevant variable selection was performed, in which irrelevant or redundant variables were eliminated. Finally, nominal variables were processed with the *OneHotEncoder* operator, which transforms each variable with *N* categories into *N* binary variables. These preprocessing steps resulted in the dataset of 438 variables.

Accounting for outliers is critical for robust and generalizable analysis. This is especially important in observational studies containing self-reported evaluations (eg, nutrition diary), a common subject to measurement error. Therefore, following the multistep variable selection and transformation, the *STAR* outlier detection algorithm was applied, a multivariate projection-based method that does not rely on the normal distribution assumption. Appendix Figure 2 demonstrates examples of the *STAR* algorithm outlier detection for NHANES variables with skewed distributions and different tail heaviness. The final step of the cleaning pipeline was multivariate

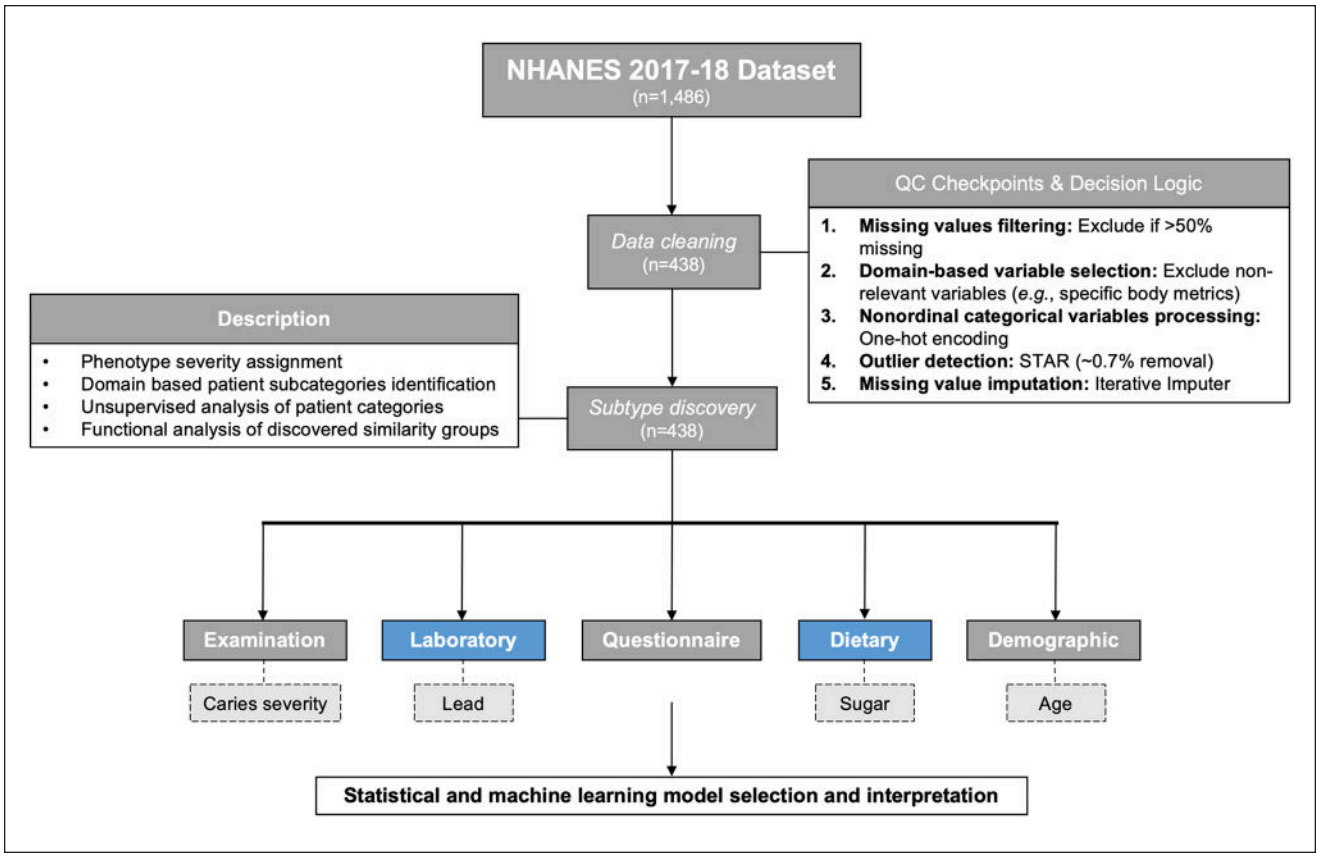


Figure 1. An overview of the data-cleaning and exploratory analysis pipeline components.

imputation with the iterative imputer. This method performs sequential imputation of every variable with missing values by fitting a machine learning model on all the remaining variables in the dataset and improves the missing value estimate by repeating this process multiple times.

Caries Subtype Discovery

Dental caries is a complex trait, and analyzing the entire disease population with a single model can oversimplify the true underlying mechanisms. Therefore, once the dataset was finalized, we investigated a series of clinical indicators for their role as drivers of caries subpopulation heterogeneity. We first assumed age, disease severity, and disease activity status to be the key drivers of clinical divergence and analyzed these subpopulations with unsupervised learning methods. The *pvclust* hierarchical clustering method assigns bootstrap resampling–derived *P* values to cluster dendrograms and outlines clusters supported by the data. The presence of large clusters of variables is of particular interest in the disease subtype heterogeneity analysis. Functional analysis of such patterns can provide an explanation of the driving processes within the subtype while serving as a variable selection foundation for datasets with high dimensionality. Diverging patterns between disease subtypes also affect the generalizability of machine learning model inference.

For caries age-stratified analysis performed with *pvclust* (Fig 2A–D), the caries subpopulation of children (≤5 y) showed the largest supported cluster among all age groups, consisting of 151 variables (Fig 2A, E). In addition, caries among the senior (≥65 y) subpopulation contained the second largest cluster discovered, consisting of 146 variables (Fig 2D and H). The remaining age groups, youth (>5 to ≤18 y) and adult (>18 to <65 y) caries subpopulations (Fig 2B, C), appeared to have a more homogenous distribution of supported clusters, with the largest clusters sizes less than 20 and 60 variables, respectively (Fig 2F, G). Interestingly, the children (CH) and senior (S) caries populations' largest cluster content is predominantly non-overlapping—only 46 common variables and 100 and 105 unique variables (Fig 3A). Functionally, the CH cluster content was 48% laboratory result variables, while the overall fraction of lab variables was only 28% (Fig 3C, D). The S cluster–affiliated variable categories distribution, however, is more similar to the overall categories distribution (Fig 3C and E). A color map of the scaled and aligned set of variables from the CH cluster demonstrates the discrepancy in variable patterns between the children and the senior age groups (Appendix Fig 3). This is also confirmed by the correlation network analysis in which strong linear correlation associations (Spearman's $|r|>0.6$) for children- and senior-age caries subpopulations vary dramatically (Appendix Fig 4 and Appendix Table 5).

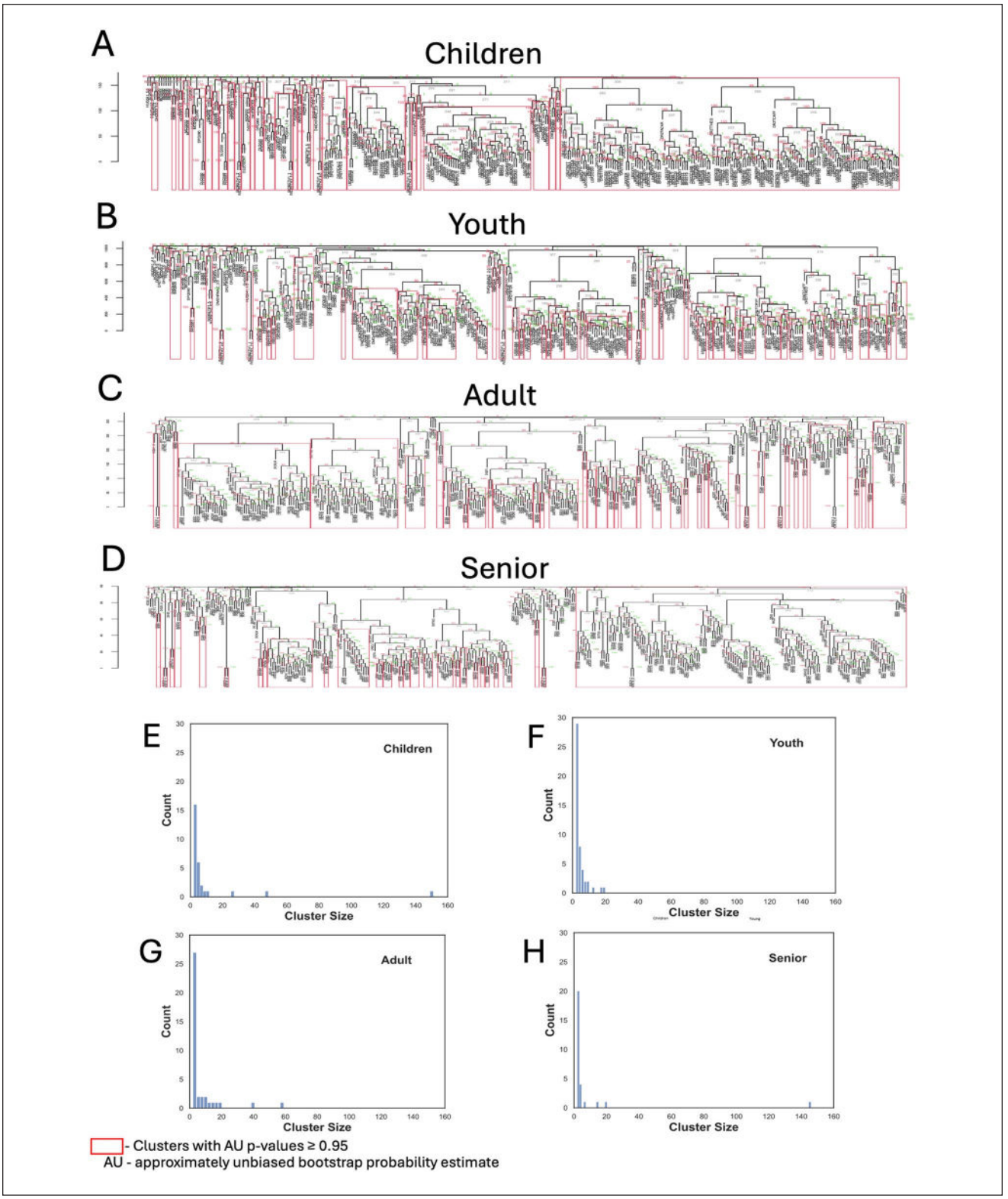


Figure 2. Hierarchical relationships between National Health and Nutrition Examination Survey (NHANES) variables for age-stratified subpopulations with caries identified by the *pvclust* clustering method. Clusters supported by the significant approximately unbiased (AU) bootstrap probability estimate are shown for (**A**) children (≤5 y), (**B**) youth (>5 to ≤18 y), (**C**) adult (>18 to <65 y), and (**D**) senior (≥65 y) age groups. Large AU significant clusters (>100 variables) are indicated for children and senior age groups, referred to as CH and S, respectively. (**E–H**) Cluster size (number of variables in AU significant cluster) distributions are shown for each age subpopulation. The statistical validity of these clustering structures was confirmed using the Calinski–Harabasz index, with detailed metrics provided in Appendix Table 4.

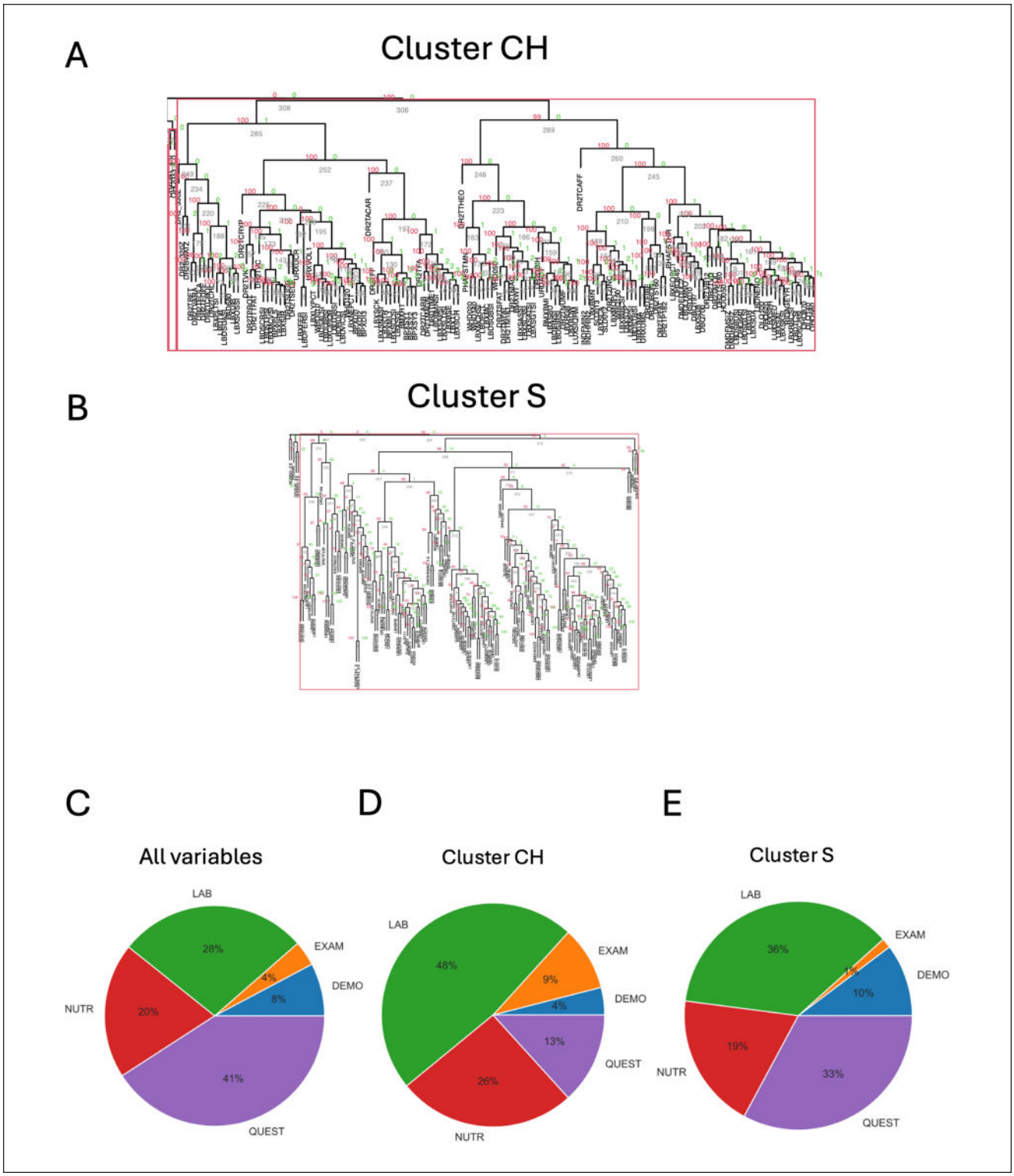


Figure 3. Functional analysis of the largest (>100 variables) approximately unbiased (AU) significant clusters in the *pvclust* analysis. Overlapping and unique features for the largest clusters for (**A**) children (≤5 y) (CH) and (**B**) senior (≥65 y). Functional variable categories distribution for (**C**) all variables, (**D**) cluster CH only, and (**E**) cluster S only. Lab, laboratory variables category; EXAM, physical examination variables category; DEMO, demographic variables category; QUEST, questionnaire variables category; NUTR, nutrition variables category.

We investigated subpopulations based on disease severity for active and historical caries. Among these, only active caries analysis revealed the most diverging cluster patterns between severity subpopulations. Specifically, a single large supported cluster (79 variables, referred to as S1) was associated with the severe caries phenotype that represents patients with more than 4 teeth affected (Appendix Fig 5). Its functional content is almost equally divided between laboratory and nutrition variable categories (Appendix Fig 6). In addition, 4 more clusters with a size greater than 20 variables were associated with an active severe caries subpopulation. Conversely, mild and no caries subpopulations appeared to have dense distributions of small supported clusters, most with fewer than 5 variables. In summary, bootstrap resampling–powered hierarchical clustering analysis identified 2 major clusters in age-stratified caries data (children and senior) with largely unique variables that may indicate distinctive phenotypes and require separate disease models, whereas a single major cluster found in the severe active caries subpopulation may reveal novel variables associated with the disease that will not be visible in the historical caries population or in the general active caries disease population.

Spectral Hierarchical Clustering Analysis

We then performed spectral hierarchical clustering analysis with *TooManyCells*, a method used to visualize single-cell clades relationships, to discover if any of the disease subtype annotations we had investigated could confirm the heterogeneity of the caries population. *TooManyCells* analysis, performed on the entire variable set, resulted into a dendrogram with 69 terminal clusters (Fig 4A). Considering the prior discovery of patterns associated with lab and nutrition variables categories, we suggested that using a complete variable set could be dimensionally challenging to infer smaller subtype patterns, and therefore, we analyzed laboratory-only and nutrition-only variable subsets, which resulted in spectral dendrograms with 110 and 74 terminal clusters correspondingly (Fig 4B and Appendix Fig 7). We then estimated the heterogeneity of the terminal clusters for age groups by mapping them on the dendrograms and analyzing the cluster diversity index.

The diversity measure was adapted from an ecological study (Heck et al 1975), where it was used to evaluate the effective number of species within a population with the minimum diversity 1 in case of a single dominant species and the maximum corresponding to a total number of evenly abundant species. Age group label diversity indexes for the entire set and nutrition-only dendrograms were substantially higher than for the laboratory-only dendrogram, which was confirmed by distribution statistics (Fig 4C, D and Appendix Fig 7). As observed in the density plot, a large portion of terminal clusters in the lab-only analysis had a diversity index less than 2 (47.3%), while this percentage was 14.5% for the entire set and 10.8% for the nutrition-only dendograms. Furthermore, in the laboratory-only dendrogram, we discovered large branches (1,024 and 722 samples) composed predominately of youth and adult caries populations, respectively, with low to medium median diversity indexes (Fig 4B). The data suggest that laboratory variables may have a discriminative effect in separating youth and adult caries subpopulations and that age stratification of the caries population has potential to improve caries risk prediction models. Age-based stratification was further confirmed by t-distributed stochastic neighbor embedding and uniform manifold approximation and projection dimensionality reduction methods conducted on the caries population with complete variable set (Appendix Fig 8).

Caries Risk Factors Targeted Analysis

We next assessed the known nutrition and laboratory variables associated with caries. Sugar is a known risk factor for dental caries (Pitts et al 2017; Lagerweij and van Loveren 2020). To evaluate its contribution to disease heterogeneity, we integrated food taxonomy data with the examined caries disease phenotypes and further stratified food variables according to their sugar content. The *pvclust* analysis of high-sugar foods demonstrated a more stable cluster structure associated with the caries subpopulation (Fig 5). Hierarchical clustering analysis revealed 19 bootstrap resampling supported clusters of meta-nutritional variables with high-sugar content specific to the caries group (Fig 5A). Further functional analysis revealed clusters with socially recognizable patterns such as a cluster with tea, pastries, cookies, candies, and sugar-additive variables (Fig 5B). Within the context of high-sugar-content foods, the caries population was uniquely associated with a specific set of beverages, dairy products, fruits, condiments, and desserts (Fig 5). Apple juice, energy drinks, and protein and nutritional powders are some examples of high-sugar caries–associated beverages. Dairy products such as flavored milk, milkshakes, yogurt, and ice cream also fell into this category. Protein, vegetable, and grain food groups were almost exclusively low in sugar content and not associated with caries in this analysis.

Lead exposure is another known risk factor that appeared to be associated with caries occurrence (Billings et al 2004; Pradeep and Hegde 2013). For example, in pediatric patients, increased salivary and enamel lead levels have demonstrated a positive correlation with an increase in dental caries severity (Pradeep and Hegde 2013). Individuals with caries were found to have significantly higher average blood lead levels (μg/dL) than those without caries (Appendix Fig 9). In this analysis, we compared correlation structures by stratifying the population into "low-lead" versus "medium-/high-lead" groups (see Appendix Methods) (Appendix Fig 10 and Appendix Table 2). Across low-lead/high-lead groups and no-caries/high-caries groups, correlation network analysis revealed distinctive trends in continuous laboratory blood values, dietary factors, and body weight metrics for each group. Of most interest are the medium/highly correlated (Spearman's $|r| > 0.4$) variables exclusive to the medium-/high-lead caries subpopulation (in relation to the low-lead no caries control). We find a positive correlation between cadmium, cotinine, and hematocrit, iron, and bilirubin (Appendix Table 2). Among both high- and

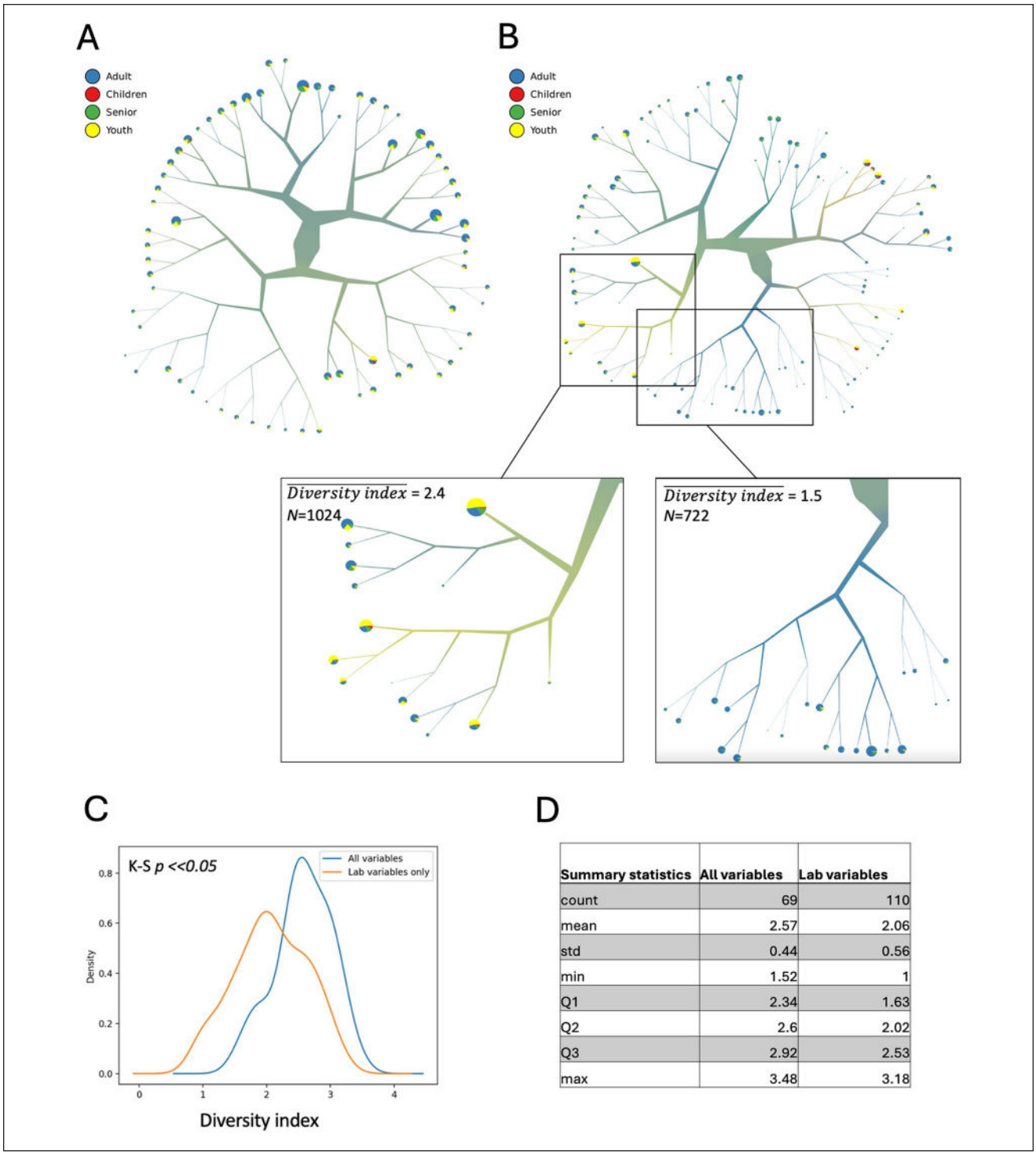
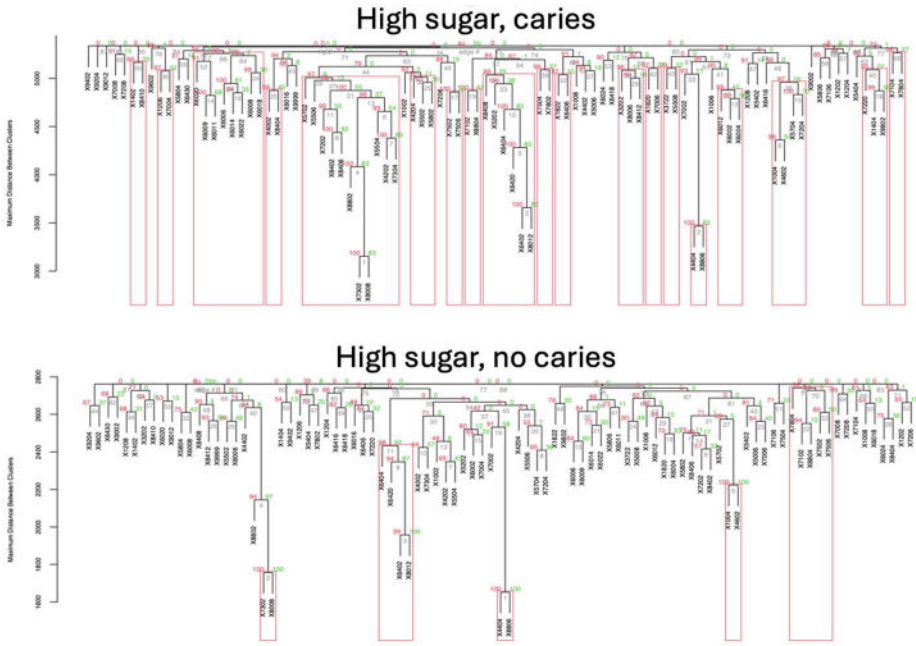


Figure 4. Spectral hierarchical clustering analysis of the caries subpopulation with age group annotation. (**A**) Dendrogram based on the full variables set and (**B**) dendrogram based on the laboratory variables–only subset. Branches with low and medium diversity indicating dominance by fewer age groups are shown in separate blocks. (**C**) Age group diversity distributions for the full and laboratory variables subsets (K-S p, Kolmogorov–Smirnov test *P* value). (**D**) Diversity summary statistics distributions for the full and laboratory variables subsets.

low-lead caries Spearman's rank correlation networks, a strong positive correlation exists in relation to dietary components such as fat and cholesterol in both caries and no-caries subgroups. In sum, the data show expected associations of sugars and lead with dental caries but also new possibilities and high granularity by pinpointing specific food types and laboratory variables that may lead to more precise predictive modeling.

A



B

VARIABLE	CLUSTER #	DESCRIPTION	CATEGORY
7004	2	Apple juice	beverages
7006	13	Other fruit juice	beverages
7104	19	Diet sport and energy drinks	beverages
7202	5	Soft drinks	beverages
7204	17	Fruit drinks	beverages
7220	18	Smoothies and grain drinks	beverages
7304	5	Tea	beverages
7504	10	Wine	beverages
7802	10	Flavored or carbonated water	beverages
9802	18	Protein and nutritional powders	beverages
1206	2	Flavored milk, lowfat	dairy
1402	1	Milk shakes and other dairy drinks	dairy
1404	18	Milk substitutes	dairy
1820	13	Yogurt, regular	dairy
1822	11	Yogurt, Greek	dairy
5802	6	Ice cream and frozen dairy desserts	dairy
8006	12	Cream cheese, sour cream, whipped cream	dairy
6002	16	Apples	fruit
6004	16	Bananas	fruit
6006	3	Grapes	fruit
6008	3	Peaches and nectarines	fruit
6009	3	Strawberries	fruit
6011	3	Blueberries and other berries	fruit
6012	16	Citrus fruits	fruit
6014	3	Melons	fruit
6018	3	Other fruits and fruit salads	fruit
6020	3	Pears	fruit
6022	3	Pineapple	fruit
3202	12	Rice mixed dishes	grains
4002	4	Rice	grains
4202	5	Yeast breads	grains
4204	6	Rolls and buns	grains
3722	14	Peanut butter and jelly sandwiches	other
5008	14	Pretzels/snack mix	other
5202	9	Crackers, excludes saltines	other
5502	6	Cakes and pies	other
5504	5	Cookies and brownies	other
5506	5	Doughnuts, sweet rolls, pastries	other
5702	5	Candy containing chocolate	other
5704	17	Candy not containing chocolate	other
8402	5	Tomato-based condiments	other
8404	4	Soy-based condiments	other
8406	5	Mustard and other condiments	other
8408	9	Olives, pickles, pickled vegetables	other
8410	1	Pasta sauces, tomato-based	other
8412	12	Dips, gravies, other sauces	other
8802	5	Sugars and honey	other
6406	11	Other red and orange vegetables	vegetables

Figure 5. Consumption patterns of high-sugar foods in caries-stratified populations. (**A**) Hierarchical clustering results for the National Health and Nutrition Examination Survey (NHANES) meta-nutritional variables with sugar content specification comparing caries and no-caries populations. (**B**) List of high-sugar variables found in approximately unbiased (AU)–significant clusters within the caries subpopulation. Cluster # indicates AU-significant cluster affiliation of each variable.

Discussion

Large population database sources, such as NHANES, offer invaluable resources for dental epidemiology but simultaneously pose challenges for machine learning approaches due to data heterogeneity, varied sample sizes, missing data, and variations in data collection. These factors necessitate thorough preprocessing before robust machine learning frameworks can be applied. Moreover, interpreting complex patterns within high-dimensional data requires effective data visualization and pattern identification. The bioinformatics pipeline developed here provides a structured approach to address these challenges, offering utility for database users and curators. By systematically handling missingness, outliers, and variable selection, the pipeline facilitates the creation of analysis-ready datasets and, crucially, aids the discovery of clinically relevant heterogeneity including nutrition and laboratory variables that can inform more precise statistical and machine-learning models. Using the caries dataset as a case study, we find that systematic preprocessing, including addressing potential redundant variable types (eg, closely related blood biochemistry markers) can potentially streamline subsequent analyses.

The application of the pipeline combined with network-based and clustering visualization approaches provided several key insights into dental caries heterogeneity within the NHANES population. We find intriguing diverging patterns of variable similarity between different age groups, particularly comparing children (<5 y) with senior (≥65 y) populations, for the complete dataset and for nutrition- and laboratory-only subsets. Interestingly, the disease has a bimodal onset, with the children and senior groups exhibiting the largest and most distinct patterns of variable similarity, suggesting different etiological factors or disease manifestations at these life stages.

The pronounced clustering of laboratory variables within the early-childhood subgroup suggests a strong link between S-ECC and systemic health. Young children are particularly vulnerable to nutritional and metabolic deficiencies that can manifest orally. Anemia has been associated with increased caries risk in children (Schroth et al 2013; Delimont et al 2021), and our findings show that indicators for nutritional status such as iron and vitamin D are prominent in this cluster. This suggests that in young children, dental caries may not be merely a localized disease but can serve as a sentinel marker of underlying systemic health issues.

The significant large cluster primarily composed of laboratory and nutrition variables identified through *pvclust* was associated with the senior (≥65 y) and children (<5 y) age groups in age-stratification cluster analysis. This pronounced age-driven heterogeneity strongly suggests that age stratification is critical for developing accurate predictive models for dental caries. Furthermore, our analysis of the nutrition-only subset, stratifying foods by sugar content and examining associations within the caries population, revealed recognizable dietary patterns. The examination of consumed food taxonomy and its sugar content revealed food patterns and types associated with caries, such as the clustering of habitual consumption of pastries and specific sugary foodstuffs/beverages, expanding known dietary risk factors at a granular level.

Analysis stratified by lead exposure showed diverging correlation network patterns in the caries and no-caries subgroups. Recent systematic reviews confirm a positive association between lead exposure and caries, particularly in primary dentition (Lee et al 2024), and the NHANES database shows higher average blood lead levels in individuals with caries. However, this relationship can be confounded by socioeconomic factors. Our analytical approach did not find a direct correlation between blood lead levels alone and caries. Instead, our network-based analysis revealed a strong correlation between cadmium and cotinine specifically in the high-lead population, extending the previously established association between smoking by-products (cotinine), heavy metals (cadmium), and caries (Akinkugbe et al 2019) to a high-lead context. This tight cluster of exposures suggests common environmental sources (eg, tobacco smoke, industrial pollutants) that are known to be prevalent in lower socioeconomic settings (Hajat et al 2015). Therefore, our data indicate that lead may be a powerful indicator of high-risk environments. While plausible biological mechanisms for lead's cariogenicity have been proposed, including disruption of enamel mineralization, impaired immune function (Billings et al 2004; Lee et al 2024), and impaired salivary function in animal models (Watson et al 1997), our findings suggest the higher caries risk is more likely explained by these co-occurring exposures rather than an independent effect of lead alone.

Beyond the insights into lead exposure, the network analysis uncovered other potentially important correlations. For instance, we identified moderate-to-high correlations between serum iron, hemoglobin, and bilirubin in the high-lead caries population. These hematologic links warrant further investigation regarding the potential interactions between lead toxicity, iron metabolism, systemic inflammation, and caries risk. Furthermore, the unexpected association between reported hours slept and age within the low-lead no-caries group opens new avenues to explore potential interactions between sleep patterns, environmental exposures, and caries susceptibility through behavioral or systemic inflammatory pathways.

Altogether, our work demonstrates a robust data-cleaning–subtype discovery pipeline that could be applied to investigate other health conditions using NHANES and similar databases for machine learning predictive modeling. This integrative approach systematically addresses data quality issues and facilitates exploratory analysis to reveal data patterns associated with subtypes and variables associated with the clinical heterogeneity of caries. The discovery of significant age-driven heterogeneity and novel factor associations highlights the value of such pipelines for generating hypotheses and guiding the development of more precise and robust machine learning predictive models in dental research. However, our analysis was limited to NHANES 2017–2018 data (the latest available release at the study conception), and future work using multiyear designs will be needed to establish longitudinal trends.

In conclusion, applying a comprehensive bioinformatics pipeline to NHANES data successfully identified substantial age-driven heterogeneity in dental caries, suggesting stratification is crucial for future predictive modeling. The approach also uncovered novel associations between caries status, lead/pollutant exposure, specific laboratory markers and food types, as well as sleep patterns, reflecting additional disease markers in susceptible populations. This work demonstrates the value of integrating data science techniques with large-scale observational data to gain deeper insights into complex, multifactorial diseases.

Author Contributions

A. Orlenko, contributed to conception and design, data acquisition, analysis, and interpretation, drafted and critically revised manuscript; J.D. Mure, J.I. Gluch, contributed to conception, data analysis and interpretation, drafted and critically revised the manuscript; J. Gregg, contributed to data analysis, drafted and critically revised the manuscript; C.W. Compher, contributed to data acquisition, drafted and critically revised the manuscript; Z. Ren, contributed to data interpretation, drafted and critically revised the manuscript; H. Koo, J.H. Moore, contributed to data conception and design, drafted and critically revised the manuscript. All authors provided final approval and agreed to be accountable for all aspects of the work.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by National Library of Medicine, National Institutes of Health grant LM010098. Z. Ren was supported by the National Institute of Dental and Craniofacial Research Postdoctoral Training Program under award R90DE031532.

ORCID iDs

Z. Ren https://orcid.org/0000-0002-3553-1958
H. Koo https://orcid.org/0000-0001-9143-2076

References

Akinkugbe AA, Moreno O, Brickhouse TH. 2019. Serum cotinine, vitamin D exposure levels and dental caries experience in U.S. adolescents. Community Dent Oral Epidemiol. 47(2):185–192. https://doi.org/10.1111/cdoe.12442

Billings RJ, Berkowitz RJ, Watson G. 2004. Teeth. Pediatrics. 113(4 Suppl):1120–1127.

Centers for Disease Control and Prevention. 2018. National Health and Nutrition Examination Survey data, 2017–2018. US Department of Health and Human Services, Centers for Disease Control and Prevention; [accessed 2025 Jun 22]. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017

Delimont NM, Carlson BN, Nickel S. 2021. Dental caries are associated with anemia in pediatric patients: a systematic literature review. J Allied Health. 50(1):73–83.

Dye BA, Afful J, Thornton-Evans G, Iafolla T. 2019. Overview and quality assurance for the oral health component of the National Health and Nutrition Examination Survey (NHANES), 2011–2014. BMC Oral Health. 19(1):95. https://doi.org/10.1186/s12903-019-0777-6

Gregg JT, Moore JH. 2023. STAR_outliers: a python package that separates univariate outliers from non-normal distributions. BioData Mining. 16(1):25. https://doi.org/10.1186/s13040-023-00342-0

Hajat A, Hsia C, O'Neill MS. 2015. Socioeconomic disparities and air pollution exposure: a global review. Curr Environ Health Rep. 2(4):440–450. https://doi.org/10.1007/s40572-015-0069-5

Heck KL, van Belle G, Simberloff D. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology. 56(6):1459–1461. https://doi.org/10.2307/1934716

Lagerweij M, van Loveren C. 2020. Chapter 7: sugar and dental caries. Monogr Oral Sci. 28:68–76. https://doi.org/10.1159/000455373

Lee G et al. 2024. The association between lead exposure and dental caries: a systematic review. Caries Res. 58(3):141–152. https://doi.org/10.1159/000537826

Marcenes W et al. 2013. Global burden of oral conditions in 1990-2010: a systematic analysis. J Dent Res. 92(7):592–597. https://doi.org/10.1177/0022034513490168

Pfeiffer CM, Lacher DA, Schleicher RL, Johnson CL, Yetley EA. 2017. Challenges and lessons learned in generating and interpreting NHANES nutritional biomarker data. Adv Nutr. 8(2):290–307. https://doi.org/10.3945/an.116.014076

Pitts NB et al. 2017. Dental caries. Nat Rev Dis Primers. 3:17030. https://doi.org/10.1038/nrdp.2017.30

Pradeep KK, Hegde AM. 2013. Lead exposure and its relation to dental caries in children. J Clin Pediatr Dent. 38(1):71–74. https://doi.org/10.17796/jcpd.38.1.lg8272w848644621

Richards D. 2013. Oral diseases affect some 3.9 billion people. Evid Based Dent. 14(2):35. https://doi.org/10.1038/sj.ebd.6400925

Schroth RJ et al. 2013. Vitamin D status of children with severe early childhood caries: a case–control study. BMC Pediatr. 13(1):174. https://doi.org/10.1186/1471-2431-13-174

Schwartz GW et al. 2020. TooManyCells identifies and visualizes relationships of single-cell clades. Nat Methods. 17(4):405–413. https://doi.org/10.1038/s41592-020-0748-5

Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 22(12):1540–1542. https://doi.org/10.1093/bioinformatics/btl117

Tinanoff N et al. 2019. Early childhood caries epidemiology, aetiology, risk assessment, societal burden, management, education, and policy: global perspective. Int J Paediatr Dent. 29(3):238–248. https://doi.org/10.1111/ipd.12484

US Department of Agriculture, US Department of Health and Human Services. 2020. Dietary guidelines for Americans, 2020–2025. 9th ed. USDA, HHS; [accessed 2025 Jun 22]. https://www.dietaryguidelines.gov/

Verardi V, Vermandele C. 2018. Univariate and multivariate outlier identification for skewed or heavy-tailed distributions. Stata J. 18(3):517–532. https://doi.org/10.1177/1536867X1801800303

Watson GE, Davis BA, Raubertas RF, Pearson SK, Bowen WH. 1997. Influence of maternal lead ingestion on caries in rat pups. Nat Med. 3(9):1024–1025. https://doi.org/10.1038/nm0997-1024

Willemink MJ et al. 2020. Preparing medical imaging data for machine learning. Radiology. 295(1):4–15. https://doi.org/10.1148/radiol.2020192224