Nicotine & Tobacco Research, 2021, 2162–2169
doi:10.1093/ntr/ntab124
Original Investigation
Received December 18, 2020; Editorial Decision June 3, 2021; Accepted June 11, 2021

Original Investigation

Multiethnic Prediction of Nicotine Biomarkers and Association With Nicotine Dependence

Andrew W. Bergen PhD[1,2], Christopher S. McMahan PhD[3], Stephen McGee MS[2], Carolyn M. Ervin PhD[2], Hilary A. Tindle, MD, MPH[4,5], Loïc Le Marchand MD, PhD[6], Sharon E. Murphy PhD[7], Daniel O. Stram PhD[8], Yesha M. Patel MS[8], Sungshim L. Park PhD[6], James W. Baurley PhD[2]

[1]Oregon Research Institute, Eugene, OR, USA; [2]BioRealm, LLC, Walnut, CA, USA; [3]School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, USA; [4]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [5]Veterans Health Administration-Tennessee Valley Healthcare System, Geriatric Research, Education and Clinical Center (GRECC), Nashville, TN, USA; [6]Cancer Epidemiology and University of Hawaii Cancer Center, University of Hawai'i, Honolulu, HI, USA; [7]Biochemistry, Molecular Biology, and Biophysics and Masonic Cancer Center, University of Minnesota, Minneapolis, MN, USA; [8]Department of Preventive Medicine, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Corresponding Author: Andrew W. Bergen, PhD, Oregon Research Institute, 1776 Millrace Drive, Eugene, OR 97403, USA. Telephone: (541) 484-2123; Fax: (541) 484-1108; E-mail: abergen@ori.org

Abstract

Introduction: The nicotine metabolite ratio and nicotine equivalents are measures of metabolism rate and intake. Genome-wide prediction of these nicotine biomarkers in multiethnic samples will enable tobacco-related biomarker, behavioral, and exposure research in studies without measured biomarkers.

Aims and Methods: We screened genetic variants genome-wide using marginal scans and applied statistical learning algorithms on top-ranked genetic variants, age, ethnicity and sex, and, in additional modeling, cigarettes per day (CPD), (in additional modeling) to build prediction models for the urinary nicotine metabolite ratio (uNMR) and creatinine-standardized total nicotine equivalents (TNE) in 2239 current cigarette smokers in five ethnic groups. We predicted these nicotine biomarkers using model ensembles and evaluated external validity using dependence measures in 1864 treatment-seeking smokers in two ethnic groups.

Results: The genomic regions with the most selected and included variants for measured biomarkers were chr19q13.2 (uNMR, without and with CPD) and chr15q25.1 and chr10q25.3 (TNE, without and with CPD). We observed ensemble correlations between measured and predicted biomarker values for the uNMR and TNE without (with CPD) of 0.67 (0.68) and 0.65 (0.72) in the training sample. We observed inconsistency in penalized regression models of TNE (with CPD) with fewer variants at chr15q25.1 selected and included. In treatment-seeking smokers, predicted uNMR (without CPD) was significantly associated with CPD and predicted TNE (without CPD) with CPD, time-to-first-cigarette, and Fagerström total score.

Conclusions: Nicotine metabolites, genome-wide data, and statistical learning approaches developed novel robust predictive models for urinary nicotine biomarkers in multiple ethnic groups. Predicted biomarker associations helped define genetically influenced components of nicotine dependence.

Implications: We demonstrate development of robust models and multiethnic prediction of the uNMR and TNE using statistical and machine learning approaches. Variants included in trained models for nicotine biomarkers include top-ranked variants in multiethnic genome-wide studies of smoking behavior, nicotine metabolites, and related disease. Association of the two predicted nicotine biomarkers with Fagerström Test for Nicotine Dependence items supports models of nicotine biomarkers as predictors of physical dependence and nicotine exposure. Predicted nicotine biomarkers may facilitate tobacco-related disease and treatment research in samples with genomic data and limited nicotine metabolite or tobacco exposure data.

Introduction

Cigarette smoking remains the largest modifiable cause of mortality in the United States, responsible for one-third of cancer and cardiovascular disease and most pulmonary disease mortality.[1] Tobacco control and cessation therapies have reduced smoking prevalence 67% over the last 50 years in the United States; yet, in 2018, there were 34 million adult cigarette smokers, with numerous use disparities by demographic, economic, and health conditions.[1]

Nicotine (NIC) is the tobacco constituent responsible for sustained tobacco use.[2] The nicotine metabolite ratio (NMR, the ratio of *trans*-3′-hydroxycotinine, 3HC, to cotinine, COT) is a biomarker of CYP2A6 metabolic activity. The ratio of these two nicotine metabolites is measured via laboratory analysis of blood, saliva, or urine.[3,4] Total nicotine equivalents (TNE) is a biomarker of nicotine consumption, defined here as the molar sum of the urinary concentrations of total NIC, total COT, total 3HC, and nicotine N-oxide, NNO ("total" refers to the molecule and its glucuronides).[5] In addition to serving as a biomarker of nicotine metabolism and consumption,[6] the NMR is associated with the efficacy of multiple tobacco cessation therapies with potential use for personalizing treatment for tobacco use disorder,[7] while TNE is associated with smoking behaviors and toxicant exposures that may account for some lung cancer risk disparities by race or ethnicity.[8,9]

Predictive genetic modeling of nicotine biomarkers promises to provide genetic signatures supporting disease, mechanistic, and treatment research. Prediction models aggregate genetic information into useful metrics, for example, a genetic score predicting lapse and response to bupropion treatment of tobacco use disorder.[10] Genetic modeling of the NMR is supported by significant twin and locus specific heritability estimates.[11,12] There are no heritability estimates of TNE. Heritability estimates of cigarettes per day (CPD), a less precise measure of consumption, are significant in twin and genome-wide approaches[13–15] but lower than NMR estimates.

Predictive genetic modeling of nicotine metabolism and initial applications have encompassed laboratory studies, research cohorts, and cessation trials; modeling focused first on candidate gene variants and then leveraged variants from genome-wide analyses. Predictive genetic models of CYP2A6-mediated nicotine metabolism have been developed that account for approximately 38% to 62% of NMR variance.[16–18] Herein, we describe the development and internal validation of prediction models of two urinary nicotine biomarkers in current smokers from five ethnic groups[19] followed by prediction and external validation in treatment-seeking smokers from two ethnic groups.[20,21] We relate findings to prior analyses and review prospects for translation. Our genome-wide modeling (variant selection, model training, and prediction) of nicotine biomarkers addresses four current research gaps: (i) multiethnic modeling of a NMR, (ii) modeling the urinary NMR (uNMR) versus the NMR, (iii) including statistical learning approaches in modeling of the uNMR, and (iv) modeling of the TNE in any ancestry using any approach.

Materials and Methods

Ethical Approval

Written informed consent was obtained from all participants. The research described herein received approvals from the Institutional Review Boards of BioRealm, the Oregon Research Institute, the University of Hawaii, and the NIH Joint Addiction, Aging, and Mental Health Data Access Committee.

Participants, Measured Biomarkers, and Nicotine Dependence Measures

We utilized participant data from two multiethnic studies in this secondary data analysis: current smokers from the Multiethnic Cohort (MEC) study, initially assembled in 1993 at the University of Hawaii Cancer Center and Department of Preventive Medicine, University of Southern California, to study diet and cancer; treatment-seeking smokers recruited by the University of Wisconsin Transdisciplinary Tobacco Use Research Center (UW-TTURC), at the Center for Tobacco Research and Intervention, established in 1992 to study nicotine dependence and deliver smoking cessation treatments. MEC and UW-TTURC participants were not compensated for providing biospecimens and data used herein.

We studied MEC current smokers who provided (2004–2006) blood and urine samples and epidemiologic data to enable research on genomics and tobacco exposures.[22] Urinary total and free NIC, COT, and 3HC, and NNO, were measured.[22] We analyzed the natural log-transformed uNMR (defined as the ratio of total 3HC and COT) and the square root-transformed TNE (creatinine-standardized molar sum of total NIC, total COT, total 3HC, and NNO). Selection of variants and training of biomarker models were performed with MEC participant data.

We studied a subset of UW-TTURC smokers recruited and randomized (2000–2010) into three smoking cessation trials,[23–25] who provided baseline demographic and behavioral data and a blood sample for research on genetics and nicotine addiction (dbGaP phs000404.v1.p1).[20] The UW-TTURC dataset included four self-administered nicotine dependence measures: the Fagerström Test of Nicotine Dependence (FTND),[26] the Tobacco Dependence Screener,[27] the Nicotine Dependence Syndrome Scale (NDSS),[28] and the Wisconsin Inventory of Smoking Dependence Motives (WISDM).[29] Prediction of biomarkers and external validation with dependence measures were performed in UW-TTURC participant data.

See Supplementary Material for details on MEC metabolite and genomic data and UW-TTURC demographic, dependence, and genomic data.

Variable Selection Phase

The first phase of our modeling process used a marginal scan to examine each genetic variant through a model of the form:

$$Y_i = X_i^t\beta \; + \; S_{ij}\alpha_0 \; + \; P_i^t\alpha + \varepsilon_i, \tag{1}$$

where Y_i is the biomarker level for the ith individual; X_i is a vector of confounding variables with corresponding regression coefficients β; Sij is the genetic variant with α_0 as the corresponding regression coefficient; P_i is a vector of principal components computed on the genotype design matrix, α is the corresponding vector of regression coefficients, and ε_i is the usual error term. We included age, sex, ethnicity, BMI, and the first 50 principal components of the genotype design matrix as confounding variables. The first 50 principal components of the genotype design matrix is a conservative approach to account for genetic relatedness and ancestry among the study participants; that is, the first 50 principal components explain 72% of principal component variance. Given its utility in prediction models of TNE,[9] we considered CPD ("with CPD") as a candidate predictor of the nicotine biomarkers in a second series of models. The model depicted in (1) was fit for each genetic variant in the MEC participant data with Smokescreen database annotation, and p-values for the test of $H_0 : \alpha_0 = 0$ *versus* $H_1 : \alpha_0 \neq 0$ were computed. This phase was completed by selecting 200 genetic variants based on the smallest p-values to move into the training phase.

Training Phase

The second phase of our modeling process makes use of a suite of penalized regression and machine learning techniques. The selected techniques represent the most common, adopted, and validated techniques in the literature. The set of penalized regression models consisted of the LASSO, elastic net, adaptive LASSO, and the adaptive elastic net.[30–33] The penalty parameters in these techniques were selected to minimize the Bayesian information criterion. In each elastic net model, we considered five settings (ie, 0.20, 0.35, 0.5, 0.65, and 0.8) for the penalty mixing parameter; in each adaptive method, we considered five weighting schemes based on a priori fits. This led to a total of 36 fitted regression models. We also trained three machine learning algorithms: a regression tree,[34] selected for the minimum number of splits and maximum depth of the tree via fivefold cross validation; bagging,[35] selected for the number of trees; and gradient boosting machine,[36,37] selected for step size of each boosting step, maximum depth of tree, minimum sum of instance weight (Hessian) needed in a child, subsample ratio of the training instance, and subsample ratio of columns when constructing each tree via fivefold cross validation. In each model, the predictor variables were the selected genetic variants, age, sex, and ethnicity, with CPD added in an additional set of models.

Prediction Phase

The third phase of our modeling process leveraged the 39 trained models to perform prediction. We formed the following predictions:

$$Y_i^{(j)} = g_j(D_i;\theta_j), \text{ for } j = 1, ..., 39, \tag{2}$$

where $Y_i^{(j)}$ denotes the predicted nicotine biomarker level for the ith subject in the UW-TTURC data, D_i denotes the demographics and genotypes available on the ith subject, $g_j(.;.)$denotes the form of the jth model, and θ_j denotes the set of trained parameters for the jth model. These predictions were used to construct an ensemble-based prediction. Briefly, ensemble methods obtain better predictive performance by aggregating over the predictions of multiple statistical and machine learning algorithms. In our application, as is the common approach, we used the following predictive aggregation:

$$Y_i^* = \frac{1}{39}\sum_{j=1}^{39} Y_i^{(j)} \tag{3}$$

In this analysis, genotypes from selected variants were extracted from African American and White UW-TTURC participants (dbGaP phs000404.v1.p1) and cross-referenced to the Smokescreen database[38] by chromosome and position. Dosages were transformed as needed to count Smokescreen alternate alleles. Modeling analyses used the R programming language.[17] Variants selected in analyses of MEC participant data but not available in UW-TTURC participant data were not used in prediction.

Variant Annotation

Variant annotation (GRCh37/hg19 assembly) used the Ensembl Variant Effect Predictor.[39] Variant-related gene associations with smoking-related phenotypes were from the NHGRI-EBI GWAS catalog.[40]

Measured and Predicted Biomarker Demographic Differences

We estimated significant differences in covariate-adjusted measured biomarkers in African American and White MEC participants and in predicted biomarkers in UW-TTURC participants by sex and by ethnicity.

Predicted Biomarkers and Nicotine Dependence Measures

Predicted uNMR and predicted TNE were individually included in linear regression of each score of the four nicotine dependence measures. Each model was adjusted for age, sex, and ethnicity. Regressions were also performed to evaluate interactions with ethnicity and with sex to evaluate potential moderation by demographics.[9]

Results

There were 2239 MEC participants in five ethnic groups with biomarker and genotype data available for modeling. There were 1864 UW-TTURC participants in two ethnic groups with genome-wide data available for prediction and 1800–1862 participants with nicotine dependence data for validation.[20] Participant age and sex distributions reflect study designs. Ethnicity distributions reflect study designs, recruitment locations and selection of African American and White treatment-seeking smokers for prediction. CPD distributions reflect study design and trial recruitment criteria. See Table 1.

Measured Nicotine Biomarkers

The two biomarkers (without or with CPD) in African American and White MEC participants were significantly positively related to each other in a linear model, adjusting for age, sex, and ethnicity (p-values < .001). We observed statistically significant higher levels of covariate-adjusted uNMR without CPD in female versus male participants (p < .001) but no significant differences between African American and White participants. There were no significant

Table 1. Samples Included in Nicotine Biomarker Modeling and Prediction

Characteristic	MEC	UW-TTURC
Participant *N*	2239	1864
Age,[a] mean (SD)	63.9 (7.2)[b]	43.4 (11.3)[b]
Female *N* (%)	1199 (53.6%)[b]	1090 (58.5%)[b]
Ethnicity *N* (%)		
African American	364 (16.3%)	260 (14.0%)
Native Hawaiian	311 (13.9%)	—
Japanese American	674 (30.1%)	—
Latinos	453 (20.2%)	—
White	437 (19.5%)	1604 (86.0%)
Cigarettes per day	[b]	[b]
1–10	1168 (52.2%)	99 (5.3%)
11–20	870 (38.9%)	988 (53.1%)
21–30	119 (5.3%)	533 (28.6%)
≥31	82 (3.5%)	242 (13.0%)

Ethnicity proportions not tested. MEC = Multiethnic Cohort; UW-TTURC = University of Wisconsin Transdisciplinary Tobacco Use Research Center.

[a]MEC age at biospecimen collection; UW-TTURC age at baseline interview.

[b]$p < .001$.

differences in covariate-adjusted uNMR with CPD by sex or ethnicity. We observed statistically significant higher levels in female participants and lower levels in African American participants of covariate-adjusted TNE without CPD, than in male or White participants, respectively (p-values < .001). We observed statistically significant differences in covariate-adjusted TNE with CPD by sex (p-values < .001) but not by ethnicity. See Table 2 and Supplementary Tables 1A and 1B.

Genome-Wide Variant Selection

The number of variants in all genome-wide analyses in MEC participants was N = 542 732. See Table 3 and Supplementary Tables 2A, 2B, 3A, and 3B for selected variant details.

The genome-wide analysis of measured covariate-adjusted uNMR without CPD identified N = 122 genome-wide significant (p-values < 5E–8) associations at chr19q13.2 and associations (p-values < 6.3E-7) at chr19q13.2 and on N = 11 additional autosomes among the top 200 variants. The most significant marginal result genome-wide was rs56113850 (C allele, β = 0.40, p = 5 × 4E-48), in the fourth intron of *CYP2A6*.

The primary genome-wide analysis of measured covariate-adjusted TNE without CPD identified variant associations (p-values < 2.6E–7) on all autosomes among the top 200 variants. The region with the most variants selected (31 variants) was chr15q25.1, and the most significant marginal result variant in this region was rs2036527 (A allele, β = 0.57, p = 1.4E-5), proximal of *CHRNA5*. The region with the top-ranked variant in the genome-wide analyses of TNE (rs56113850, C allele, β = 0.43, p = 2.6E-7) was chr19q13.2 with 20 variants selected.

Results of genome-wide analysis of the uNMR with CPD were nearly identical to the analysis without CPD, for example, 87% of selected variants in both uNMR analyses were found at chr19q13.2. The genome-wide analysis of TNE with CPD exhibited reduced marginal significance (p-values < 3.4E-6), reduced numbers of variants in the chr15q25.1 and chr19q13.2 regions, and a different region with the most variants selected (chr10q25.3).

Table 2. Measured (MEC) and Predicted (UW-TTURC) Nicotine Biomarkers by Sex and Ethnicity, African American and White

	Female		Male		African American		White	
Biomarker	Mean	(SE)	Mean	(SE)	Mean	(SE)	Mean	(SE)
Measured	*N* = 500		*N* = 301		*N* = 364		*N* = 437	
uNMR[a]	1.48[b]	(0.03)	1.36[b]	(0.04)	1.45	(0.04)	1.40	(0.04)
$uNMR_{CPD}$	1.20	(0.06)	1.09	(0.06)	1.44	(0.07)	1.32	(0.07)
TNE[c]	8.03[b]	(0.12)	7.71[b]	(0.13)	7.27[b]	(0.12)	8.32[b]	(0.11)
TNE_{CPD}	7.43[b]	(0.18)	6.42[b]	(0.18)	6.99	(0.20)	7.42	(0.20)
Predicted	*N* = 1090		*N* = 774		*N* = 260		*N* = 1604	
uNMR	1.46	(0.01)	1.45	(0.01)	1.52[b]	(0.01)	1.45[b]	(0.01)
TNE	8.46[b]	(0.04)	7.91[b]	(0.04)	7.43[b]	(0.06)	8.36[b]	(0.03)

For measured nicotine biomarker values by sex, values are adjusted by age and ethnicity (and CPD, where indicated), and ethnicity strata values are adjusted by age and sex (and CPD, where indicated). For predicted nicotine biomarker values, age, sex, and ethnicity (and CPD, where indicated) were included in the models. CPD = cigarettes per day; MEC = Multiethnic Cohort; TNE = total nicotine equivalents; uNMR = urinary nicotine metabolite ratio; UW-TTURC = University of Wisconsin Transdisciplinary Tobacco Use Research Center.

[a]Natural log transformed, no units.

[b]$p < .001$.

[c]Square root transformed, nmol/mg creatinine.

Model Training, Variants, Covariates, and Associated Genes

See Table 3 and Supplementary Tables 2A, 2B, 3A, and 3B for variant and annotated gene details.

As expected, most variants included in the uNMR models without CPD and associated protein-coding genes (43/63 variants and 10/19 genes) were located on chr19q13.2. Clinical covariates included in 38 trained uNMR models included age (22 models), sex (27 models), and ethnicity (38 models). Several chr19q13.2 SNPs were trained in the two machine learning models reviewed (Supplementary Figures) with rs56113850 included in all models reviewed. In one machine learning method (Supplementary Figure 1), Japanese American ethnicity dichotomized uNMR, with chr19q13.2 variants defining the remaining tree structure.

Training TNE models without CPD resulted in 124 included variants located on all autosomes. Included variants were found most often on chromosomes 1, 8, 11, and 15. The regions with the largest number of included variants were chr15q25.1 (eight variants) and chr19q13.2 (six variants). Clinical covariates included in 38 trained TNE models reviewed included age (1 model), sex (37 models), and ethnicity (38 models). In one machine learning model (Supplementary Figure 2), a chr15q25.1 variant dichotomized TNE, sex dichotomized lower values, and Latino ethnicity and a chr22q13.2 variant trichotomized higher values. Included variants were annotated to 53 protein-coding genes distributed over all autosomes. Thirty-six of 47 annotated protein-coding genes have GWAS catalog associations with smoking-related behaviors, diseases, or traits, and five have associations with kidney function (data not shown).

In uNMR models without and with CPD, most (58 of 63) included variants were identical, and there were only minor differences in the frequency of variant inclusion of trained models. In uNMR models with CPD, CPD was included in 36 of 38 trained models reviewed, and age

Table 3. Variants Selected and Included[a] in Penalized Regression Models, by Chromosome, MEC

	uNMR		$uNMR_{CPD}$		TNE		TNE_{CPD}	
C[b]	Selected	Included	Selected	Included	Selected	Included	Selected	Included
1	4	3/3	4	3/4	25	12/22	24	15/22
2	4	3/4	5	3/5	4	4/4	4	4/4
3	5	4/4	3	3/3	5	5/5	3	3/3
4	0	—/—	0	—/—	10	9/10	13	12/13
5	3	3/3	4	4/4	8	6/7	17	11/14
6	0	—/—	0	—/—	6	6/9	9	9/9
7	1	1/1	1	1/1	5	5/5	9	9/9
8	3	1/1	3	1/1	24	14/23	25	17/23
9	0	—/—	0	—/—	4	4/4	5	4/5
10	1	1/1	1	1/1	16	8/13	16	10/13
11	1	1/1	0	—/—	14	11/14	20	10/19
12	0	—/—	0	—/—	3	3/3	3	3/3
13	1	1/1	1	1/1	2	2/2	6	6/6
14	0	—/—	1	1/1	1	1/1	1	1/1
15	1	1/1	1	1/1	34	9/29	7	4/7
16	1	1/1	1	1/1	1	1/1	3	3/3
17	0	—/—	0	—/—	3	3/3	4	3/4
18	0	—/—	0	—/—	4	4/4	4	4/4
19	175	43/151	174	43/148	20	6/18	15	7/12
20	0	—/—	0	—/—	7	7/7	7	6/6
21	0	—/—	0	—/—	1	1/1	0	—/—
22	0	—/—	0	—/—	3	3/3	5	5/5

CPD = cigarettes per day; MEC = Multiethnic Cohort; TNE = total nicotine equivalents; uNMR = urinary nicotine metabolite ratio.
[a]The number of variants included in trained models/number of variants available.
[b]Chromosome.

and sex were included in three and nine additional models. However, in TNE models with CPD, the number of included variants increased and the frequency of variant inclusion in trained models decreased. In TNE models with CPD, CPD was included in all 38 models reviewed, age and ethnicity and were included in six additional and eight fewer models, respectively.

Training and Internal Validation of Nicotine Biomarker Models

For each of the 39 models, we evaluated the final form of the model via standard model diagnostic techniques, for example, residual plots. From these diagnostics, we discovered no evidence that the assumed forms of the models were invalid. To assess model fit, the correlation between measured and fitted nicotine biomarkers was computed for each model and biomarker in the MEC participant data. The ensemble values of these correlations, r, and variance explained (r^2) for uNMR and TNE without CPD were 0.6695 (0.4482) and 0.6450 (0.4160), and for uNMR and TNE with CPD, 0.6760 (0.4570) and 0.7162 (0.5129), respectively (see Supplementary Table 4 for correlation estimates across all 39 models). For three of four sets of models, these values indicate good fit and do not point to overfitting issues. For the models of TNE with CPD, individual penalized regression model correlations dropped from ~0.73 to ~0.42 as penalty parameters increased (Supplementary Table 4), reflecting the loss of variants correlated with CPD. The similar correlation values across penalized regression models for three of four analyses supports equal weighting for each contributing model in constructing our ensemble-based estimators.

Predicted Biomarkers in the UW-TTURC

Given minimal differences in model and ensemble correlations between the two analyses for the uNMR, and evidence for confounding in penalized regression TNE models with CPD, we focus further reporting on predicted biomarkers modeled without CPD to emphasize the utility of genome-wide models for nicotine biomarkers.

Using the ensemble-based models without CPD generated in the MEC, predictions were obtained for both nicotine biomarkers for all UW-TTURC participants (Table 2). Predicted uNMR and predicted TNE in participants were significantly related to each other (β(SE) = 0.017(.005), $p < .001$). Predicted uNMR was significantly higher in African American than White participants ($p < .001$), but there was no significant difference in predicted uNMR by sex ($p = 0.28$). Predicted TNE was significantly larger in female than male participants and significantly smaller in African American than White participants (p-values < .001).

Predicted uNMR and Nicotine Dependence

Predicted uNMR was positively associated with FTND CPD ($p = .002$), WISDM Automaticity ($p = .049$), and NDSS Tolerance ($p = .022$) (Table 4). In additional analyses, interactions of ethnicity and of sex with predicted uNMR (ethnicity $p = .041$, sex $p = .024$) were observed with NDSS Continuity and of sex with predicted uNMR ($p = .045$) were observed with NDSS Stereotypy (Supplementary Table 5).

Predicted TNE and Nicotine Dependence

Predicted TNE was positively associated with FTND total score ($p = .027$), CPD ($p = .014$), and time-to-first-cigarette ($p = .022$); with WISDM Tolerance ($p = .042$) and NDSS Stereotypy ($p = .003$) (Table 4). In additional analyses, interaction of ethnicity with predicted TNE ($p = .0036$) was observed with NDSS Stereotypy (Supplementary Table 5).

Discussion

Genome-Wide Variant Selection, Model Training, and Explanatory Power

Our analyses describe the first genome-wide selection, training, and prediction (genome-wide modeling) of the uNMR using statistical and machine learning techniques, and the first genome-wide modeling of TNE using any technique, as far as we are aware. These analyses demonstrate internal validity in current smokers and external validity in treatment-seeking smokers with prior genome-wide, biomarker, and nicotine dependence findings. We modeled the two nicotine biomarkers throughout the analysis workflow without and with self-reported CPD coded as in the FTND, as CPD has previously been identified as a significant predictor of TNE and the NMR. Inclusion of CPD in modeling of TNE resulted in selection and inclusion of CPD in all models reviewed, reductions in the significance of variants selected, and reduced numbers of variants included from regions strongly associated with the TNE. Modeling of the uNMR with CPD had very limited effects on the selection and inclusion of variants or on the predictive validity of the models. We concentrate discussion on the results of modeling the two biomarkers without CPD.

As expected from prior genome-wide studies, most variants selected and included in uNMR modeling were from the chr19q13.2 *CYP2A6* region. We previously identified rs56113850 in the MEC as the top-ranked variant for uNMR in all ethnic groups tested,[19] and as a *cis* expression Quantitative Trait Locus (*cis* eQTL) for *CYP2A6*.[41] rs56113850 was top ranked in genome-wide studies of the NMR in smokers of European ancestry.[12,13] However, nearly a third (31%) of variants included in our uNMR models were located in non-chr19q13.2 genomic regions. Four of six non-chr19q13.2 protein-coding genes with variants included in uNMR models have associations with smoking-related behaviors and disease in the GWAS catalog (data not shown), adding to the non-chr19q13.2 genes with variants included in models of nicotine metabolism.[17]

Variant selection and inclusion in TNE modeling was more polygenic than for the uNMR, consistent with our understanding of nicotine pharmacology[6] and dominant nicotine-related loci characterized in genome-wide studies.[42] We previously identified rs2036527 (included in seven penalized regression TNE models), as top ranked in genome-wide studies of CPD and of lung cancer in African Americans.[43,44] This variant was identified as the top-ranked variant in genome-wide studies of blood-based COT and of COT + 3HC levels in European ancestry smokers and a *cis* eQTL for *CHRNA5* and other chr15q25.1 genes.[13] However, among chr15q25.1 variants, only rs55676755 was included in all penalized regression models of TNE. Association of rs55676755 with pulmonary disease and function in multiethnic genome-wide studies[45] supports inclusion of rs55676755 in our multiethnic TNE models.

We and others identified the chr19q13.2 region variant rs12459249 (included in six penalized regression models of TNE) as the top-ranked variant in genome-wide analyses of the laboratory-based NMR in three ethnicities[41] and the blood-based NMR in African American smokers.[46] Among six chr19q13.2 region included variants, only rs56113850 and rs73038469 were included in all penalized regression models. Both variants are *cis* eQTLs for protein-coding and noncoding genes in multiple tissues and *cis* QTLs for methylated cytosine–guanine dinucleotides, supporting possible functional roles in gene regulation.[13] While multiple chr19q13.2 variants were included in models for each biomarker, only rs56113850 was included in models of both biomarkers.

The explanatory power of the models in our uNMR ensemble is comparable to those of the in vivo NMR model ensemble we developed based on *CYP2A6–CYP2B6* and related regulatory gene variants (uNMR r^2 = 0.36–0.77 vs. NMR r^2 = 0.37–0.62).[17] Our analysis goals here were to estimate ensemble values for nicotine biomarker models; for the uNMR ensemble, the r^2 was 0.45. Another genetic model, based on the plasma COT/(NIC + COT) ratio, had a comparable r^2 = 0.52.[16] Twin heritability estimates of the NMR are greater than those of the uNMR,[11] providing another perspective for model comparisons. Estimates of genetic constructs for NMRs [12,13,18,19] involve different study designs, ancestries, and validation procedures, making direct comparisons of explanatory power difficult. Our findings of variants in annotated genes with GWAS catalog associations with kidney function in trained TNE models suggest that our models incorporate the greater mechanistic complexity of a urinary biomarker.

Table 4. Predicted Biomarkers and Nicotine Dependence Measures, UW-TTURC

Dependence		uNMR		TNE	
Measure	N	Coefficient	SE	Coefficient	SE
FTND					
Total	1843	0.129	0.195	0.099[a]	0.045
CPD	1862	0.211[b]	0.068	0.039[a]	0.016
TTFC	1861	−0.008	0.078	0.041[a]	0.018
WISDM					
Automaticity	1800	0.297[a]	0.151	−0.011	0.035
Loss of control	1800	−0.021	0.127	0.033	0.029
Craving	1800	−0.156	0.120	0.025	0.028
Tolerance	1800	0.147	0.127	0.060[a]	0.029
Total PDM	1800	−0.003	1.183	−0.065	0.273
NDSS					
Drive	1809	−0.062	0.096	0.005	0.022
Priority	1820	0.019	0.097	−0.033	0.022
Tolerance	1814	0.239[a]	0.104	0.032	0.024
Continuity	1815	0.033	0.094	0.015	0.022
Stereotypy	1813	−0.012	0.096	0.066[b]	0.022
NDSS-T	1800	0.008	0.086	0.022	0.020

CPD = cigarettes per day; FTND = Fagerström Test of Nicotine Dependence; NDSS = Nicotine Dependence Syndrome Scale; NDSS-T = NDSS Total; TNE = total nicotine equivalents; Total PDM = sum of four Primary Dependence Motives; TTFC = time-to-first-cigarette; uNMR = urinary nicotine metabolite ratio; UW-TTURC = University of Wisconsin Transdisciplinary Tobacco Use Research Center; WISDM = Wisconsin Inventory of Smoking Dependence Motives.

[a]$p < .05$.

[b]$p < .005$.

Biomarkers, Demographics, and Dependence

These are the first analyses to relate predicted uNMR and predicted TNE to each other, to ethnicity and sex, to major FTND items, and to WISDM and NDSS subscales. Predicted uNMR and TNE in treatment-seeking smokers were significantly associated with each other as were measured uNMR and TNE in current smokers.[19] Significant differences for both measured and predicted TNE by ethnicity and by sex were observed in the expected directions for creatinine-standardized TNE.[47]

Prior findings provide support for the associations with nicotine dependence measures we observed using predicted nicotine biomarkers. A systematic review found measured NMRs significantly correlated with CPD in 9 of 15 studies overall and in 3 of 4 using the

measured uNMR.[48] Predictive genetic models of two NMRs have shown significant associations with CPD in ordinal and continuous coding.[11,16,18] Measured TNE (24 hour urine, molar sum of NIC, COT, 3HC, and glucuronides, unadjusted for creatinine) was significantly associated with CPD, time-to-first-cigarette, and total FTND score in current smokers.[49]

The associations of predicted nicotine biomarkers with components of the WISDM and NDSS measures we observed are novel. However, prior associations of smoking constructs provide support for the observed associations. For example, WISDM Automaticity and Tolerance and NDSS Stereotypy and Tolerance correlations with the FTND and CPD were among the largest correlations of 13 WISDM and 5 NDSS subscales tested in treatment-seeking smokers from 2 UW-TTURC cessation trials.[21] NDSS Stereotypy and Tolerance were significantly correlated with multiple physical dependence variables in daily smokers recruited for laboratory studies of smoking cessation medications.[28]

Strengths and Limitations

Use of a MEC for modeling nicotine biomarkers will support translation to studies of smokers of multiple ethnicities in behavioral, disease, and treatment research. Further research is needed to assess performance of multiethnic models in specific ethnic populations.

Our uNMR genome-wide variant selection and model training included multiple variants at and outside the chr19q13.2 region. Selection and training of models predicting the uNMR in larger samples may clarify the role of non-chr19q13.2 genes in nicotine metabolism and clearance. Genome-wide modeling and comparison of NMR and uNMR models may provide clues to differences in model explanatory power[17] and reduced correlation between measured blood NMR and uNMR.[50]

Our TNE genome-wide variant selection and model training included top-ranked variants at chr15q25.1 and chr19q13.2 identified in recent genome-wide studies of smoking behaviors, nicotine metabolites, and related disease. Research in additional cohorts with diverse smoking behavior and measured metabolite data may elucidate how behavior, metabolite source, measurement, and standardization influence model development and power.

Conclusions

Concordances observed between our nicotine biomarker modeling and recent genome-wide studies support our goal of developing robust genome-wide prediction models for nicotine biomarkers. Meta-analysis of larger and more diverse samples with respect to participant ancestries, behaviors, biomarkers, and clinical data will improve the predictive power of models and enable out-of-sample model validations. The associations we observed between predicted urinary biomarkers and measures of dependence are supported by prior analyses of biomarkers, dependence measures, and models of predicted NMR with similar measures. Availability of smoking cessation trial data will provide an opportunity to characterize relations between genetically determined components of dependence and cessation outcomes and assess translational relevance.

Supplementary Material

A Contributorship Form detailing each author's specific involvement with this content, as well as any supplementary data, are available online at https://academic.oup.com/ntr.

Funding

This work was supported by the National Institute on Alcohol Abuse and Alcoholism (R44 AA027675 to AWB, CSM, SM, CME, SLP, and JWB) and by the National Cancer Institute (R01 CA232516 to HAT; U01 CA164973 and P01 CA138338 to LLM, DOS, SEM, YMP, and SLP). The sponsors had no role in the analysis of data, writing of the report, or in the decision to submit the paper for publication.

Declaration of Interests

AWB is an employee of Oregon Research Institute and Oregon Community and Evaluation Services and serves as a Scientific Advisor and Consultant to BioRealm, LLC. CME is a co-owner and the Principal Biostatistician for BioRealm, LLC. HAT has served as PI on NIH-supported studies for smoking cessation in which the medication was donated by the manufacturer (eg, Pfizer, varenicline). SM, LLM, DOS, SEM, YMP, and SLP have no conflicts of interest to report. JWB is an employee and an owner of BioRealm, LLC. JWB, CSM, and AWB are coinventors on a related patent application "Biosignature Discovery for Substance Use Disorder Using Statistical Learning," assigned to BioRealm, LLC. BioRealm, LLC offers services related to the Smokescreen Genotyping Array and analysis of nicotine biomarkers.

Acknowledgments

The authors thank participants of the Multiethnic Cohort study and of the University of Wisconsin cessation trials.

The authors acknowledge the contribution of data from Genetic Architecture of Smoking and Smoking Cessation accessed through dbGAP (phs000404.v1.p1). Support for genotyping, which was performed at the Center for Inherited Disease Research (CIDR), was provided by 1 X01 HG005274-01. CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. Assistance with genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies (GENEVA) Coordinating Center (U01 HG004446). Funding support for collection of datasets and samples was provided by the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392) and the University of Wisconsin Transdisciplinary Tobacco Use Research Center (P50 DA019706, P50 CA084724).

References

1. U.S. Department of Health and Human Services. *Smoking Cessation: A Report of the Surgeon General*. Rockville, MD: U.S. Department of Health and Human Services; 2020. https://www.hhs.gov/sites/default/files/2020-cessation-sgr-full-report.pdf. Accessed February 16, 2020.
2. Benowitz NL, Henningfield JE. Establishing a nicotine threshold for addiction. The implications for tobacco regulation. *N Engl J Med*. 1994;331(2):123–125.
3. Benowitz NL, Pomerleau OF, Pomerleau CS, Jacob P 3rd. Nicotine metabolite ratio as a predictor of cigarette consumption. *Nicotine Tob Res*. 2003;5(5):621–624.
4. Dempsey D, Tutka P, Jacob P 3rd, et al. Nicotine metabolite ratio as an index of cytochrome P450 2A6 metabolic activity. *Clin Pharmacol Ther*. 2004;76(1):64–72.
5. Benowitz NL, St Helen G, Nardone N, Cox LS, Jacob P. Urine metabolites for estimating daily intake of nicotine from cigarette smoking. *Nicotine Tob Res*. 2020;22(2):288–292. doi:10.1093/ntr/ntz034.
6. Benowitz NL, Hukkanen J, Jacob P 3rd. Nicotine chemistry, metabolism, kinetics and biomarkers. *Handb Exp Pharmacol*. 2009;(192):29–60. doi:10.1007/978-3-540-69248-5_2.
7. Lerman C, Schnoll RA, Hawk LW Jr, et al.; PGRN-PNAT Research Group. Use of the nicotine metabolite ratio as a genetically informed biomarker of response to nicotine patch or varenicline for smoking cessation: a randomised, double-blind placebo-controlled trial. *Lancet Respir Med*. 2015;3(2):131–138.

8. St Helen G, Benowitz NL, Ko J, et al. Differences in exposure to toxic and/or carcinogenic volatile organic compounds between Black and White cigarette smokers. *J Expo Sci Environ Epidemiol*. 2021;31(2):211–223.
9. Stram DO, Park SL, Haiman CA, et al. Racial/ethnic differences in lung cancer incidence in the Multiethnic Cohort study: an update. *J Natl Cancer Inst*. 2019;111(8):811–819.
10. David SP, Strong DR, Leventhal AM, et al. Influence of a dopamine pathway additive genetic efficacy score on smoking cessation: results from two randomized clinical trials of bupropion. *Addiction*. 2013;108(12):2202–11. doi:10.1111/add.12325.
11. Swan GE, Lessov-Schlaggar CN, Bergen AW, He Y, Tyndale RF, Benowitz NL. Genetic and environmental influences on the ratio of 3′hydroxycotinine to cotinine in plasma and urine. *Pharmacogenet Genomics*. 2009;19(5):388–398.
12. Loukola A, Buchwald J, Gupta R, et al. A genome-wide association study of a biomarker of nicotine metabolism. *PLoS Genet*. 2015;11(9):e1005498. doi:10.1371/journal.pgen.1005498.
13. Buchwald J, Chenoweth MJ, Palviainen T, et al. Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent [published online ahead of print March 10, 2020]. *Mol Psychiatry*. doi:10.1038/s41380-020-0702-z.
14. Liu M, Jiang Y, Wedow R, et al.; 23andMe Research Team, HUNT All-In Psychiatry. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. 2019;51(2):237–244.
15. Xu K, Li B, McGinnis KA, et al. Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. *Nat Commun*. 2020;11(1):5302. doi:10.1038/s41467-020-18489-3.
16. Bloom J, Hinrichs AL, Wang JC, et al. The contribution of common CYP2A6 alleles to variation in nicotine metabolism among European-Americans. *Pharmacogenet Genomics*. 2011;21(7):403–416.
17. Baurley JW, McMahan CS, Ervin CM, Pardamean B, Bergen AW. Biosignature discovery for substance use disorders using statistical learning. *Trends Mol Med*. 2018;24(2):221–235.
18. El-Boraie A, Taghavi T, Chenoweth MJ, et al. Evaluation of a weighted genetic risk score for the prediction of biomarkers of CYP2A6 activity. *Addict Biol*. 2020;25(1):e12741.
19. Patel YM, Park SL, Han Y, et al. Novel association of genetic markers affecting CYP2A6 activity and lung cancer risk. *Cancer Res*. 2016;76(19):5768–5776.
20. Bierut LJ, Baker T, Breslau N, et al. *The Genetic Architecture of Smoking and Smoking Cessation. dbGaP Genotypes and Phenotypes*; 2011. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000404.v1.p1. Accessed April 5, 2019.
21. Piper ME, McCarthy DE, Bolt DM, et al. Assessing dimensions of nicotine dependence: an evaluation of the Nicotine Dependence Syndrome Scale (NDSS) and the Wisconsin Inventory of Smoking Dependence Motives (WISDM). *Nicotine Tob Res*. 2008;10(6):1009–1020.
22. Murphy SE, Park SS, Thompson EF, et al. Nicotine N-glucuronidation relative to N-oxidation and C-oxidation and UGT2B10 genotype in five ethnic/racial groups. *Carcinogenesis*. 2014;35(11):2526–2533.
23. Piper ME, Federman EB, McCarthy DE, et al. Efficacy of bupropion alone and in combination with nicotine gum. *Nicotine Tob Res*. 2007;9(9):947–954.
24. McCarthy DE, Piasecki TM, Lawrence DL, et al. A randomized controlled clinical trial of bupropion SR and individual smoking cessation counseling. *Nicotine Tob Res*. 2008;10(4):717–729.
25. Piper ME, Smith SS, Schlam TR, et al. A randomized placebo-controlled clinical trial of 5 smoking cessation pharmacotherapies. *Arch Gen Psychiatry*. 2009;66(11):1253–1262.
26. Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom K-O. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Addiction*. 1991;86(9):1119–1127. doi:10.1111/j.1360-0443.1991.tb01879.x.
27. Kawakami N, Takatsuka N, Inaba S, Shimizu H. Development of a screening questionnaire for tobacco/nicotine dependence according to ICD-10, DSM-III-R, and DSM-IV. *Addict Behav*. 1999;24(2):155–166.
28. Shiffman S, Waters A, Hickcox M. The Nicotine Dependence Syndrome Scale: a multidimensional measure of nicotine dependence. *Nicotine Tob Res*. 2004;6(2):327–348. doi:10.1080/1462220042000202481.
29. Piper ME, Piasecki TM, Federman EB, et al. A multiple motives approach to tobacco dependence: the Wisconsin Inventory of Smoking Dependence Motives (WISDM-68). *J Consult Clin Psychol*. 2004;72(2):139–154.
30. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58(1):267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.
31. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301–320. doi:10.1111/j.1467-9868.2005.00503.x.
32. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418–1429. doi:10.1198/016214506000000735.
33. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat*. 2009;37(4):1733–1751.
34. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med*. 2003;26(3):172–181.
35. Friedman JH, Hall P. On bagging and nonlinear estimation. *J Stat Plan Inference*. 2007;137(3):669–683. doi:10.1016/j.jspi.2006.06.002.
36. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot*. 2013;7:21. doi:10.3389/fnbot.2013.00021.
37. Chen T, He T, Benesty M, Khotilovich V, Tang Y. *Xgboost: Extreme Gradient Boosting. R Package Version 0 4-2*; 2015:1–4. http://cran.fhcrc.org/web/packages/xgboost/vignettes/xgboost.pdf. Accessed April 22, 2021.
38. Baurley JW, Edlund CK, Pardamean CI, Conti DV, Bergen AW. Smokescreen: a targeted genotyping array for addiction research. *BMC Genomics*. 2016;17:145. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2495–7
39. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122.
40. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005–D1012.
41. Baurley JW, Edlund CK, Pardamean CI, et al. Genome-wide association of the laboratory-based nicotine metabolite ratio in three ancestries. *Nicotine Tob Res*. 2016;18(9):1837–1844.
42. Bierut LJ, Tyndale RF. Preparing the way: exploiting genomic medicine to stop smoking. *Trends Mol Med*. 2018;24(2):187–196.
43. David SP, Hamidovic A, Chen GK, et al. Genome-wide meta-analyses of smoking behaviors in African Americans. *Transl Psychiatry*. 2012;2:e119. https://www.nature.com/articles/tp201241
44. Zanetti KA, Wang Z, Aldrich M, et al. Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African-American population. *Lung Cancer*. 2016;98:33–42. https://pubmed.ncbi.nlm.nih.gov/27393504/
45. Sakornsakolpat P, Prokopenko D, Lamontagne M, et al.; SpiroMeta Consortium, International COPD Genetics Consortium. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet*. 2019;51(3):494–505.
46. Chenoweth MJ, Ware JJ, Zhu AZX, et al.; PGRN-PNAT Research Group. Genome-wide association study of a nicotine metabolism biomarker in African American smokers: impact of chromosome 19 genetic influences. *Addiction*. 2018;113(3):509–523.
47. Carroll DM, Cigan S, Ikuemonisan J, et al. Relationships between race, gender, and spot urine levels of biomarkers of tobacco exposure vary based on how creatinine is handled in analyses. *Nicotine Tob Res*. 2020;22(11):2109–2113.
48. West O, Hajek P, McRobbie H. Systematic review of the relationship between the 3-hydroxycotinine/cotinine ratio and cigarette dependence. *Psychopharmacology (Berl)*. 2011;218(2):313–322.
49. Muhammad-Kah RS, Hayden AD, Liang Q, Frost-Pineda K, Sarkar M. The relationship between nicotine dependence scores and biomarkers of exposure in adult cigarette smokers. *Regul Toxicol Pharmacol*. 2011;60(1):79–83.
50. St Helen G, Novalen M, Heitjan DF, et al. Reproducibility of the nicotine metabolite ratio in cigarette smokers. *Cancer Epidemiol Biomarkers Prev*. 2012;21(7):1105–1114.