Multi-Site Validation of a Simple Electronic Health Record Algorithm for Defining Obstructive Sleep Apnea

Brendan T. Keenan, MS[1,†]; H. Lester Kirchner, PhD[2,†]; Olivia J. Veatch, PhD[1,6]; Kenneth M. Borthwick, MS[2]; Vicki A. Davenport, HSD[2]; John C. Feemster, BA[3]; Maged Gendy, BS[4]; Thomas R. Gossard, BA[3]; Frances M. Pack, RN[1]; Laura Sirikulvadhana, MPH[5]; Luke N. Teigen, BS[3]; Paul C. Timm, MNM[3]; Beth A. Malow, MS, MD[6]; Timothy I. Morgenthaler, MD[3]; Phyllis C. Zee, MD, PhD[4]; Allan I. Pack, MBChB, PhD[1]; Janet D. Robishaw, PhD[7,‡]; Stephen F. Derose, MS, MD[5,‡]

[†]co-lead authors; [‡]joint-senior authors; [1]Division of Sleep Medicine/Department of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [2]Biomedical & Translational Informatics, Geisinger, Danville, PA 17822; [3]Center for Sleep Medicine, Mayo Clinic, Rochester, MN 55905; [4]Center for Circadian and Sleep Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611; [5]Department of Research & Evaluation, Kaiser Permanente Southern California, Pasadena, CA 91107; [6]Sleep Disorders Division, Vanderbilt University Medical Center, Nashville, TN 37232; [7]Charles E Schmidt College of Medicine, Florida Atlantic University, Boca Raton, FL 33431

SUPPLEMENTAL METHODS

Study Populations

We performed clinical chart reviews in patients from each of the six institutions participating in the Sleep Apnea Genetics Study (SAGS), which include: Geisinger, Kaiser Permanente Southern California (KPSC), Mayo Clinic, Northwestern University, University of Pennsylvania, and Vanderbilt University Medical Center. Additional details on the specific populations from which participants were obtained at each site are provided below.

Geisinger

Patients undergoing chart review were randomly selected from the current MyCode biobank participants. MyCode is a major resource for research that combines information obtained from DNA and serum with health information from the electronic health record (Epic) and other sources intended to improve the prevention, diagnosis, and treatment of disease.[1] MyCode participants are consented under an IRB-approved protocol that allows research across a broad range of clinical conditions and permits the sharing of data consistent with the NIH data sharing policy. No specific clinics or practices are targeted by MyCode and there is a high consent rate, suggesting samples are representative of the overall health system.[1] All included patients had available whole exome sequencing data. EHR-defined cases were additionally required to have at least 1 year of activity post their first OSA diagnosis and be between 18-88 years of age at time of the OSA diagnosis. EHR-defined non-cases were required to have at least 2 years of activity in the health system between January 1, 2008 and December 31, 2016 and be between 18-88 years of age as of December 31, 2016.

Kaiser Permanente Southern California

Participants randomly selected for chart reviews were members of the Kaiser Permanente Southern California (KPSC) health system. KPSC is a prepaid, integrated health system with about 4.5 million members of diverse race, ethnicity and socioeconomic status cared for at 15 medical centers and 231 medical offices throughout Southern California. Members have very similar health coverage benefits. All clinical data used in this study were captured by an EHR system (Epic), which was used to identify cases, non-cases, and indeterminate cases among all KPSC members from January 1, 2002 through December 31, 2017. Potential study participants had the following inclusion criteria: (1) cases or non-cases, (2) aged 18-88 years, (3) a genetic sample available through Kaiser Permanente Northern California's Research Program on Genes, Environment and Health (https://divisionofresearch.kaiserpermanente.org/genetics/rpgeh; RPGEH), and (4) one or more years of membership, including membership after January 1, 2007, to assure sufficient EHR data for case or non-case confirmation.

Mayo Clinic

Patients undergoing chart reviews for this validation study were selected from the population in the Mayo Clinic Health System. Data regarding diagnosis of OSA, OSA treatment, polysomnographic information, and demographics was extracted from Mayo Clinic's medical record system (Epic) or the database for the Center for Sleep Medicine at Mayo Clinic. Each individual's data were linked to the corresponding genetic data from the Mayo Clinic eMERGE database. The eMERGE network is a biorepository with EHR data that is collected for large-scale genomic research, including genome-wide association studies (GWAS). Participants at the Mayo Clinic are recruited during routine medical appointments of all kinds, unsolicited volunteers are accepted, and efforts are taken to enhance community engagement.[2] Patients

between ages 18-88 years old, with two or more years of EHR data between January 1, 1990 and April 5, 2018, a recorded BMI, and whole-genome genotyping and/or sequence data or a previously collected DNA sample were chosen for this project. Patients lacking any of these components were excluded from analysis.

Northwestern University

Participants undergoing chart review at Northwestern University were obtained from the NUgene biobank (https://www.cgm.northwestern.edu/cores/nugene/). The biobank contains information from over 14,000 patients seen at Northwestern Medicine or its affiliates at any time. Participants are recruited regardless of age, sex, ethnicity and state of health with no specific clinics or practices targeted. Subjects in the biobank are aged 18 or older, and provide a sample of blood from which DNA is extracted, and also consent to their electronic health record (Epic) being queried and de-identified data obtained for research. For validation of the EHR-based algorithm, data from January 1, 2003 to December 31, 2017 were queried on a subset of biobank subjects aged between 18 and 88 years, both with and without a diagnosis of OSA, and with and without sleep studies, were randomly selected and had their charts reviewed.

University of Pennsylvania

Participants for inclusion in the validation sample at the University of Pennsylvania were selected from the Hospital of the University of Pennsylvania health system population. The health system includes clinical data from the electronic medical record (Epic) on all patients seen at the University Hospital. A subset of the validation sample (8.2%; all were EHR-defined cases) were also consented into the Penn Medicine BioBank (http://www.itmat.upenn.edu/biobank/; PMBB). The PMBB operates under two IRB-approved protocols that coordinate the ethical collection, storage, annotation, and distribution of tissue and peripheral blood samples. Consent

for the PMBB is obtained with a consistent informed consent document which includes permission to use participant data for future research opportunities. Blood and tissue samples (when applicable) are obtained from patients recruited at the University Hospital during clinical visits or from outpatient blood laboratories. Clinical data were obtained through multiple sources, such as electronic medical records for abstraction by trained study personnel, and the Penn Data Store (PDS), Penn Medicine's Clinical Data Warehouse. Cases and non-cases for chart review were randomly selected through the data analytics center (DAC) at the University of Pennsylvania. A waiver of HIPAA authorization was received from the IRB to allow solely the two individuals conducting chart reviews the ability to view protected health information (PHI) for the exclusive purpose of manually reviewing charts in this study; no PHI was stored during this process. Eligible participants were defined as those aged 18-88 years and with at least 2 years of available data in the EHR between January 1, 2008 and September 11, 2017 (the date of data extraction).

Vanderbilt University Medical Center

Chart reviews were conducted using records from individuals included in Vanderbilt's biorepository linked to electronic health records (BioVU). BioVU is a biorepository of DNA extracted from discarded blood collected during routine clinical testing and linked to de-identified medical records in the Synthetic Derivative (SD). BioVU samples are obtained from every clinic that collects blood for routine laboratory tests at Vanderbilt University Medical Center; thus, we expect minimal bias with regard to the clinical aspects of these samples.[3] The SD is a de-identified copy of the main hospital medical record databases created for research purposes. The de-identification of SD records was achieved primarily through the application of a commercial electronic program, which was applied and assessed for acceptable effectiveness in

scrubbing identifiers. The Medical Center Ethics Committee was consulted during the planning phase of the BioVU biorepository and continues to provide oversight. The Vanderbilt IRB have on-going oversight; all patients had the right to receive information about this project through a comprehensive education plan, as well as the right to refuse to participate by opting out of the program. Individuals included in the study were required to have available genome-wide genotyping data and at least two years of activity in the health system between January 1, 2001 and June 6, 2017. EHR-defined cases were additionally required to have the first OSA-related ICD code used on or after March 1, 2005 and be between 18-88 years of age at time of the first code usage. EHR-defined non-cases were required to be between 18-88 years of age as of June 6, 2017.

Table S1. Percent agreement and Kappa coefficients from chart review quality control analysis

Site	Percent Agreement	Kappa ± Standard Error
All Participants	97.5%	0.950 ± 0.090
Geisinger	95.0%	0.902 ± 0.207
Kaiser Permanente Southern California	95.0%	0.900 ± 0.222
Mayo Clinic	100.0%	1.000 ± 0.224
Northwestern University	100.0%	1.000 ± 0.224
University of Pennsylvania	95.0%	0.900 ± 0.222
Vanderbilt University Medical Center	100.0%	1.000 ± 0.224

Table S2. Demographic Characteristics of EHR-defined Cases and Non-cases in the Validation Sample at each Site

Measure	Geisinger			Kaiser Permanente Southern California			Mayo Clinic		
	Non-cases	Cases	p†	Non-cases	Cases	p†	Non-cases	Cases	p†
N	100	120	–	100	120	–	100	120	–
Male, %	27.0%	53.3%	0.0001	46.0%	45.8%	0.980	47.0%	67.5%	0.002
Age, years	53.0±17.6	54.7±12.3	0.423	47.9±15.1	61.6±11.8	<0.0001	66.7±14.3	61.2±14.0	0.004
BMI, kg/m^2	30.4±6.8	39.7±9.1	<0.0001	30.2±7.0	33.0±6.9	0.004	28.8±6.6	33.9±7.7	<0.0001
Race, %			0.778			0.0004			0.008
Caucasian	98.0%	98.3%		72.5%	88.7%		94.0%	100.0%	
A. American	1.0%	1.7%		6.1%	6.1%		1.0%	0.0%	
Asian	1.0%	0.0%		7.1%	5.2%		0.0%	0.0%	
Other	0.0%	0.0%		14.3%	0.0%		5.0%	0.0%	
Hispanic, %	2.0%	0.0%	0.205	33.3%	14.7%	0.001	0.0%	0.9%	>0.999
Measure	**Northwestern University**			**University of Pennsylvania**			**Vanderbilt University Medical Center**		
	Non-cases	Cases	p†	Non-cases	Cases	p†	Non-cases	Cases	p†
N	100	120	–	100	120	–	100	120	–
Male, %	17.0%	38.3%	0.0005	28.0%	37.0%	0.159	48.0%	46.7%	0.844
Age, years	57.1±13.7	57.0±10.3	0.967	37.5±12.5	55.6±12.2	<0.0001	62.1±13.2	55.8±11.6	0.0002
BMI, kg/m^2	28.7±6.9	35.3±9.7	<0.0001	30.0±8.0	38.6±11.0	<0.0001	27.7±6.2	36.9±8.7	<0.0001
Race, %			0.191			<0.0001			0.085
Caucasian	61.5%	68.9%		59.0%	32.5%		87.0%	85.0%	
A. American	23.1%	24.4%		33.7%	66.7%		8.0%	14.2%	
Asian	1.5%	0.0%		3.2%	0.0%		0.0%	0.0%	
Other	13.9%	6.7%		4.2%	0.8%		5.0%	0.8%	
Hispanic, %	7.0%	6.7%	0.922	3.0%	0.0%	0.092	0.0%	0.0%	–

†p-value from T-test and chi-squared or Fisher's exact tests comparing EHR– and EHR+ patients; *Abbreviations:* EHR: electronic health record; BMI: body mass index; A. American: African American.

SUPPLEMENTAL REFERENCES

1. Carey DJ, Fetterolf SN, Davis FD, et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med 2016;18:906-13.
2. Lemke AA, Wu JT, Waudby C, Pulley J, Somkin CP, Trinidad SB. Community engagement in biobanking: Experiences from the eMERGE Network. Genomics, society, and policy 2010;6:50.
3. Crawford DC, Goodloe R, Farber-Eger E, et al. Leveraging Epidemiologic and Clinical Collections for Genomic Studies of Complex Traits. Hum Hered 2015;79:137-46.