SLEEP, 2023, 1–3
https://doi.org/10.1093/sleep/zsac312
Advance access publication 15 December 2022
Editorial

Editorial

Overcoming the underdiagnosis of obstructive sleep apnea to empower genetic association analyses

Tamar Sofer[1,2,3,*]

[1]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA,
[2]Department of Medicine, Harvard Medical School, Brigham and Women's Hospital, Boston, MA, USA and
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[*]Corresponding author. Tamar Sofer, Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, 221 Longwood Ave, Boston, MA 02115, USA. Email: tsofer@bwh.harvard.edu.

Genome-wide association studies (GWAS) of complex traits, that is, phenotypes that are influenced by many genetic variants, have discovered thousands of genetic loci [1] underlying blood pressure, diabetes, lipids, psychiatric, and other traits, including sleep-related phenotypes [2]. However, studies of obstructive sleep apnea (OSA) have been less fruitful in that fewer discoveries have been made. Two major factors have limited OSA GWAS. First, few epidemiologic studies have measured OSA, and those that did so often studied a subset of participants rather than the full sample of large cohort studies, resulting in low sample sizes for GWAS (in comparison with other phenotypes) [3–6]. Second, while the availability of large biobanks that collected genotyping data in conjunction with electronic health records, including the UK Biobank (UKB), FinnGen, and Biobank Japan, accelerated GWAS of many phenotypes, including OSA [7], challenges lingered because OSA is underdiagnosed [8, 9]. As a result, many individuals with OSA are misclassified as "controls". Thus, while current estimates of OSA prevalence in the United States are around 17% in women and 34% in men [10], and similarly, high prevalence is reported elsewhere, the prevalence of OSA status in the UKB is only about 1% [11] and about 8% in FinnGen [7] (gender combined).

Misclassification of OSA reduces the power to discover genetic associations and biases effect-size estimates, in a manner depending on the OSA prevalence and on the misclassification rate. Figure 1 provides a schematic tabulation of the true OSA status compared to the OSA status observed in a population. Out of $n_{1s} = n_{10} + n_{11}$ individuals with OSA in the population, n_{11} individuals are indeed observed to have OSA, and n_{10} individuals appear to have no OSA, despite having OSA. Define the misclassification rate as $\pi = n_{10}/n_{1s}$, the proportion of individuals with OSA who are erroneously classified. Using the same notation, the prevalence of OSA in the healthcare system or study is $n_{1s}/(n_{1s} + n_{0s}) = n_{1s}/n$. I performed a simulation study to demonstrate how misclassification of OSA may bias genetic effect estimates and reduce power (see https://github.com/tamartsi/OSA_misclassification for code). Using a simple logistic regression model, I assumed that OSA probability depends on a population-based constant, the intercept β_0 (which may be thought of as the average of many factors, including genetic ones), and on a single-modeled genetic variant g via the standard logistic model equation:

$$\text{logit}\left(\Pr\left(\text{OSA} = 1\right)\right) = \beta_0 + g \times \beta_g.$$

The simulations had β_g, the log odds ratio (OR), set to 0.1, corresponding to an OR of 1.10, while β_0 took the values –1.5, –1, and –0.5, corresponding to true underlying OSA prevalence of about 19%, 28%, and 39%. The genetic variant g was sampled from a binomial distribution with probability 0.3 and a count of 0, 1, or 2, representing a genetic allele with frequency 0.3 across two chromosomes. Using the equation above, in each iteration of the simulation OSA probability was computed, and next true OSA status was sampled from the resulting probability. The next step induced misclassification, where individuals with true OSA = 1 had observed OSA with probability $1 - \pi$. Misclassification rate took the values 0.4, 0.6, and 0.8. For context, if the true OSA population prevalence in the UKB and FinnGen is 25%, their misclassification rates are 96% and 68%, respectively. The simulations iterated 1000 times for each combination of true OSA prevalence and misclassification rate, with a total sample size of $n = 20\,000$ in each simulation iteration.

	Observed: no OSA	**Observed: yes OSA**	
True: no OSA	n_{00}	$n_{01} \approx 0$	$n_{0s}=n_{00}+n_{01}$
True: yes OSA	n_{10}	n_{11}	$n_{1s}=n_{10}+n_{11}$
	$n_{s0}=n_{00}+n_{10}$	$n_{s1}=n_{01}+n_{11}$	

Figure 1. True OSA status versus observed OSA status. Tabulation of the observed OSA status against the true, underlying OSA in a given study. The number of individuals in the study is decomposed into individuals in each of the table cells. The marginal, dark gray, cells sum the individuals in the rows and in the columns. In health records, typically we expect that the number of individuals who in truth do not have OSA yet are observed as having OSA is near zero.

Table 1 provides the simulation results. Indeed, the power to detect the association of the genetic variant with OSA is reduced as the misclassification rate is increased: for a modest OSA prevalence of about 19%, a misclassification rate of $\pi = 0.4$ results in 0.76 power while with $\pi = 0.8$, the power is reduced to 0.29. When the true OSA prevalence is higher, the power is higher (when using both the true and the misclassified OSA). Yet, even with a true OSA prevalence of 39%, with $\pi = 0.8$, the power is still very low at 0.37. Further, the estimated variant effect size is reduced toward the null as the misclassification rate increases, with a higher reduction when the true OSA prevalence is higher.

To address the reduced power caused by OSA misclassification, Campos et al. [12] performed a multi-trait analysis, combining OSA GWAS with a GWAS of snoring, and discovered 49 loci associated with OSA, snoring, or both. Multi-trait analyses have been used to discover genetic associations with other trait groups, including blood pressure, anthropometric, psychiatric traits, and others [13–15]. Such approaches are limited in that identified genetic associations cannot be attributed with confidence to any one trait. Importantly, Campos et al. [12] addressed this limitation via an OSA-specific replication analysis. They replicated 29 of the 49 discovered associations in a BMI-adjusted OSA GWAS in 23andMe, which had an OSA prevalence of ~11%. This suggests that the 29 replicated loci are indeed associated with OSA, and not only with snoring. This replication rate is higher than the replication rate reported when using US-based healthcare systems to estimate the genetic association of variants that were reported in OSA-focused studies with substantially smaller sample sizes [16].

The principle of leveraging genetic associations with OSA-related traits to discover OSA-specific genetic associations is useful. It could be extended to excessive daytime sleepiness (EDS), the most common presenting symptom of OSA [17], to insomnia, as we recently found that a polygenic risk score of insomnia is associated with OSA [18], and to other OSA-associated phenotypes. However, it remains important to validate associations with OSA in independent studies, and preferably in studies that correctly classify OSA cases and controls (as much as possible given the variability in OSA indices such as the apnea–hypopnea index [19]).

As shown in Table 1, the misclassification of OSA results in biased genetic effect estimates. The simulated example is simplistic, as it assumes that OSA misclassification does not depend on the genetic variant. In reality, it is expected that misclassification will be more or less severe depending on the mechanism underlying the genetic variant's association with OSA, and how it manifests in other phenotypes. OSA is heterogeneous, and some OSA subtypes manifest in higher daytime sleepiness or other symptoms [20, 21], leading to higher likelihood of diagnosis. The study of Campos et al. may have better captured genetic variants corresponding to OSA subtypes that also manifest in snoring.

Knowledge about the specific OSA consequences associated with the variant can be leveraged, with the development of an appropriate statistical method, to compute unbiased effect size estimates for the variant-OSA association. Figure 2 shows a directed acyclic graph where a genetic variant g is known to be associated with EDS, with an estimated odds ratio $\mathrm{OR}_g^{\mathrm{eds}}$. In a given population, it should be possible to estimate the association of OSA with EDS: $\mathrm{OR}_{\mathrm{osa}}^{\mathrm{eds}}$. Assuming that g is associated with EDS only via its effect on OSA, that is, OSA completely mediates the association of g with EDS, one should be able to "reverse" the standard mediation analysis to estimate $\mathrm{OR}_g^{\mathrm{osa}}$. Whether such an estimate will be more accurate than an estimate of variant–OSA association obtained in a small study with unbiased OSA classification, is a topic that warrants further statistical and empirical research. Nonetheless, obtaining more accurate estimates of OSA effect sizes, that are not biased by OSA misclassification, is important for downstream applications such as Mendelian randomization analysis.

Table 1. Simulation results demonstrating bias and power loss caused by misclassification of OSA cases

π Misclassification rate	Mean estimated β_g (true OSA)	Mean estimated β_g (observed OSA)	Bias of β_g estimates (observed OSA)	Power (true OSA)	Power (observed OSA)
True OSA prevalence 19%					
0.4	0.100	0.090	0.010	0.96	0.76
0.6	0.100	0.086	0.014	0.96	0.57
0.8	0.099	0.081	0.019	0.95	0.29
True OSA prevalence 28%					
0.4	0.100	0.086	0.014	0.99	0.84
0.6	0.099	0.080	0.020	0.98	0.65
0.8	0.101	0.077	0.023	0.98	0.38
True OSA prevalence 39%					
0.4	0.099	0.079	0.021	0.99	0.86
0.6	0.100	0.071	0.029	0.99	0.64
0.8	0.099	0.065	0.035	1.00	0.37

For each combination of parameters determining OSA prevalence and its rate of misclassification, the simulations compare the estimated effect size (log odds ratio) when using the real OSA status and when using the observed OSA status, that suffers from misclassification, as mean estimates across 1000 simulation repetitions. The power is computed as the proportion of simulations in which the p-value of the genetic variant effect estimate was <.05.

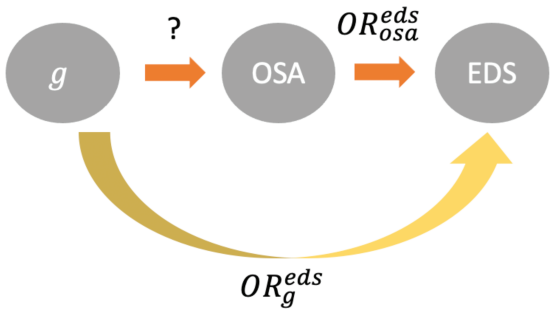

Figure 2. Directed acyclic graph connecting OSA, excessive day time sleepiness, and a genetic variant. The directed acyclic graph presents a potential mediation relationship between a genetic variant, OSA, and excessive daytime sleepiness (EDS). Assuming that the effect of g on EDS is only mediate through OSA, given appropriate methodology one can use the estimated association of g with EDS $\mathrm{OR}_g^{\mathrm{eds}}$and the estimated association of OSA with EDS $\mathrm{OR}_{\mathrm{osa}}^{\mathrm{eds}}$ to estimate $\mathrm{OR}_g^{\mathrm{osa}}$.

Acknowledgments

The author thanks Dr. Daniel J. Gottlieb for reviewing a draft of this manuscript and for providing helpful suggestions.

Funding

Dr. Sofer is supported in part by the National Heart Lung and Blood Institute grant R01HL61012.

Data availability

No data were analyzed in support of this manuscript.

Disclosure statement

Financial disclosure: none.

Non-financial disclosure: none.

References

1. Mills MC, *et al*. A scientometric review of genome-wide association studies. *Commun Biol*. 2019;**2**(1):9. doi:10.1038/s42003-018-0261-x.
2. Lane JM, *et al*. Genetics of circadian rhythms and sleep in human health and disease. *Nat Rev Genet*. 2022;**24**(1):4–20.
3. Cade BE, *et al*. Genetic associations with obstructive sleep apnea traits in hispanic/latino americans. *Am J Respir Crit Care Med*. 2016;**194**(7):886–897. doi:10.1164/rccm.201512-2431OC.
4. Chen H, *et al*. Multi-ethnic meta-analysis identifies RAI1 as a possible obstructive sleep apnea related quantitative trait locus in men. *Am J Respir Cell Mol Biol*. 2017;**58**(3):391–401.
5. Farias Tempaku P, *et al*. Genome-wide association study reveals two novel risk alleles for incident obstructive sleep apnea in the EPISONO cohort. *Sleep Med*. 2020;**66**:24–32. doi:10.1016/j.sleep.2019.08.003.
6. Xu H, *et al*. Genome-wide association study of obstructive sleep apnea and objective sleep-related traits identifies novel risk loci in Han Chinese individuals. *Am J Respir Crit Care Med*. 2022;**2016**(12):1534–1545.
7. Strausz S, *et al*. Genetic analysis of obstructive sleep apnoea discovers a strong association with cardiometabolic health. *Eur Respir J*. 2021;**57**(5):2003091.
8. Kapur V, *et al*. Underdiagnosis of sleep apnea syndrome in U.S. communities. *Sleep Breath*. 2002;**6**(2):49–54.
9. Redline S, *et al*. Sleep-disordered breathing in Hispanic/Latino individuals of diverse backgrounds. The Hispanic Community Health Study/Study of Latinos. *Am J Respir Crit Care Med*. 2014;**189**(3):335–344. doi:10.1164/rccm.201309-1735OC.
10. Benjafield AV, *et al*. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. 2019;**7**(8):687–698. doi:10.1016/S2213-2600(19)30198-5.
11. Campos AI, *et al*. Insights into the aetiology of snoring from observational and genetic investigations in the UK Biobank. *Nat Commun*. 2020;**11**(1):817. doi:10.1038/s41467-020-14625-1.
12. Campos A, *et al*. Discovery of genomic loci associated with sleep apnoea risk through multi-trait GWAS analysis with snoring. *Sleep*. 2023; **46**(3). doi:org/10.1093/sleep/zsac308.
13. Liang J, *et al*. Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations. *PLoS Genet*. 2017;**13**(5):e1006728. doi:10.1371/journal.pgen.1006728.
14. Park H, *et al*. multivariate analysis of anthropometric traits using summary statistics of genome-wide association studies from GIANT consortium. *PLoS One*. 2016;**11**(10):e0163912e0163912. doi:10.1371/journal.pone.0163912.
15. Wu Y, *et al*. Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Transl Psychiatry*. 2020;**10**(1):209. doi:10.1038/s41398-020-00902-6.
16. Veatch OJ, *et al*. Characterization of genetic and phenotypic heterogeneity of obstructive sleep apnea using electronic health records. *BMC Med Genomics*. 2020;**13**(1):105. doi:10.1186/s12920-020-00755-4.
17. Gottlieb DJ, *et al*. Diagnosis and Management of Obstructive Sleep Apnea: A Review. *JAMA*. 2020;**323**(14):1389–1400. doi:10.1001/jama.2020.3514.
18. Zhang Y, *et al*. Genetic determinants of cardiometabolic and pulmonary phenotypes and obstructive sleep apnoea in HCHS/SOL. *eBioMedicine*. 2022;**84**:104288. doi:10.1016/j.ebiom.2022.104288.
19. Lechat B, *et al*. Multinight prevalence, variability, and diagnostic misclassification of obstructive sleep apnea. *Am J Respir Crit Care Med*. 2021;**205**(5):563–569.
20. Allen AJH, *et al*. Symptom subtypes and risk of incident cardiovascular and cerebrovascular disease in a clinic-based obstructive sleep apnea cohort. *J Clin Sleep Med*. 2022;**18**(9):2093–2102. doi:10.5664/jcsm.9986.
21. Keenan BT, *et al*. Recognizable clinical subtypes of obstructive sleep apnea across international sleep centers: a cluster analysis. *Sleep*. 2018;**41**(3):zsx214. doi:10.1093/sleep/zsx214.