



Sources of bias in genomics research of oral and dental traits

[Special issue of Community Dental Health, to be disseminated at ‘Random and Systematic Bias in Population Oral Health Research’ IADR symposium, March 2020, Washington DC.]

Cary S Agler and Kimon Divaris

Adams School of Dentistry, University of North Carolina Chapel Hill, Chapel Hill, NC, United States.

Abstract: Evidence regarding the genomic basis of oral/dental traits and diseases is a fundamental pillar of the emerging notion of precision health. During the last decade, technological advances have improved the feasibility and affordability of conducting genome-wide association studies (GWAS) and studying the associations of emanating data with both common and rare oral conditions. Most evidence thus far emanates from GWAS of dental caries and periodontal disease that have tested the associations of several million single nucleotide polymorphisms (SNPs) with typically binary, health vs. disease phenotypes. GWAS offer advantages over the previous candidate-gene studies, mainly owing to their agnostic (*i.e.*, unbiased, or hypothesis-free) nature. Nevertheless, GWAS are prone to virtually all sources of random and systematic error. Here, we review common sources of bias in genomics research with focus on GWAS including: type I and II errors, population stratification and heterogeneity, selection bias, adjustment for heritable covariates, appropriate reference panels for imputation, and gene annotation. We argue that valid and precise phenotype measurement is a key requirement, as GWAS sample sizes and thus statistical power increase. Finally, we stress that the lack of diversity of populations with phenotypes and genotypes is a major limitation for the generalizability and ultimate translation of the emerging genomics evidence-base into oral health promotion for all.

Keywords: genome-wide association studies, genetic epidemiology, bias, dental caries, periodontitis.

Introduction

Evidence regarding the genomic basis of oral/dental traits and diseases is a fundamental pillar of the emerging notion of precision health (Divaris, 2017). During the last decade, technological advances have improved the feasibility and affordability of generating genome-wide association (GWAS) data and studying the genetic underpinning of both common and rare oral conditions. In the oral health domain, most evidence has thus far emanated from GWAS of dental caries and periodontal disease that have tested the associations of several million single nucleotide polymorphisms (SNPs) with typically binary, health vs. disease phenotypes (Morelli *et al.* 2019). GWAS offer advantages over the previous candidate-gene studies, mainly owing to their agnostic (*i.e.*, unbiased, or hypothesis-free) nature. Nevertheless, GWAS are prone to virtually all sources of random and systematic error, as well as reporting bias. Here, we review common sources of bias in genomics research, focusing specifically on GWAS, including: 1) type I and II errors, 2) population stratification and heterogeneity, 3) selection bias, 4) adjustment for heritable covariates, 5) appropriate reference panels for imputation, gene annotation and genotyping, 6) lack of racial/ethnic diversity at the international level in the available cohorts and samples. We conclude that recognizing and adequately

controlling for those known biases will help build a stronger evidence base for the genomic underpinning of oral and dental traits and will ultimately contribute to better individual and population health.

Definition and measurement of analytical endpoints (*i.e.*, GWAS phenotypes)

Similar to any other research study design, valid and precise measurement of the analytical endpoint cannot be overemphasized in the context of GWAS (van der Sluis *et al.*, 2010). Clinical measures of oral and dental diseases, periodontitis and dental caries being the most common, are traditionally challenging to measure with precision and validity, especially in large sample sizes and population cohorts. All GWAS reports of dental caries and periodontitis to-date have been based on cross-sectional data, that are somewhat limited in their potential to accurately identify the sources of tooth loss—which can lead to the under-estimation of periodontitis history. Large-scale studies may also rely on partial-mouth examinations, screenings versus comprehensive examinations, health records, or even self-reported and proxy data for oral diseases (Shungin *et al.*, 2019). All these measurement issues likely introduce non-differential bias in GWAS, diluting potentially true association signals and influencing the replicability of reported findings.

Type I and type II errors

Several approaches exist for quantifying the impact of measurement error and outcome misclassification on study power. Liao and colleagues (2014) demonstrate these effects quantitatively using simulated and real data, and suggest that the impact on power is much larger in the context of misclassification (i.e., in case control studies) versus measurement error (i.e., in quantitative traits). A recent report by Gordon and colleagues (2015) introduced a method and accompanying visualization tools for the estimation of power in genetic association case-control studies, that allows the consideration of different scenarios of (among other features) outcome misclassification error rate.

In terms of study designs that can accommodate genome-wide association analyses, the typical biases associated with selection of participants, cases and controls apply. Cohorts may be relatively underpowered compared to case-control studies for analyses of binary traits, because the latter sample participants based on disease status. However, they offer benefits in terms of possible longitudinal or repeated measurements and opportunities to leverage pleiotropy (i.e., the examination of multiple, related outcomes that may naturally co-occur in the population). These may not be observed in a focused, case-control sample selection that is optimally designed to test a narrower hypothesis. Regardless of the 'parent' study design, it is recommended that information on the accuracy (e.g., repeatability) of examined traits in GWAS is known or estimated, ideally before the study execution (Barendse, 2011), as it may influence or inform downstream experimental procedures and analyses.

Study sample characteristics

The importance of sample size in GWAS cannot be overemphasized (Cantor *et al.*, 2010). Thousands of individuals are needed for GWAS because most allele effects identified for common, complex diseases are modest or small. Small p-values generated from small sample sizes do not necessarily imply trustworthy findings—they could very well represent extreme findings, unlikely to be observed, or be indicative of model misspecification. As mentioned earlier, most GWAS are based on case-control designs. The selection of case and control samples is important and while it may seem advantageous selecting severe or extreme cases in terms of power, especially when there are logistical limitations, it can have the opposite effect for GWAS (McCarthy *et al.*, 2008). The selection of controls in a case-control study is subject to Berkson's bias, a form of selection bias due to the inclusion of participants from specific subpopulations such as those from clinics and hospitals. Alternatively, the use of common 'healthy' controls for contrast against multiple disease outcomes is less likely to induce bias. Latent population substructure (i.e., stratification) can also induce spurious associations unless controlled for (Li and Yu, 2008). These spurious associations are typically a result of varying patterns of racial/ethnic admixture in the study sample. Several well-established methods using ancestry-informative genetic markers exist to account for population substructure, as well as other forms of known or cryptic relatedness that might violate assumptions of independence in GWAS (Aglér *et al.*, 2019).

Type I error is commonly understood as a false positive and type II error as a false negative finding. Balancing the potential for these two types of error is a fundamental requirement in GWAS for two main reasons: the likely modest or weak genetic effects underlying common-complex oral/dental diseases, and the large number of tests conducted. Specifically, it is not uncommon for allelic effects to be in the range of 1.1-1.3 relative magnitude, while 1 million independent tests are conducted. The requirement of implementing a very stringent p-value criterion (typically 5×10^{-8}) for genome-wide significance (protecting from a type I error inflation) comes at the expense of a study's ability to detect small effects (increased type II error). The issue magnifies when study sample sizes are modest, in the range of 10,000 at best, in the case of most single clinical cohorts with dental phenotypes and genotypes.

Another form of bias is the "winner's curse", a term used to describe the relatively common phenomenon wherein the initially discovered measure of association is inflated in the first GWAS compared to its true magnitude (Kraft, 2008). A related issue is the use of a discovery sample for the development of polygenic (i.e., multi-locus) risk scores that may be similarly exaggerated, due to model "over-fitting". To ameliorate the issue with false positive findings, overestimation (Zhong and Prentice, 2008) and overfitting, replication of genetic findings from GWAS in independent, external samples and cohorts is a requirement (McCarthy *et al.*, 2008). It must be acknowledged that non-replication does not necessarily imply lack of a true association but may suggest additional complexity in sample ascertainment, between-study heterogeneity (Nakaoka and Inoue, 2009), population substructure and genetic architecture. In principle, efforts to generalize signals across populations are desirable.

It is important to stress that the primary goal of GWAS is to identify loci of relevance to traits and not the precise or unbiased measurement of specific SNP associations (i.e., effect estimation) within these loci. Effect estimates may be substantially biased and arguably, in most instances, the causal makers remain unknown until substantial follow-up work has been completed in these loci (e.g., bioinformatics annotation, fine mapping, re-sequencing, experimental follow-up, etc.). So far, biological information or prior existing evidence of association (i.e., prior probability of association) are not explicitly incorporated in the discovery stage of most GWAS (Broer *et al.*, 2013)—this may inevitably lead to some promising candidates being missed under the stringent threshold of multiple testing correction. On the other hand, only a small fraction of genetic associations reported by candidate gene studies appear to replicate in the GWAS setting (Siontis *et al.*, 2010), suggesting a substantial false positive rate in the earlier, candidate-gene study, literature.

Adjustment for heritable covariates in genetic models

The role of adjustment in genetic models employed in GWAS is infrequently discussed. Covariates that are typically included *a priori* in these models include study design characteristics (i.e., study site or cluster), popu-

lation substructure (i.e., ancestry principal components, family structure/relatedness), age and sex. The inclusion of additional terms for covariates that are known to be associated with the outcome has been proposed as a strategy for reducing the residual variance in the outcome (thus increasing statistical power for the discovery GWAS) and accounting for potential confounding. However, Aschard and colleagues (Aschard *et al.*, 2015) recently demonstrated how adjusting for heritable covariates (e.g., smoking and diabetes for GWAS of periodontitis) can introduce bias in the GWAS effect estimation. This “collider bias” is induced by the inclusion of a causally associated covariate in the genetic model, creating apparently robust but otherwise spurious associations (Day *et al.*, 2016).

Selection of reference panels for imputation and annotation databases

It is important to acknowledge that although GWAS include a large number of markers (nowadays, several million SNPs), they still cover a small fraction of the entire human genome sequence. Carefully-selected, directly genotyped SNPs cover a substantial proportion of the known, mostly common, variation of the human genome. These “tagging” SNPs can then be used to infer (i.e., impute) non-typed SNPs and haplotypes, using information from panels that include fully-sequenced “reference” genomes. The imputation process offers gains in efficiency (i.e., fewer SNPs needing to be directly measured) and improves the resolution for the characterization of candidate loci. However, imputation can itself introduce biases. For instance, using reference genomes from “healthy” individuals has been shown to significantly bias SNP associations of disease-associated markers, i.e., favouring health-associated alleles (Khankhanian *et al.*, 2015), that are under-represented in the reference panels. These issues may be accentuated when considering trans-ethnic populations or heavily admixed samples.

The annotation of the human genome is far from uniform or balanced. Traditionally, genes with known biological function and appearance in experimental and candidate-gene studies are more likely to be annotated. It is conceivable, and has actually been shown (Haynes *et al.*, 2018), that these genes may be favoured when reporting associations over markers and genes for which less is known, even if the molecular evidence of association is strong. In other words, investigators themselves can introduce biases in their GWAS reports, related to the qualitative interpretation of their findings. (Kraft, 2008). Haynes and colleagues (2018) suggest that the research community can overcome this form of bias by prioritizing empirically derived hypotheses and inferences. On the other hand, other areas of the human genome [e.g., the human leukocyte antigen (HLA) region] are highly polymorphic (Brandt *et al.*, 2015) and are commonly excluded altogether from the reporting of GWAS results.

Selection bias

Selection bias, a known threat to the validity of most types of biomedical research, is also relevant to the GWAS domain. Conceivably, markers associated with severe

health outcomes impacting longevity, may be systematically under-represented in a cross-sectional study of middle-age adults, as they are being selectively removed from the population allele pool. In a similar fashion, selection on or exclusion of specific sub-types of disease from a GWAS may also introduce bias, as it is equivalent to conditioning on a collider (Munafò *et al.*, 2018). Theoretically, this bias can be accounted for, if the selection effect can be quantified (Xiao and Boehnke, 2010) and examined in population-based birth cohorts with longitudinal follow-up. An interesting scenario arises when longitudinal outcomes (e.g., survival, prognosis or incident events) that are conditional on outcome diagnosis and thus susceptibility are interrogated in the context of a GWAS. Such analyses are prone to “index event bias” wherein this form of selection bias can introduce spurious associations, unless accounted for (Dudbridge *et al.*, 2019).

Genotype information quality

High-density genotyping entails the determination (i.e., “calling”) of often millions of single nucleotide polymorphisms and this comes with an unavoidable error rate. Poorly genotyped or imprecisely imputed markers can induce both spurious associations and result in decreased power to detect true associations. For this reason, GWAS employ stringent quality assessment and quality control procedures beginning at the genotyping stage. For instance, to address genotyping platform batch effects, cases and controls may be equally distributed across plates, while other important study variables may be randomized for the same reason. Other sources of error can be attributed to possibly different DNA extraction methods between cases and controls (if they have been ascertained separately or asynchronously). Additional quality filters and exclusions are conventionally applied at the SNP level (i.e., excluding markers that do not meet pre-specified criteria for call rate, imputation quality, Hardy-Weinberg equilibrium, etc.) and at the individual participant level (i.e., sex mismatches, genetic outliers, etc.). Detection of genomic inflation due to residual population stratification or other systematic sources of error can be determined by the generation and inspection of quantile-quantile (Q-Q) plots of observed versus expected association p-values. A consensus report of all analytical procedures, including quality control, for GWAS in the oral/dental domain has been recently reported (Agler *et al.*, 2019).

Lack of diversity

An astonishing figure--in 2009, 96% of participants in GWAS were of European descent, and in 2016 only 20% of participants were not of European descent (Popejoy and Fullerton, 2016). There are several reasons behind this persistent issue including but not limited to available research funding allocation and prioritization, unequal inclusion in biomedical research and historic reasons. Hispanic/Latinos, African, and Indigenous populations continue to be greatly under-represented in the genomics evidence base to-date. The systematic exclusion of population segments from the evidence base of genetic associations with health outcomes is problematic from multiple standpoints, ranging from biological to social justice.

Importantly, genetic association signals discovered from GWAS in populations of European descent do not always transfer or generalize to non-European populations. This is due to differences in genetic architecture (i.e., linkage disequilibrium), association of risk polymorphisms with ancestry-informative markers, allele frequency, study power, as well as other reasons (Zanetti and Weale, 2018). These authors suggest that, although transethnic differences may be at play in some instances, strong causal effects are largely shared among human populations, motivating the use of transethnic data for fine mapping of these regions. For this reason, the current lack of racial/ethnic diversity overall, and specifically in the available cohorts and samples with oral/dental phenotypes and genotypes, is a major limitation that must be addressed. This issue hampers the generalizability and ultimate translation of the emerging genomics evidence-base into what is aspirationally envisaged as oral health promotion for all.

Conclusion and recommendations

While numerous sources of bias exist in performing and interpreting GWAS of oral and dental traits, these are analogous to most other study designs. Here, we emphasize that measurement is a primary source of bias in the oral/dental domain, as both dental caries and periodontal disease are subject to important and variable sources of measurement error and misclassification. Efforts to improve measurement are best invested prior to study execution, whereas quantification of the possible magnitude of error introduced can be carried out *post hoc*. We caution that genetic models employed in GWAS are subject to known issues that are relevant in observational research, including collider bias (i.e., when adjusting for heritable covariates) and non-differential misclassification bias towards the null. Additional issues that mainly result in necessary caution in interpretation include type I and II errors, available panels for imputation and information on gene annotation. We stress that the lack of racial/ethnic diversity in the currently available cohorts and samples with oral/dental traits and genotypes is a critical issue that must be addressed with concerted, international efforts. Recognizing and adequately controlling for those known biases will help expand our understanding of the genetic underpinnings of oral and dental traits and ultimately help improve oral health and care.

Authors' individual contributions

CSA and KD contributed to the conceptualization of the manuscript, literature review, evidence synthesis, drafting and critical revisions of the manuscript

Acknowledgments

CSA and KD acknowledge support from a grant from the National Institutes of Health/National Institute of Dental and Craniofacial Research U01DE025046.

References

- Agler, C. S., Shungin, D., Ferreira Zandona, A. G., Schmadeke, P., Basta, P. V., Luo, J., . . . Divaris, K. (2019): Protocols, Methods, and Tools for Genome-Wide Association Studies (GWAS) of Dental Traits. *Methods in Molecular Biology* **1922**, 493-509.
- Aschard, H., Vilhjalmsón, B. J., Joshi, A. D., Price, A. L. and Kraft, P. (2015): Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *American Journal of Human Genetics* **96**, 329-339.
- Barendse, W. (2011): The effect of measurement error of phenotypes on genome wide association studies. *BMC Genomics*, **12**, 232-232.
- Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J. and Meyer, D. (2015): Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3 (Bethesda)* **5**, 931-941.
- Broer, L., Lill, C. M., Schuur, M., Amin, N., Roehr, J. T., Bertram, L., Ioannidis, J.P.A. and van Duijn, C. M. (2013): Distinguishing true from false positives in genomic studies: p values. *European Journal of Epidemiology* **28**, 131-138.
- Cantor, R. M., Lange, K. and Sinsheimer, J. S. (2010): Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American Journal of Human Genetics* **86**, 6-22.
- Day, F. R., Loh, P. R., Scott, R. A., Ong, K. K. and Perry, J. R. (2016): A Robust Example of Collider Bias in a Genetic Association Study. *American Journal of Human Genetics* **98**, 392-393.
- Divaris, K. (2017): Fundamentals of Precision Medicine. *Compendium of Continuing Education in Dentistry* **38** (8 Suppl), 30-32.
- Dudbridge, F., Allen, R. J., Sheehan, N. A., Schmidt, A. F., Lee, J. C., Jenkins, R. G., . . . Patel, R. S. (2019): Adjustment for index event bias in genome-wide association studies of subsequent events. *Nature Communications* **10**, 1561-1510.
- Gordon, D., Haynes, C., Blumenfeld, J. and Finch, S. J. (2005): PAWE-3D: Visualizing power for association with error in case-control genetic studies of complex traits. *Bioinformatics* **21**, 3935-3937.
- Haynes, W. A., Tomczak, A. and Khatri, P. (2018): Gene annotation bias impedes biomedical research. *Scientific Reports* **8**, 1362.
- Khankhanian, P., Din, L., Caillier, S. J., Gourraud, P. A. and Baranzini, S. E. (2015): SNP imputation bias reduces effect size determination. *Frontiers in Genetics* **6**, 30.
- Kraft, P. (2008): Curses--winner's and otherwise--in genetic epidemiology. *Epidemiology* **19**, 649-651; discussion 657-648.
- Li, Q. and Yu, K. (2008): Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology* **32**, 215-226.
- Liao, J., Li, X., Wong, T.-Y., Wang, J. J., Khor, C. C., Tai, E.S., Aung, T., Teo, Y.-Y. and Cheng, C.-Y. (2014): Impact of measurement error on testing genetic association with quantitative traits. *PLoS One*, **9**, e87044.
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. and Hirschhorn, J. N. (2008): Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, **9**, 356-369.
- Morelli, T., Agler, C. S. and Divaris, K. (2019): Genomics of Periodontal Disease and Tooth Morbidity. *Periodontology* **2000**, DOI: 10.1111/prd.12320.
- Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. and Davey Smith, G. (2018): Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology* **47**, 226-235.

- Nakaoka, H. and Inoue, I. (2009): Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse. *Journal of Human Genetics* **54**, 615-623.
- Popejoy, A. B. and Fullerton, S. M. (2016): Genomics is failing on diversity. *Nature* **538**, 161-164.
- Shungin, D., Haworth, S., Divaris, K., Agler, C. S., Kamatani, Y., Keun Lee, M., . . . Johansson, I. (2019): Genome-wide analysis of dental caries and periodontitis combining clinical and self-reported data. *Nature Communications* **10**, 2773.
- Siontis, K. C., Patsopoulos, N. A. and Ioannidis, J. P. (2010): Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *European Journal of Human Genetics* **18**, 832-837.
- van der Sluis, S., Verhage, M., Posthuma, D. and Dolan, C. V. (2010): Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PLoS One*, **5**, e13929.
- Xiao, R. and Boehnke, M. (2009): Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology* **33**, 453-462.
- Zanetti, D. and Weale, M. E. (2018): Transethnic differences in GWAS signals: A simulation study. *Annals of Human Genetics* **82**, 280-286.
- Zhong, H. and Prentice, R. L. (2008): Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621-634.