nature portfolio

Corresponding author(s): Miguel Simancas-Pallares

Last updated by author(s): Feb 18, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For the discovery cohort, data collection was completed via on-site clinical exams which provided the surface-level caries information. For the first replication cohort (NHANES - US), data was extracted from their repository. For the second replication cohort (Swedish registry on caries and periodontal disease) data was extracted from electronic records, and the third replication cohort (Wide-Smiles - Australia) their data was collected via on-site clinical exams. In addition, supragingival plaque samples were obtained during this clinical exam for the discovery cohort (subsample of 5%). Additional demographic and oral health-related information was collected in ZOE 2.0 via questionnaires administered to children's guardians in the language of their preference. Genotyping was performed using DNA from saliva samples at the Center for Inherited Disease Research (CIDR, Johns Hopkins University) using the Infinium Global Diversity Array-8 v1.0, offering 1,905,000 genotyped single nucleotide polymorphisms (SNPs). Paired-end reads were trimmed of adapter sequences using Trim Galore (Babraham Institute, Cambridge, UK) and classified with Kraken2 and Bracken 2.5 using a custom database including human, fungal, bacterial, and the expanded Human Oral Microbiome Database (eHOMD) genomes to produce an initial taxonomic composition profile. Additional details on data collection can be found in PMID: 30838597; 30838958
Data analysis	Latent class analysis was conducted using Mplus and R. All other analyses were performed using Stata. No particular/unique codes (conducive to these results) were used in this study. Statistical codes for latent class analysis are publicly available.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Clinical data of the ZOE 2.0 study are available via an institutional data repository: https://cdr.lib.unc.edu/concern/data_sets/kk91fv385 Clinical data of the NHANES are available via: https://wwwn.cdc.gov/nchs/nhanes/ Clinical data of the Swedish and Australian cohorts are not publicly available and cannot be shared internationally; investigators interested in collaborating with these cohorts are encouraged to contact the study authors. Molecular data (human genotypes and microbiome sequencing files) are available via the dbGaP repository under the umbrella "Trans-Omics for Precision Dentistry and Early Childhood Caries-TOP-DECC" and accession: phs002232.v1.p1.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	In this study, parent's of participants reported children's sex. However, genotyping data was available for the discovery cohort as well.
Population characteristics	This study included four population-based cohorts of preschool (3-to-5 years of age); one discovery and three replication cohorts. The first called discovery cohort, was a well-characterized sample of 6,404 preschool-age enrolled in the ZOE 2.0 study (2016-2019) in North Carolina, US, and that served as the main analysis cohort. Participating children attended public preschools (Head Start) in a state-wide sample in NC. From this cohort, the mean age was 53 months and 50% were girls. The first replication cohort was comprised by 3,958 3-to-5 year-old children from 7 cycles of the National Health and Nutrition Examination Survey (NHANES). In this cohort, the mean age was 53 months and 49% were girls. The second replication cohort comprised 208,112 age-matched children (mean age 50 months; 48% girls) participating in the Swedish Quality Registry for Caries and Periodontal Diseases (SKaPa). The third replication cohort, included 7,997 age-matched participants from Australia with a mean age of 54 months and 49% girls. Additional details on the discovery cohort can be found in: PMID: 33139633
Recruitment	In the ZOE Study, participating children attended public preschools (Head Start centers) that were invited to participate in the study via their local coordinators. Details of the study protocol including recruitment strategy can be found in: PMID: 30838597 and 30566750
Ethics oversight	Ethics approval were obtained from all participating institutions/studies

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Total sample size for this study included 226,471 3-to-5 year-old children. From this overall sample size, 6,404 participants were from the discovery cohort (ZOE 2.0 - United States), 3,958 from the first replication cohort (NHANES - United States), 208,112 from the second replication cohort (Swedish registry for caries and periodontal disease), and 7,997 from the third replication cohort (Wide Smiles- Australia). Details of sample size calculation for the discovery cohort was based on the study main objective and details of such calculations can be found in PMID: 33139633. The sample size for microbiome analyses was determined based on a convenience sample from the parent study as a case-control with a 1:1 ratio (150 cases and 150 non-cases).
Data exclusions	No data was excluded
Replication	This study included a discovery cohort and three replication cohorts. These three replication cohorts were age-matched and they had different prevalences of disease (which was one of the study hypotheses). The latent structure that was identified in the discovery cohort was replicated in the additional cohorts. In all cohorts, analyses began from identifying from class 1 until non-identification was concluded. The replication strategy started from an exploratory fashion for all cohorts-- starting from 1-class model until non-identification was concluded.

Randomization	There were no randomization of participants in this study as this is an observational study
Blinding	Blinding was not considered in this study given the observational nature of the study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This was an observational study and therefore not registered in any website as a clinical trial
Study protocol	Relevant references for study protocols can be found in PMIDs: 30838597; 30838958
Data collection	Data collection for ZOE study took place between 2016 and 2019. Additional details on protocols for data collection can be accessed in PMID: 33139633
Outcomes	For the discovery cohort, the main outcome in this study was dental caries experience using ICDAS criteria for caries lesion detection at the surface level. The same outcome was considered for data from Sweden and Australia. Although the NHANES data did not measure caries using this criteria, their clinical criteria resembles the ICDAS established/severe definition from the other cohorts as well. Additional details for clinical outcomes in this study can be found in PMID: 30838597. The complete study description and overarching goal can be found in PMID: 33139633. We used this surface-level caries information to identify disease subtypes and their association with behavioral (oral health behaviors) and biological (microbiome and heritability) risk factors.