Clustered Tooth Surfaces into Biologically Informative Caries Outcomes

ABSTRACT
Dental caries affects most adults worldwide; however, the risk factors for dental caries do not necessarily exert their effects uniformly across all tooth surfaces. Instead, the actions of some risk factors may be limited to a subset of teeth/surfaces. Therefore, we used hierarchical clustering on tooth surface-level caries data for 1,068 Appalachian adults (ages 18-75 yrs) to group surfaces based on co-occurrence of caries. Our cluster analysis yielded evidence of 5 distinct groups of tooth surfaces that differ with respect to caries: (C1) pit and fissure molar surfaces, (C2) mandibular anterior surfaces, (C3) posterior non-pit and fissure surfaces, (C4) maxillary anterior surfaces, and (C5) mid-dentition surfaces. These clusters were replicated in a national dataset (NHANES 1999-2000, N = 3,123). We created new caries outcomes defined as the number of carious tooth surfaces within each cluster. We show that some cluster-based caries outcomes are heritable (i.e., under genetic regulation; p < 0.05), whereas others are not. Likewise, we demonstrate the association between some cluster-based outcomes and potential risk factors such as age, sex, educational attainment, and toothbrushing habits. Together, these results support the notion that permanent dentition can be subdivided into groups of tooth surfaces that are useful for understanding the factors influencing cariogenesis. Abbreviations: COHRA, Center for Oral Health in Appalachia, the principal study sample; C1-5, clusters 1-5, groups of similarly behaving tooth surfaces identified through hierarchical clustering; DMFS index, decayed, missing, or filled surfaces, a traditional caries measure representing the number of affected surfaces across the entire dentition; DMFS<i>1-5</i>, partial DMFS indices representing the number of affected surfaces within a hierarchical cluster; and NHANES, National Health and Nutrition Examination Survey, the secondary study sample.

KEY WORDS: dental caries, permanent dentition, white spots, hierarchical clustering, cluster analysis, heritability.

INTRODUCTION
Dental caries, which affects the great majority of adolescents and adults throughout the world, is a multi-factorial disease caused by the effects of numerous environmental, behavioral, and genetic factors. Many risk factors have been identified, such as: host genetics (Horowitz et al., 1958); environmental exposures, including fluoride, cariogenic bacteria, and pH-altering agents; behavioral factors, including diet and oral hygiene; characteristics of the dentition, including enamel composition and positions and morphology of teeth; characteristics of the oral environment, including saliva composition, flow rate, and pH buffering capacity; and demographic factors, including age, sex, race, ethnicity, socio-economic status, and access to oral health care (Hunter, 1988). This complexity is further compounded by the disease phenotype, which may manifest as innumerable combinations of caries lesions across tooth surfaces of the full dentition. Because caries risk factors may exert differential effects across tooth surfaces of the permanent dentition, measurable caries experience may be modeled as the cumulative result of multiple superimposed patterns of decay due to the various risk factors (Batchelor and Sheiham, 2004; Shaffer et al., 2012a).

In epidemiological studies, however, caries experience is typically reduced to a single measure of decay, such as the widely used DMFT/S index (calculated as the sum of decayed, missing due to decay, or filled/restored teeth/surfaces). Such global measures of caries experience ignore the fact that categories of tooth surfaces exhibit differences in susceptibility to decay and are differentially affected by risk factors. Because they ignore the patterns of dental caries across the dentition, global measures of decay such as DMFT/S index may be limited in their ability to identify caries risk factors that exert their effects on specific categories of tooth surfaces. Indeed, previous studies have demonstrated that modeling the patterns of tooth decay is beneficial for epidemiological (Psoter et al., 2003, 2009) and genetic studies (Shaffer et al., 2012a, 2012b). We used hierarchical clustering analysis to group tooth surfaces into categories based on co-occurrence of caries lesions. We then generated novel caries outcomes reflecting these tooth-surface categories, and explored their utility for studying caries etiology.
METHODS

COHRA Sample Recruitment and Data Collection

The principal sample used in this study was ascertained through the Center for Oral Health Research in Appalachia (COHRA), a joint effort established in 2000 between the University of Pittsburgh and West Virginia University (Polk et al., 2008). COHRA was designed to study the individual-, familial-, and community-based factors affecting oral health in the Appalachian population, an underserved, low-income group with disproportionately poorer oral health than the general US population (Purnell and Counts, 1998). Participants were recruited according to a household-based protocol requiring at least one biological parent-offspring pair and extending participation to all other household members regardless of legal or biological relationships. In total, 732 households were recruited, which comprised 2,663 participants. Written informed consent was provided by adult participants for themselves and for their participating children. All procedures were approved by the Institutional Review Boards of the relevant Universities. Full details on sample recruitment have previously been described (Polk et al., 2008).

Evidence of dental caries experience was visually assessed with a dental explorer for each surface of each tooth during intra-oral examinations conducted by trained and calibrated dentists or research dental hygienists. Inter- and intra-examiner concordances of caries assessments were high (Polk et al., 2008; Wendell et al., 2010). Tooth surfaces were scored as sound, non-cavitated carious, cavitated carious, restored, missing due to caries, or missing due to other reasons. Caries assessment was performed in accordance with the National Center for Health Statistics Dental Examiners Procedures Manual (Section 4.9.1.3) to maximize comparability with other national datasets. This method yields high-quality, reproducible data for research purposes.

Data on several covariates were also collected. Questionnaire and/or interview was used to assess self-reported data on sex, race, age, birth year, education, toothbrushing frequency, and home water source (city/public vs. well). Three-minute timed expectoration was used to assess unstimulated saliva flow rate (mL/min). A fluoride-specific electrode was used to measure home water fluoride (mg/L) in water samples.

The present study, which aimed to model the patterns of caries lesions across surfaces of the permanent dentition, was limited to the subset of participants (N = 1,068) aged 18 to 75 yrs. We re-coded each tooth surface (of 128 surfaces, i.e., buccal, distal, lingual, and mesial surfaces on all 28 permanent teeth, plus occlusal surfaces on pre-molars and molars, excluding third molars) as 0 for sound or missing due to reasons other than decay, or 1 for non-cavitated, cavitated, missing due to decay, or restored. Note that previous work on this sample has shown that modeling non-cavitated lesions is beneficial for genetics studies (Wang et al., 2010). Applying this coding scheme yielded a matrix of 1,068 participants by 128 surface-level caries affection statuses, which was used as input for cluster analysis.

Hierarchical Clustering

We used Ward's minimum variance method (Ward, 1963), a type of agglomerative hierarchical clustering, to group similarly affected tooth surfaces (i.e., surfaces that exhibit the same caries affection status), starting with each surface as its own cluster, and then successively merging similar clusters to form a clustering hierarchy. The Euclidean distance (i.e., a common measure of statistical similarity between 2 vectors) was used to determine similarity among clusters. Clustering was performed with the R statistical environment (R Foundation for Statistical Computing, Vienna, Austria).

Statistical Analysis

The data matrix of 1,068 participants by 128 tooth-surface-specific caries affection statuses was used as input for hierarchical clustering, which generated clusters of similar tooth surfaces (with respect to occurrence of caries lesions). To help determine stability of the clusters, we repeated the hierarchical clustering on ten randomly chosen subsets of 50% of the full sample. Likewise, because our sample contained related individuals, we repeated our hierarchical clustering using the maximal set of unrelated individuals. Overall, clusters were very robust (see Appendix).

To create novel dental caries outcomes, we calculated the partial DMFS index (i.e., the number of decayed, missing due to decay, or filled surfaces) restricted to surfaces included in each cluster. To test the association of these novel caries outcomes with several potential risk factors, we calculated non-parametric Spearman correlations. Sex, race, home water source (which is related to urban vs. rural residency), and toothbrushing frequency were treated as binary categorical variables, with race dichotomized as white vs. non-white, and toothbrushing frequency dichotomized as 1 or less vs. 2 or more brushings per day. Age, age², birth year (included to assess cohort effects), saliva flow, and home water source fluoride were treated as continuous variables. Educational attainment was coded as 0 to 2, respectively, for up to high school diploma, some college, and undergraduate or advanced degree. For all association tests, unadjusted p values (shown) and family-adjusted p values [calculated with SOLAR (Almasy and Blangero, 1998)] were nearly identical.

We used variance decomposition methods as implemented in SOLAR (Almasy and Blangero, 1998) to calculate the heritability of caries outcomes while adjusting for age and age² (which were the strongest risk factors observed in the correlational analysis). This method models phenotypic variance among all types of relatives as a function of the theoretical degree of genetic sharing (i.e., that parents and offspring share 50% of their genome, siblings share 50%, half-siblings share 25%, unrelated individuals share 0%, etc.). Details for this method as applied to the COHRA sample have previously been reported (Wang et al., 2010; Shaffer et al., 2012a, 2012b). The heritability estimate is interpreted as the proportion of phenotype variance attributable to the cumulative effect of all genes.

NHANES Data

For comparison with our COHRA sample, we also performed hierarchical clustering on the dental caries data available for the NHANES 1999-2000 cohort. Analysis was limited to the subset of 3,123 participants aged 18 to 75 yrs for whom dental caries assessments were made. Caries data were recoded to form a
matrix of 3,123 participants by 128 surfaces, where 0 represents no observed evidence of caries and 1 represents caries or restoration. Third molars were excluded.

RESULTS

Descriptive characteristics of the COHRA study population are shown in the Appendix Table. The sample (N = 1,068) was 63% female and 90% white, with a mean age of 34.7 yrs. Educational attainment was low (16% college graduates, 25% some college). Prevalence of dental caries for each tooth surface, shown in Fig. 1A, was higher in COHRA than in the general US population.

Fig. 2 shows the dendrogram representing the hierarchical clustering of tooth surfaces in COHRA. We identified 5 clusters (C1-C5) of tooth surfaces based on similarity in occurrence of dental caries: (C1) pit and fissure molar surfaces, (C2) mandibular anterior surfaces, (C3) posterior non-pit and fissure surfaces, (C4) maxillary anterior surfaces, and (C5) the mid-dentition surfaces (Fig. 1B). It is important to note that we generated these surface clusters from caries data only, without any information regarding their relative anatomic positions in the permanent dentition. Interestingly, cluster analysis generated contiguous and symmetrical clusters that sensibly recreated the spatial and morphological relationships among tooth surfaces. Cross-validation showed that clusters were robust (see Appendix). Moreover, clusters generated from COHRA were nearly identical to clusters generated from the 1999-2000 NHANES data (Fig. 1C). For both the full COHRA sample and NHANES sample, cluster 5 was further partitioned into sub-categories comprised of maxillary vs. mandibular teeth, although this separation was not consistent across random sub-samples comprised of 50% of the COHRA sample (results not shown).

To create novel dental caries outcomes, we calculated for each participant the number of carious surfaces within each of the 5 clusters, which we denote as DMFS1, DMFS2, DMFS3, DMFS4, and DMFS5. We tested the univariate associations of DMFS1-DMFS5 with a number of potential caries risk factors (Table 1). All caries outcomes were strongly associated with age, age², and birth year (p < 0.001). Other risk factors were associated only with specific clusters, such as the associations of education with DMFS1, DMFS2, and DMFS4 (p values < 0.001), and the suggestive associations of DMFS4 with tooth-brushing frequency, race, and home water source (p values < 0.05). Several associations (such as those with education) were observed for specific clusters that were not observed for the global DMFS index. Many significant covariates had weak effect sizes, which is consistent with the multi-factorial nature of dental caries.

Based on trait similarity among biological relatives in our sample, we estimated the heritability (h², i.e., the proportion of trait variation due to the cumulative effect of all genetic factors) of DMFS1-DMFS5 (Table 2). DMFS2 (h² = 54%, p = 0.002), DMFS3 (h² = 43%, p = 0.004), and DMFS5 (h² = 40%, p = 0.008) were significantly heritable, whereas DMFS1 was suggestively heritable (h² = 27%, p = 0.06), and DMFS4 was not heritable (h² = 0%, p = 0.5).
In this study, we applied hierarchical clustering to tooth-surface-level caries data to group tooth surfaces based on similarity in dental caries experience. Five clusters were observed in our COHRA sample, which were nearly identical to the results of cluster analysis in the NHANES sample, suggesting that these clusters are reproducible in national datasets, and are not unique to the Appalachian population. Moreover, the 5 clusters were very consistent across randomly partitioned halves of our sample, further indicating their robustness to the inclusion of any specific individuals. The clusters exhibited perfect left-right symmetry, and were comprised of sensible, mostly contiguous, tooth surfaces. Despite the fact that the clustering method did not use any information on the positions of the teeth, 4 of the 5 clusters reflected the positional relationships of surfaces across the permanent dentition: anterior vs. posterior and maxillary vs. mandibular (whereas one cluster was comprised of highest-risk pit and fissure molar surfaces).

We generated cluster-specific caries outcomes (DMFS1-DMFS5; representing the number of carious surfaces in a cluster) and showed that some were heritable, indicating the causal role of genetic factors on caries experience, whereas others were not. In particular, DMFS2 (comprised of the anterior mandibular surfaces) was the most heritable outcome, which corresponds to the group of surfaces with the lowest prevalences of caries. In contrast, DMFS4 (comprised of the maxillary incisors), was not heritable (i.e., $h^2 = 0\%$), and corresponds to surfaces with fairly high prevalence of caries. Interestingly, in young children, caries

<table>
<thead>
<tr>
<th>Variable</th>
<th>DMFS</th>
<th>DMFS1</th>
<th>DMFS2</th>
<th>DMFS3</th>
<th>DMFS4</th>
<th>DMFS5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>0.04</td>
<td>0.17</td>
<td>-0.02</td>
<td>0.52</td>
<td>-0.04</td>
<td>0.16</td>
</tr>
<tr>
<td>Race (white vs. other)</td>
<td>0.01</td>
<td>0.79</td>
<td>0.02</td>
<td>0.42</td>
<td>-0.04</td>
<td>0.21</td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>0.19</td>
<td>4E-10</td>
<td>0.12</td>
<td>4E-5</td>
<td>0.27</td>
<td>5E-19</td>
</tr>
<tr>
<td>Age$^2$ (yrs)</td>
<td>0.19</td>
<td>4E-10</td>
<td>0.12</td>
<td>4E-5</td>
<td>0.27</td>
<td>5E-19</td>
</tr>
<tr>
<td>Birth yr</td>
<td>-0.15</td>
<td>6E-7</td>
<td>-0.11</td>
<td>2E-4</td>
<td>-0.24</td>
<td>8E-15</td>
</tr>
<tr>
<td>Educational attainment</td>
<td>0.01</td>
<td>0.76</td>
<td>0.01</td>
<td>0.69</td>
<td>-0.02</td>
<td>0.46</td>
</tr>
<tr>
<td>Saliva flow</td>
<td>-0.03</td>
<td>0.33</td>
<td>0.01</td>
<td>0.69</td>
<td>0.01</td>
<td>0.72</td>
</tr>
<tr>
<td>Home water source fluoride</td>
<td>0.06</td>
<td>0.16</td>
<td>0.09</td>
<td>0.03</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>Toothbrushing frequency</td>
<td>-0.02</td>
<td>0.44</td>
<td>0.00</td>
<td>0.93</td>
<td>0.00</td>
<td>0.92</td>
</tr>
<tr>
<td>Water source (public vs. well)</td>
<td>0.00</td>
<td>0.94</td>
<td>0.03</td>
<td>0.28</td>
<td>0.04</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Table 1. Univariate (Spearman) Correlations between Caries Outcomes and Possible Risk Factors

Bold = associations with $p$ values $\leq 0.05$.
Bonferroni-adjusted alpha for 6 caries outcomes = 0.008.
Bonferroni-adjusted alpha for 10 risk factors = 0.005.
Bonferroni-adjusted alpha for 60 individual tests = 0.0008.
of the maxillary incisors is associated with nighttime bottle feeding (Douglass et al., 2001) and access to sweets (Johnsen et al., 1984); we speculate that environmental or behavioral factors may similarly affect maxillary incisors in adults, perhaps overshadowing any genetic predispositions. This reasoning is supported by the observed associations of DMFS4 with tooth-brushing frequency and home water source. The high heritability of some caries outcomes indicates that they may be better suited for gene mapping than the traditional DMFS index (h² = 42%). Indeed, in another study we have used the cluster-based caries outcomes in a genome-wide association study (Shaffer et al., 2013). Likewise, results showing that different clusters were associated with different potential risk factors, including associations that were not observed for the traditional DMFS index, suggest that clusters may serve as useful outcomes for traditional epidemiological studies, in addition to genetics studies. (See the Appendix for an expanded discussion of these results.) One issue warranting consideration is how to decide on the number of distinct clusters without over-fitting. Various criteria can be imposed on the clustering hierarchy to delineate clusters—that is, the level of within-cluster similarity and between-cluster dissimilarity that defines the clusters. Rather than choosing an a priori distance threshold for determining this, we used two-fold cross-validation in which we repeated hierarchical clustering in randomly chosen halves of the sample (see Appendix). This process demonstrated 5 stable clusters. However, in the full COHRA sample, the next would-be clusters were subdivisions of DMFS5 into maxillary and mandibular surfaces; these subdivisions were also observed for NHANES. Interestingly, heritability of DMFS5 as a whole was greater than each of these subdivisions.

In addition to cross-validation, this study benefits from several other strengths, including our large sample of related individuals with surface-level caries assessments, which facilitated hierarchical clustering and heritability estimation. Also, we were able to replicate our clusters in an independent national dataset. Despite these strengths, however, some limitations merit discussion. First, the method of caries assessment by visual inspection (in both COHRA and NHANES), while appropriate for research purposes, may under-represent caries prevalence. Additionally, teeth missing due to caries (for which all surfaces are counted as carious) and two-surface restorations (i.e., whereby approximal lesions are accessed through the occlusal surface) may have resulted in the erroneous coding of truly non-carious surfaces adjacent to carious surfaces. The effect of such assessment errors would bias the clustering analysis in favor of including multiple surfaces of the same tooth in the same cluster. However, given that 4 of the 5 observed clusters reflected the positional relationships of surfaces in the permanent dentition (e.g., anterior vs. posterior, and maxillary vs. mandibular), this potential bias was likely not detrimental to the clustering.

In conclusion, this study demonstrates the utility of clustering methods for grouping tooth surfaces with similar caries experience, and reinforces the complexity of dental caries etiology. Refining caries phenotype definitions based on the type of surface may assist in identifying the environmental and genetic risk factors that manifest as specific decay patterns, ultimately leading to improved understanding of the multifactorial nature of dental caries. This study is one of few but much-needed efforts to develop novel, biological data-driven outcomes for investigating dental caries.

**ACKNOWLEDGMENTS**

Funding support for this work was provided by the National Institute of Dental and Craniofacial Research, including grants R01-DE014899 and R03-DE021425. Additional support was provided by the University of Pittsburgh School of Dental Medicine, the West Virginia University School of Dentistry, and Eberly College of Arts and Sciences. The content presented herein is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Dental and Craniofacial Research or the National Institutes of Health. The funding sources had no input in study design, data collection, analysis, decision to publish, or manuscript preparation. (See the Appendix for full acknowledgments.) The authors declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

**REFERENCES**


