

RESEARCH

Open Access



Binary classification of users of electronic cigarettes and smokeless tobacco through biomarkers to assess similarity with current and former smokers: machine learning applied to the population assessment of tobacco and health study

Hiromi Ohara^{1*}, Shigeaki Ito¹ and Yuichiro Takanami¹

Abstract

Background Exposure to harmful and potentially harmful constituents in cigarette smoke is a risk factor for cardiovascular and respiratory diseases. Tobacco products that could reduce exposure to these constituents have been developed. However, the long-term effects of their use on health remain unclear. The Population Assessment of Tobacco and Health (PATH) study is a population-based study examining the health effects of smoking and cigarette smoking habits in the U.S. population. Participants include users of tobacco products, including electronic cigarettes and smokeless tobacco. In this study, we attempted to evaluate the population-wide effects of these products, using machine learning techniques and data from the PATH study.

Methods Biomarkers of exposure (BoE) and potential harm (BoPH) in cigarette smokers and former smokers in wave 1 of PATH were used to create binary classification machine-learning models that classified participants as either current (BoE: N = 102, BoPH: N = 428) or former smokers (BoE: N = 102, BoPH: N = 428). Data on the BoE and BoPH of users of electronic cigarettes (BoE: N = 210, BoPH: N = 258) and smokeless tobacco (BoE: N = 206, BoPH: N = 242) were input into the models to investigate whether these product users were classified as current or former smokers. The disease status of individuals classified as either current or former smokers was investigated.

Results The classification models for BoE and BoPH both had high model accuracy. More than 60% of participants who used either one of electronic cigarettes or smokeless tobacco were classified as former smokers in the classification model for BoE. Fewer than 15% of current smokers and dual users were classified as former smokers. A similar trend was found in the classification model for BoPH. Compared with those classified as former smokers,

*Correspondence:

Hiromi Ohara
hiromi.ohara@jt.com

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

a higher percentage of those classified as current smokers had cardiovascular disease (9.9–10.9% vs. 6.3–6.4%) and respiratory diseases (19.4–22.2% vs. 14.2–16.7%).

Conclusions Users of electronic cigarettes or smokeless tobacco are likely to be similar to former smokers in their biomarkers of exposure and potential harm. This suggests that using these products helps to reduce exposure to the harmful constituents of cigarettes, and they are potentially less harmful than conventional cigarettes.

Keywords Population study, Smoking, E-cigarette, Oral tobacco, Supervised learning, Biomarker of exposure, Biomarker of potential harm.

Background

Cigarette smoking has been reported to increase the risk of several diseases, including chronic obstructive pulmonary disease (COPD) and cardiovascular disease (CVD) [1]. These diseases have been suggested to be caused by the harmful constituents of cigarette smoke [2, 3]. Tobacco companies have been developing products that may reduce disease risks because they emit fewer harmful and potentially harmful constituents [4]. Cigarette smokers switching to these products showed reductions in their biomarkers of exposure (BoE), which are derived from constituents in tobacco smoke [5–13]. Biomarkers of potential harm (BoPH) (e.g., oxidative stress, inflammation, lipid metabolism, and platelet activation/coagulation) have also been reported to be closer to those of non-smokers or former smokers in those who switch from conventional cigarettes to tobacco heating systems [14, 15], and nonconventional vapor products [16]. However, further studies are needed to verify the risk reduction achieved by switching to these products, including with a sufficiently large sample and good background information.

The Population Assessment of Tobacco and Health (PATH) study is a joint project of the Food and Drug Administration (FDA) and the National Institutes of Health (NIH), and is one of the largest studies that track tobacco product use and health effects over time [17]. In the PATH study, questionnaire data on smoking status and health status, biomarkers of exposure and potential harm, and other background information were obtained from participants and registered with the Inter-University Consortium for Political and Social Research (ICPSR). The PATH study includes dual and exclusive users of potentially lower-risk products (e.g., electronic cigarettes and smokeless tobacco) as well as current and former smokers and non-smokers. The study data will therefore allow for a more extensive analysis of tobacco smoke exposure and the biological effects of using potentially lower-risk products. Several reports have compared biomarkers among smokers and users of electronic cigarettes and smokeless tobacco products using the PATH study data [18–20], making statistical group comparisons for each biomarker. These statistical methods may yield different comparisons and different results for some

biomarkers within the same report: for example, the concentrations of nicotine metabolites and tobacco-specific nitrosamines were higher in users of smokeless products than in current smokers, while those of polycyclic aromatic hydrocarbons and volatile organic compounds were lower [18].

Machine learning optimizes technologically advanced computational power and statistical tools for the analysis of big data, with advanced algorithms capable of handling multicollinearity, nonlinearity, and higher-order interactions among variables [21]. In general, supervised learning techniques are used in classification models. The model is trained with labeled data to properly classify non-labeled data. Where binary classification is used, the model shows whether the new observations are closer to one or another option based on the features of the training dataset. Previous studies reported that users of electronic cigarettes and smokeless tobacco could have levels of several biomarkers that are closer to those found in non-smokers or former smokers than current smokers. However, these studies evaluated the biomarkers individually, and not comprehensively. Machine learning can evaluate the difference between participants by integrating all tested biomarkers. This study aimed to develop two classification models to discriminate between current and former smokers based on their biomarkers of exposure and potential harm. The classification models that had learned the features of current and former smokers were then used to evaluate markers from users of electronic cigarettes and smokeless tobacco products to see if they were similar to either current or former smokers. It is also possible to evaluate the differences in disease prevalence in individuals, including users of electronic cigarettes and smokeless tobacco products, between those classified as current and former smokers. This may suggest whether electronic cigarettes and smokeless tobacco products genuinely pose a potentially lower risk to health.

Methods

Study design We used data for adults (18 to 90 years old) from the public-use files (ICPSR (a), ICPSR 36,498) and restricted-use files (ICPSR (b), ICPSR 36,231) from Wave 1 (September 12, 2013 to December 14, 2014) of the PATH

Study. To protect the rights, welfare, and well-being of all human participants in this study, the Westat Institutional Review Board approved the study design and protocol, and the Office of Management and Budget approved data collection. The detailed study design and data collection of the PATH study have previously been published [17].

Participants Groups were defined using self-reports in the PATH study questionnaire. Participants who reported smoking cigarettes daily but not using any other tobacco products were defined as exclusive cigarette smokers (CS group). Exclusive users of electronic cigarettes and smokeless tobacco products were defined similarly (EPRODS and SMKLS groups). Participants who reported using both conventional and electronic cigarettes, or conventional cigarettes and smokeless tobacco products, were defined as dual users (dual-EPRODS and dual-SMKLS groups). People who reported that they had quit smoking and did not use any tobacco products were defined as former smokers (ExSM group). We also extracted information on age, gender, whether they drank alcohol, lived in urban or rural areas, and health status (cardiovascular disease; respiratory disease; high blood pressure, high cholesterol, and diabetes) from the questionnaire data. For health status, those who answered “Yes” to the question “Has a doctor, nurse or other health professional told you that you have COPD?” were considered to have COPD. People who answered “Yes” to questions on any of COPD, chronic bronchitis, emphysema, or other pulmonary or respiratory diseases were considered to have a respiratory disease. Anyone who answered “Yes” to questions about congestive heart failure, stroke, heart attack, or some other heart condition were considered to have CVD. A total of 2707 participants were included in the analysis for biomarkers of exposure, and 3466 in the analysis for biomarkers of potential harm.

Biomarkers For the biomarkers of exposure included in the restricted data collected by the PATH study, we selected those that corresponded to harmful and potentially harmful constituents listed by the FDA [4] and evaluated in previous tobacco evaluation studies [6, 10]. These included 4-(methylnitrosamino)-4-(3-pyridyl)-1-butanol (NNAL), N'-nitrosonornicotine (NNNT), total nicotine equivalents (TNE7), N-acetyl-S-(2-carboxyethyl)-L-cysteine (CEMA), N-acetyl-S-(2-hydroxyethyl)-L-cysteine (HEMA), N-acetyl-S-(3-hydroxypropyl)-L-cysteine (HPMA), N-acetyl-S-(phenyl)-L-cysteine (PMA), 1-naphthol or 1-hydroxynaphthalene (P01), 2-naphthol or 2-hydroxynaphthalene (P02), and 1-hydroxypyrene (P10). For the biomarkers of potential harm, we used 8-isoprostane (8PGFT), high-sensitivity C-reactive protein (hsCRP), interleukin 6 (IL6), soluble intercellular adhesion module (sICAM), and fibrinogen (FIB). The creatinine

correction was used for each participant for biomarkers in urine samples to adjust for daily excretion.

Machine learning To predict the exposed constituent and biological effects of tobacco smoke from the use of each tobacco product, we created models with features consisting of biomarkers of either exposure (for exposure [BoE] classification model) or potential harm (for potential harm [BoPH] classification model). We divided data from current and former smokers into training datasets (80%) and test datasets (20%). These test datasets were evaluated as references for users of electronic cigarettes and smokeless tobacco products. The total number of participants and the number used for machine training are shown in Tables 1 and 2. Using 80% of the training data, the two biomarker classification models were created by machine learning, and used to classify participants into current or former smokers. A random forest model (five-fold, 100-repeated cross-validation) was used for the algorithm. The cross-validation accuracy of the models obtained in the machine learning process was evaluated by the receiver operating characteristic curve-area under curve (ROC-AUC). We input data from the two test datasets on current and former smokers, and also on dual and single users of electronic cigarettes and smokeless tobacco products. The percentage of data classified as current and former smokers was tabulated for each group. There were more current than former smokers and the current smokers were therefore randomly sampled by random number generation, to align the number of participants in each group, because using imbalanced data leads to poorer model performance [22]. The “feature importance” was calculated as the percentage of each feature that influenced the classification into current or former smokers in the cross-validation.

To assess the prevalence of disease in participants classified as current or former smokers, we calculated the percentage of those labelled with CVD or respiratory diseases among those classified as current or former smokers by the classification model, using data from the questionnaire.

Data analysis, including machine learning, was performed in R software version 4.2.1 using the caret package [23].

Results

Demographic information and biomarkers The demographic information and biomarker characteristics of all the participants in this study and those randomly selected for machine learning for the two classification models are shown in Tables 1 and 2. In accordance with ICPSR data release rules, tables with cell sizes smaller than the threshold for the specific dataset will not be released.

Table 1 Demographic information and geometric mean of biomarkers for the exposure biomarkers classification model

		Training data			Test data		Products			
		Full CS ¹	CS ²	ExSM	test-CS ²	test-ExSM	dual EPRODS	EPRODS	dual SMKLS	SMKLS
Participants (N)		2189	81	81	21	21	68	142	37	169
Age	Mean	39.1	38.8	30.1	38.7	31.1	36.0	38.1	32.1	42.0
	SD	15.0	14.6	14.8	13.8	14.2	14.2	14.2	15.5	17.0
Gender (N)	Male	1080	35	56	8	12	42	61	-	159
	Female	1109	46	25	13	9	26	81	-	10
Ethnicity (N)	White	1681	65	43	-	-	-	127	-	148
	Black	299	8	27	-	-	-	5	-	6
	Other	209	8	11	-	-	-	10	-	15
Alcohol (N) ³		1980	76	70	19	19	61	134	32	155
Urban (N) ³		1957	75	77	18	20	62	134	23	136
High blood pressure (N) ³		547	23	14	-	-	15	23	9	56
High cholesterol (N) ³		386	12	15	0	-	9	27	-	41
Diabetes (N) ³		289	14	6	-	-	6	11	-	23
Cardiovascular disease (N) ⁴		230	10	5	-	-	-	10	-	13
Respiratory disease (N) ⁴		551	24	11	6	-	10	29	5	28
CEMA (ug/g*CRE)	Mean	368.2	361.0	131.0	278.5	96.0	287.9	169.4	261.1	126.8
	SD	281.0	272.2	97.7	132.3	44.9	186.2	203.7	162.2	93.7
HEMA (ug/g*CRE)	Mean	5.2	5.6	1.7	4.3	1.2	3.9	1.8	3.0	1.2
	SD	6.6	9.9	1.7	3.2	0.8	5.3	1.6	2.6	1.1
HPMA (ug/g*CRE)	Mean	1708.8	1646.3	364.5	1378.5	251.6	1412.2	564.2	1075.3	406.7
	SD	1452.9	1130.8	430.9	1141.4	190.4	1039.8	660.5	740.5	674.5
NNAL (ng/g*CRE)	Mean	410.4	366.8	94.7	355.4	69.7	290.6	76.1	912.5	1480.8
	SD	365.9	247.7	266.6	289.3	159.5	209.8	186.6	1044.4	3142.6
NNNT (ng/g*CRE)	Mean	25.7	24.0	5.6	14.0	3.5	19.2	9.9	40.1	53.1
	SD	78.9	69.7	14.8	12.4	7.6	22.1	16.8	40.3	87.0
P01 (ug/g*CRE)	Mean	86.1	14.5	3.1	12.3	2.1	62.6	44.3	10.3	4.4
	SD	768.7	8.7	4.1	8.0	1.9	375.1	349.9	7.1	20.7
P02 (ug/g*CRE)	Mean	17.2	16.9	7.0	15.5	4.6	15.3	11.5	12.5	8.3
	SD	10.9	7.6	5.7	8.5	2.3	8.1	35.7	5.2	13.4
P10 (ng/g*CRE)	Mean	471.1	386.9	202.0	344.4	354.1	401.8	234.8	399.8	245.4
	SD	1420.5	232.9	154.1	167.6	611.3	410.2	196.2	396.4	236.0
PMA (ng/g*CRE)	Mean	1311.2	1150.4	947.0	1681.7	846.2	1268.6	1543.7	1151.0	1226.2
	SD	1448.9	732.3	817.9	1797.6	510.9	1294.2	1792.3	1028.2	1081.3
TNE7 (ug/g*CRE)	Mean	77.9	73.2	12.1	67.7	6.0	80.2	64.5	98.6	119.4
	SD	51.7	41.8	32.2	37.5	10.6	53.2	53.5	66.8	100.5

Abbreviations: CEMA N-acetyl-S-(2-carboxyethyl)-L-cysteine, CS cigarette smoker, CRE creatinine, dual-EPRODS user of both conventional and electronic cigarettes, dual-SMKLS user of both conventional cigarettes and smokeless tobacco, ExSM former smoker, EPRODS electronic cigarettes user, test-CS test dataset of current smokers, test-ExSM test dataset of former smokers, HEMA N-acetyl-S-(2-hydroxyethyl)-L-cysteine, HPMA N-acetyl-S-(3-hydroxypropyl)-L-cysteine, NNAL 4-(methylnitrosamino)-4-(3-pyridyl)-1-butanol, NNNT N'-nitrosonornicotine, SD standard deviation, SMKLS user of smokeless tobacco products, TNE7 total nicotine equivalents, PMA N-acetyl-S-(phenyl)-L-cysteine, P01 1-naphthol or 1-hydroxynaphthalene, P02 2-naphthol or 2-hydroxynaphthalene, P10 1-hydroxypyrene

¹Full data of current smokers

²Random selection from full data of current smokers to match the number of former smokers for machine learning

³Number of participants who answered "Yes" in the questionnaire

⁴Number of participants who answered "Yes" in the relevant disease questionnaire

- In accordance with ICPSR data release rules, tables with cell sizes smaller than the threshold for the specific dataset will not be released

Machine-learning models The percentage of users of each tobacco product classified as current or former smokers by (a) the biomarkers of exposure model and (b) the biomarkers of potential harm model, and the order of importance of the features are shown in Figs. 1 and 2. The ROC-AUC, a cross-validation model performance score, was 95.0% for the exposure biomarkers classification

model and 82.0% for potential harm biomarkers model. For both the classification models, the scores for both test datasets were more than 75%. A higher percentage of users of either electronic cigarettes or smokeless tobacco products were classified as former than current smokers. Both the dual user groups (dual-EPRODS and dual-SMKLS) had a higher percentage classified as cur-

Table 2 Demographic information and geometric mean of biomarkers for the potential harm classification model

		Training data			Test data		Products			
		Full CS ¹	CS ²	ExSM	test-CS ²	test-ExSM	dual EPRODS	EPRODS	dual SMKLS	SMKLS
Participants (N)		2538	343	343	85	85	82	176	44	198
Age	Mean	39.5	38.8	29.5	37.1	30.0	39.0	39.2	33.4	42.2
	SD	15.0	14.5	13.7	14.0	13.5	15.2	14.9	16.1	16.8
Gender (N)	Male	1217	166	205	40	45	47	74	-	187
	Female	1321	177	138	45	40	35	102	-	11
Ethnicity (N)	White	1975	277	236	62	65	71	154	-	172
	Black	329	40	53	13	12	4	11	-	11
	Other	234	26	54	10	8	7	11	-	15
Alcohol (N) ³		2280	306	316	78	82	73	165	38	183
Urban (N) ³		2263	306	336	71	81	74	167	29	164
High blood pressure (N) ³		659	80	49	20	6	18	28	11	65
High cholesterol (N) ³		471	50	50	12	6	11	35	5	47
Diabetes (N) ³		337	50	21	8	-	10	17	-	26
Cardiovascular disease (N) ⁴		272	34	17	10	5	9	13	-	16
Respiratory diseases (N) ⁴		635	100	61	16	14	18	39	6	28
FIB (mg/dL)	Mean	341.5	347.5	304.6	340.0	307.8	319.0	317.6	317.6	310.4
	SD	103.1	109.8	93.1	101.0	80.6	88.6	102.2	99.8	89.9
hsCRP (mg/L)	Mean	4.0	5.2	2.5	4.2	2.1	3.1	3.2	2.4	3.2
	SD	7.8	10.9	4.2	7.0	3.0	4.1	5.1	3.7	6.6
IL6 (pg/mL)	Mean	2.3	2.5	1.7	2.0	1.4	2.2	1.8	2.5	2.1
	SD	2.1	2.4	1.8	1.9	1.4	2.3	1.8	2.7	2.0
sICAM (ng/mL)	Mean	299.0	307.4	217.3	285.7	218.9	298.6	254.5	272.0	242.0
	SD	113.7	125.8	69.0	119.1	56.5	103.0	90.0	88.6	85.3
8PGFT (ng/g*CRE)	Mean	688.1	682.3	391.1	605.2	417.1	680.6	541.7	560.7	485.9
	SD	403.4	441.2	200.9	232.7	242.4	460.9	348.3	246.7	516.0

Abbreviations: 8PGFT 8-isoprostane, CS cigarette smoker, CRE creatinine, dual-EPRODS current smoker and user of electronic cigarettes, dual-SMKLS current smoker and user of smokeless tobacco, ExSM former smoker, EPRODS user of electronic cigarettes, FIB fibrinogen, hsCRP high-sensitivity C-reactive protein, IL6 interleukin 6, sICAM soluble intercellular adhesion module, SD standard deviation, SMKLS user of smokeless tobacco, test-CS test dataset of current smokers, test-ExSM test dataset of former smokers

¹ Full data of all current smokers

² Random selection from data of all current smokers to match the number of former smokers for machine learning

³ Number of participants who answered "Yes" in the questionnaire

⁴ Number of participants who answered "Yes" in the relevant disease questionnaire

- In accordance with ICPSR data release rules, tables with cell sizes smaller than the threshold for the specific dataset will not be released

rent smokers than exclusive users of a single product. In the exposure biomarkers classification model, the most important feature was TNE7, followed by HPMa. In the potential harm biomarkers classification model, 8PGFT and sICAM showed the highest feature importance, followed by IL6.

Disease prevalence Among those indicated by either classification model as a current or former smoker, we calculated the percentage with CVD or respiratory diseases for each group. The percentage with CVD or respiratory diseases was consistently higher among those classified as current than former smokers in both models (Table 3).

Discussion

We first developed the biomarkers of exposure-based classification model to classify participants as either current or former smokers. The result should reflect the overall exposure level to constituents of cigarette smoke. This model predicted that users of electronic cigarettes and smokeless tobacco products would be more like former than current smokers, suggesting that the use of these products reduced exposure. Consistent with previous reports [18, 20], TNE7, the total amount of nicotine metabolites [24], was identified as the most important variable in distinguishing between current and former smokers. TNE7 was much lower in former than current smokers. However, the level in users of either electronic cigarettes or smokeless tobacco products was higher than in current smokers, suggesting that the classification of these two groups did not depend on the TNE7

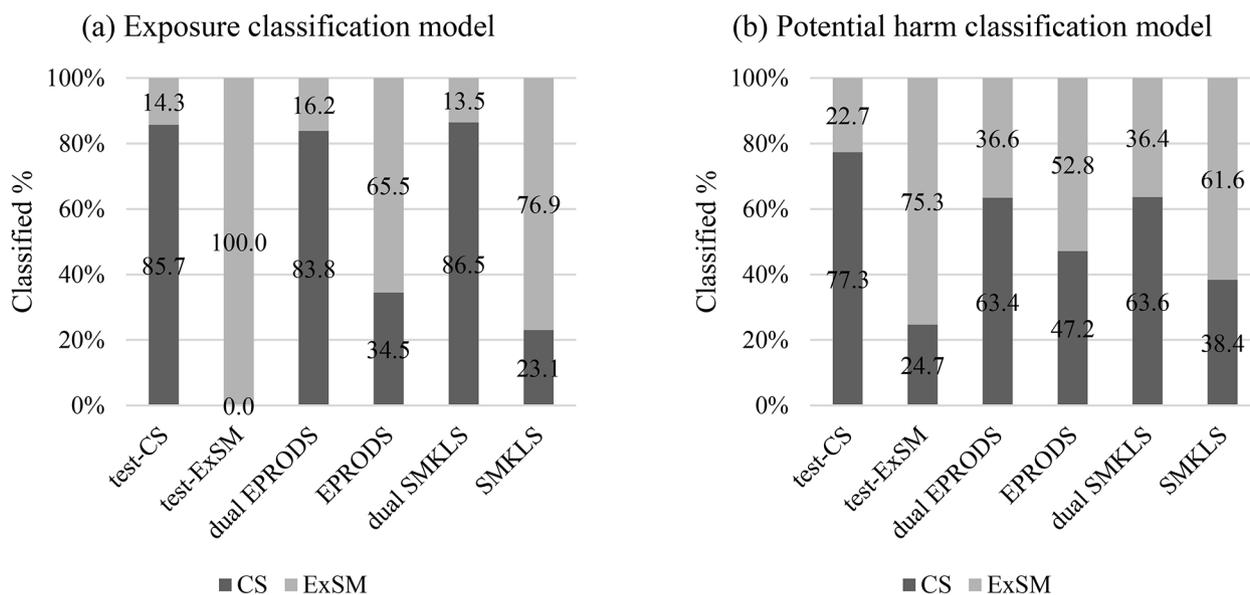


Fig. 1 Percentage of users of each tobacco and nicotine product classified as current or former smokers by the two classification models. *Abbreviations:* CS current cigarette smoker, dual-EPRODS user of conventional cigarettes and electronic cigarettes, dual-SMKLS user of conventional cigarettes and smokeless tobacco products, ExSM former smoker, EPRODS user of electronic cigarettes, SMKLS user of smokeless tobacco, test-CS models using test data for current smokers, test-ExSM models using test data for ex-smokers

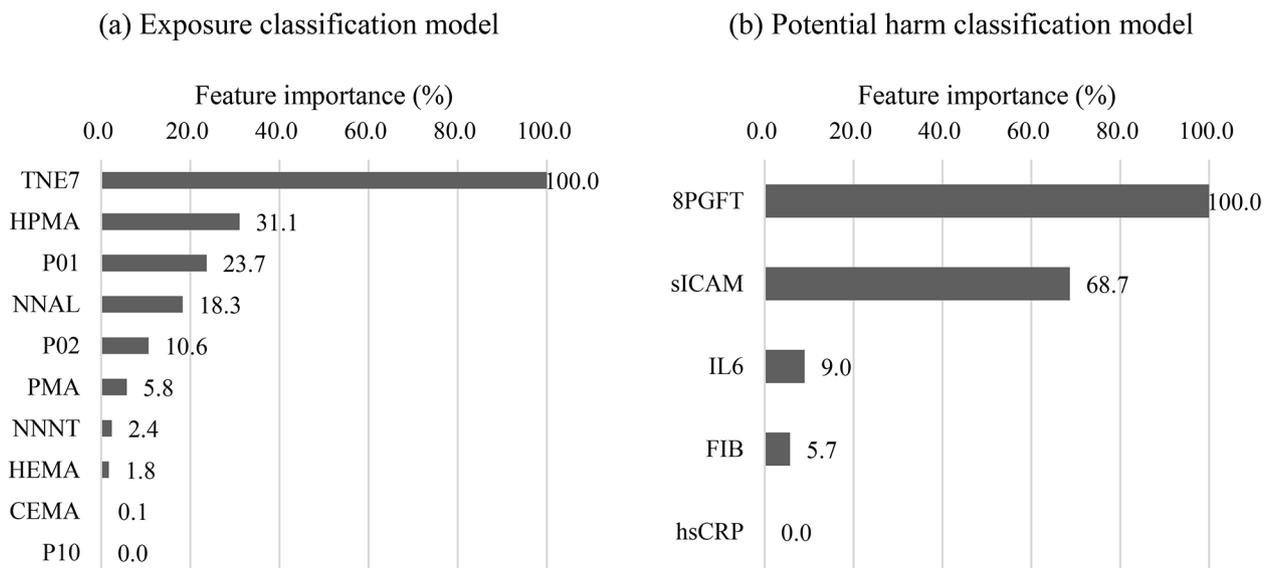


Fig. 2 The features contributing to classification as current or former smokers in cross-validation of the exposure and potential harm classification models (feature importance %). *Abbreviations:* 8PGFT 8-isoprostane, CEMA N-acetyl-S-(2-carboxyethyl)-L-cysteine, FIB fibrinogen, HEMA N-acetyl-S-(2-hydroxyethyl)-L-cysteine, HPMA N-acetyl-S-(3-hydroxypropyl)-L-cysteine, hsCRP high-sensitivity C-reactive protein, IL6 interleukin 6, NNAL 4-(methylnitrosamino)-4-(3-pyridyl)-1-butanol, NNNT N'-nitrosornicotine, sICAM soluble intercellular adhesion module, TNE7 total nicotine equivalents, PMA N-acetyl-S-(phenyl)-L-cysteine, P01 1-naphthol or 1-hydroxynaphthalene, P02 2-naphthol or 2-hydroxynaphthalene, P10 1-hydroxypyrene

level but on other biomarkers of exposure. High TNE7 levels in users of potentially lower-risk tobacco products are expected because they provide access to the parent chemical (i.e., nicotine). It is therefore likely that other biomarkers of exposure contribute more to the

classification of users of electronic cigarettes and smokeless tobacco products.

Similarly, the classification model based on biomarkers of potential harm also classified users of electronic cigarettes and smokeless tobacco products as more like

Table 3 The percentage of participants classified as current or former smokers by exposure or potential harm classification models who had cardiovascular or respiratory diseases

	Predicted by exposure classification model		Predicted by potential harm classification model	
	Current smoker	Former smoker	Current smoker	Former smoker
Cardiovascular disease	9.9%	6.4%	10.9%	6.3%
Respiratory diseases	19.4%	16.7%	22.2%	14.2%

former than current smokers, but to a lesser extent than the exposure model. The difference in the results between users of electronic cigarettes and smokeless products from the two different classification models may be due to the difference in characteristics as biomarkers. The exposure biomarkers are for cigarette smoke-specific constituents, but the biomarkers of potential harm could be influenced by a variety of factors other than smoking. The most important features of the potential harm model were 8PGFT and sICAM. Both biomarkers are known to be associated with oxidative stress and/or inflammation [25, 26] and also recognized as predisposing factors for CVD [27, 28]. IL-6 was also important in the potential harm biomarkers model. IL-6 has a short half-life and inter-individual variability, and its association with risk for CVD and respiratory diseases is still controversial. However, it is among the most frequently studied biomarkers and several studies have identified its importance for disease risk [29, 30]. The biomarkers of potential harm are important, but no single one can explain all disease risk. Our model therefore includes several biomarkers of potential harm, and could approximate the relative disease risk between cigarette smoking and smoking cessation. Our potential harm classification model is based on the hypothesis that smoking cessation could minimize the risk of smoking-associated diseases. If a user of potentially lower-risk tobacco products is classified as a former smoker based on their biomarkers of potential harm, this suggests that using these products does indeed reduce the disease risk compared with smoking conventional cigarettes. Our results therefore suggest that both electronic cigarettes and smokeless tobacco products could contribute to reduce smoking-associated disease risk. Disease prevalence was slightly lower in participants classified as former smokers than those classified as current smokers. Our results show that participants classified as current smokers in either model had relatively higher rates of CVD or respiratory diseases than those classified as former smokers. There may therefore be a reduced risk of disease among users of potentially lower-risk products. The results of the feature

importance, and of previous reports of biomarkers, suggest an association with CVD.

We also evaluated dual users of both conventional cigarettes and either electronic cigarettes or smokeless tobacco products. The exposure model recognized these dual users as current smokers. If participants use both cigarettes and electronic cigarettes or smokeless tobacco products on a daily basis, the biomarkers of exposure could be closer to those of cigarette smokers than those who use only potentially lower-risk products. The result of the exposure model reflects this hypothesis. However, the model based on biomarkers of potential harm showed an intermediate result for dual users, between smokers and users of only electronic cigarettes or smokeless tobacco products, suggesting that dual use reduces the risk of harm. However, longitudinal observation is necessary to detect significant changes in biomarkers of potential harm, and to establish whether robust homeostasis in the human body suppresses rapid changes. Multiple exposures and cumulative effects possibly contribute to the changes in the biomarkers of potential harm as well as the manifestation of health impacts. The extent of risk reduction in dual users would therefore depend on the proportion of cigarette smoking and use of electronic cigarettes and smokeless tobacco products.

This study had several limitations. In this study, we used strict definitions for each group of tobacco product users. Exclusive users had to use only one product on a daily basis. Users of electronic cigarettes and smokeless tobacco products who also smoked were defined as dual-users. Some similarities were seen between users of electronic cigarettes and smokeless tobacco products and former smokers in the population used for modeling. However, this should be further investigated with more study participants to reflect the real-world population. The health impact of e-cigarettes is still controversial [31, 32] and the subject of ongoing study, and we cannot conclude that the risk of disease is reduced even if e-cigarette users were more likely to be classified as former smokers in this study. The variables used in this study are absent from the succeeding waves, and we therefore cannot evaluate the time-related changes in the classification results. However, longitudinal analyses could provide more insights into the health impact of these new tobacco products because the pathogenesis of smoking-related chronic diseases takes time to manifest. Expansion of variables to include items such as “Frequency of product use” and “Start of product use” could also provide additional insights, although we could not obtain sufficient data for this analysis from Wave 1 alone. These limitations could be addressed if the variables used in this study are available in future waves of the PATH study.

Conclusion

The results of our classification models based on biomarkers of exposure and potential harm showed that the biomarker profiles of people who used electronic cigarettes or smokeless tobacco products were more like those of former than current smokers. This suggests that exposure to constituents in cigarette smoke and the resulting biological effects have potential to be reduced with the use of these products, at least among those involved in the PATH study.

Abbreviations

8PGFT	8-isoprostane
BoE	biomarkers of exposure
BoPH	biomarkers of potential harm
CEMA	N-acetyl-S-(2-carboxyethyl)-L-cysteine
COPD	chronic obstructive pulmonary disease
CS	cigarette smokers
CVD	cardiovascular disease
EPRODS	electronic cigarettes
ExSM	former smokers
FDA	Food and Drug Administration
FIB	fibrinogen
HEMA	N-acetyl-S-(2-hydroxyethyl)-L-cysteine
HPHC	harmful and potentially harmful constituent
HPMA	N-acetyl-S-(3-hydroxypropyl)-L-cysteine
hsCRP	high-sensitivity C-reactive protein
ICPSR	Inter-University Consortium for Political and Social Research
IL6	interleukin 6
NIH	National Institutes of Health
NNAL	4-(methylnitrosamino)-4-(3-pyridyl)-1-butanol
NNNT	N'-nitrosonornicotine
PATH	Population Assessment of Tobacco and Health [study]
PMA	N-acetyl-S-(phenyl)-L-cysteine
P01	1-naphthol or 1-hydroxynaphthalene
P02	2-naphthol or 2-hydroxynaphthalene
P10	1-hydroxypyrene
RESP	respiratory diseases
sICAM	soluble intercellular adhesion molecule
SMKLS	smokeless tobacco products
TNE7	total nicotine equivalents
ROC-AUC	receiver operating characteristic curve-area under curve

Acknowledgements

We thank Melissa Leffler, MBA, from Edanz (<https://jp.edanz.com/ac>) for editing a draft of this manuscript.

Authors' contributions

HO conceptualized the study design, interpreted the results and wrote the first draft of the manuscript. YT contributed to the technical guidance and results discussion of machine learning for smokers and nonsmokers, and SI contributed significantly to the writing of the manuscript on the biological results discussion. All authors read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public and not-for-profit sectors. This work was supported by Japan Tobacco, Inc.

Data availability

Data are available in a public, open-access repository, the National Addiction and HIV Data Archive: <https://doi.org/10.3886/Series606>.

Declarations

Ethics approval and consent to participate

The PATH study was approved by the Westat Institutional Review Board. All participants aged 18 and older provided informed consent to participate in the PATH study. The Institutional Review Board of the Research Institute

of Healthcare Data Science approved our study design and data collection protocol (RI2019014). The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

All authors are employees of Japan Tobacco, Inc.

Author details

¹Scientific Product Assessment Center, Japan Tobacco Inc, 6-2 Umegaoka, Aoba-ku, Yokohama 227-8512, Kanagawa, Japan

Received: 20 October 2022 / Accepted: 23 March 2023

Published online: 29 March 2023

References

- Ambrose JA, Barua RS. The pathophysiology of cigarette smoking and cardiovascular disease. *Journal of the American College of Cardiology*. 2004 May;43(10):1731–7. <https://doi.org/10.1016/j.jacc.2003.12.047>
- Health Canada G. of C. Determination of "Tar", Nicotine and Carbon Monoxide in Mainstream Tobacco Smoke-Official Method. Ottawa, Canada: Health Canada (SOR-2000-272). Health Canada. 1999. Available from: <https://healthy-canadians.gc.ca/en/open-information/tobacco/t100/nicotine>
- Products C, U.S. Food and Drug Administration. for T. Reporting Harmful and Potentially Harmful Constituents in Tobacco Products and Tobacco Smoke Under Sect. 904(a)(3) of the Federal Food, Drug, and Cosmetic Act. 2020. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/reporting-harmful-and-potentially-harmful-constituents-tobacco-products-and-tobacco-smoke-under>
- Center for Tobacco Products. Established List of HPHCs in Tobacco Products and Tobacco Smoke. U.S. Food and Drug Administration. 2019. Available from: <https://www.fda.gov/tobacco-products/rules-regulations-and-guidance/harmful-and-potentially-harmful-constituents-tobacco-products-and-tobacco-smoke-established-list>
- Roethig HJ, Feng S, Liang Q, Liu J, Rees WA, Zedler BK. A 12-Month, Randomized, Controlled Study to Evaluate Exposure and Cardiovascular Risk Factors in Adult Smokers Switching From Conventional Cigarettes to a Second-Generation Electrically Heated Cigarette Smoking System. *The Journal of Clinical Pharmacology*. 2008 May;48(5):580–91. <https://doi.org/10.1177/0091270008315316>
- Miura N, Yuki D, Minami N, Kakehi A, Futamura Y. A study to investigate changes in the levels of biomarkers of exposure to selected cigarette smoke constituents in Japanese adult male smokers who switched to a non-combustion inhaler type of tobacco product. *Regul Toxicol Pharmacol*. 2015 Apr;71(3):498–506. <https://doi.org/10.1016/j.yrtph.2015.02.007>
- Ogden MW, Marano KM, Jones BA, Morgan WT, Stiles MF. Switching from usual brand cigarettes to a tobacco-heating cigarette or snus: Part 3. Biomarkers of biological effect. *Biomarkers*. 2015 Oct 3;20(6–7):404–10. <https://doi.org/10.3109/1354750X.2015.1094135>
- Haziza C, de La Bourdonnaye G, Merlet S, Benzimra M, Ancerewicz J, Donelli A et al. Assessment of the reduction in levels of exposure to harmful and potentially harmful constituents in Japanese subjects using a novel tobacco heating system compared with conventional cigarettes and smoking abstinence: A randomized controlled study in confinement. *Regulatory Toxicology and Pharmacology*. 2016 Nov;81:489–99. <https://doi.org/10.1016/j.yrtph.2016.09.014>
- Haziza C, de La Bourdonnaye G, Skiada D, Ancerewicz J, Baker G, Picavet P, et al. Biomarker of exposure level data set in smokers switching from conventional cigarettes to Tobacco Heating System 2.2, continuing smoking or abstaining from smoking for 5 days. *Data in Brief*. 2017 Feb;10:283–93. <https://doi.org/10.1016/j.dib.2016.11.047>
- Yuki D, Takeshige Y, Nakaya K, Futamura Y. Assessment of the exposure to harmful and potentially harmful constituents in healthy Japanese smokers using a novel tobacco vapor product compared with conventional cigarettes and smoking abstinence. *Regul Toxicol Pharmacol*. 2018 Jul;96:127–34. <https://doi.org/10.1016/j.yrtph.2018.05.001>

11. Haziza C, De La Bourdonnaye G, Donelli A, Poux V, Skiada D, Weitkunat R et al. Reduction in Exposure to Selected Harmful and Potentially Harmful Constituents Approaching Those Observed Upon Smoking Abstinence in Smokers Switching to the Menthol Tobacco Heating System 2.2 for 3 Months (Part 1). *Nicotine & Tobacco Research*. 2019 Feb 5;22(4):539–548. <https://doi.org/10.1093/ntr/ntz013>
12. Sakaguchi C, Nagata Y, Kikuchi A, Takeshige Y, Minami N. Differences in Levels of Biomarkers of Potential Harm Among Users of a Heat-Not-Burn Tobacco Product, Cigarette Smokers, and Never-Smokers in Japan: A Post-Marketing Observational Study. *Nicotine & Tobacco Research*. 2021 Jan 27;23(7):1143–1152. <https://doi.org/10.1093/ntr/ntab014F>
13. Edmiston JS, Webb KM, Wang J, Oliveri D, Liang Q, Sarkar M. Biomarkers of Exposure and Biomarkers of Potential Harm in Adult Smokers Who Switch to e-Vapor Products Relative to Cigarette Smoking in a 24-week, Randomized, Clinical Trial. *Nicotine & Tobacco Research*. 2022 Feb 4;24(7):1047–54. <https://doi.org/10.1093/ntr/ntac029>
14. Lüdicke F, Picavet P, Baker G, Haziza C, Poux V, Lama N et al. Effects of Switching to the Menthol Tobacco Heating System 2.2, Smoking Abstinence, or Continued Cigarette Smoking on Clinically Relevant Risk Markers: A Randomized, Controlled, Open-Label, Multicenter Study in Sequential Confinement and Ambulatory Settings (Part 2). *Nicotine & Tobacco Research*. 2017 Apr 21;20(2):173–82. <https://doi.org/10.1093/ntr/ntx028>
15. Haziza C, de La Bourdonnaye G, Donelli A, Skiada D, Poux V, Weitkunat R et al. Favorable Changes in Biomarkers of Potential Harm to Reduce the Adverse Health Effects of Smoking in Smokers Switching to the Menthol Tobacco Heating System 2.2 for 3 Months (Part 2). *Nicotine & Tobacco Research*. 2019 May 24;22(4):549–59. <https://doi.org/10.1093/ntr/ntz084>
16. Sakaguchi C, Miura N, Ohara H, Nagata Y. Effects of reduced exposure to cigarette smoking on changes in biomarkers of potential harm in adult smokers: results of combined analysis of two clinical studies. *Biomarkers*. 2019 May 14;24(5):457–68. <https://doi.org/10.1080/1354750X.2019.1609579>
17. Hyland A, Ambrose BK, Conway KP, Borek N, Lambert E, Carusi C et al. Design and methods of the Population Assessment of Tobacco and Health (PATH) Study. *Tobacco Control*. 2016 Aug 8 [cited 2020 Jan 16];26(4):371–8. <https://doi.org/10.1136/tobaccocontrol-2016-052934>
18. Cheng Y-C, Reyes-Guzman CM, Christensen CH, Rostron BL, Edwards KC, Wang L et al. Biomarkers of Exposure among Adult Smokeless Tobacco Users in the Population Assessment of Tobacco and Health Study (Wave 1, 2013–2014). *Cancer Epidemiology, Biomarkers & Prevention*. 2020 Mar 1;29(3):659–67. <https://doi.org/10.1158/1055-9965.EPI-19-0766>
19. De Jesús VR, Bhandari D, Zhang L, Reese C, Capella K, Tevis D et al. Urinary Biomarkers of Exposure to Volatile Organic Compounds from the Population Assessment of Tobacco and Health Study Wave 1 (2013–2014). *International Journal of Environmental Research and Public Health*. 2020 Jul 28;17(15):5408. <https://doi.org/10.3390/ijerph17155408>
20. Anic GM, Rostron BL, Hammad HT, van Bommel DM, Del Valle-Pinero AY, Christensen CH et al. Changes in Biomarkers of Tobacco Exposure among Cigarette Smokers Transitioning to ENDS Use: The Population Assessment of Tobacco and Health Study, 2013–2015. *International Journal of Environmental Research and Public Health*. 2022 Jan 27;19(3):1462. <https://doi.org/10.3390/ijerph19031462>
21. Müller AC, Guido S. Introduction to machine learning with Python: a guide for data scientists. Newton, MA, USA: O'Reilly Media, Inc.; 2016.
22. Bourel M, Segura AM, Crisci C, López G, Sampognaro L, Vidal V, et al. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Res*. 2021 Sep;202:117450. <https://doi.org/10.1016/j.watres.2021.117450>
23. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A et al. caret: Classification and Regression Training. R-Packages. 2020. Available from: <https://cran.r-project.org/web/packages/caret/index.html>
24. Scherer G, Engl J, Urban M, Gilch G, Janket D, Riedel K. Relationship between machine-derived smoke yields and biomarkers in cigarette smokers in Germany. *Regul Toxicol Pharmacol*. 2007 Mar;47(2):171–83. <https://doi.org/10.1016/j.yrtph.2006.09.001>
25. Roberts LJ Jackson, Morrow JD. Measurement of F2-isoprostanes as an index of oxidative stress in vivo. *Free Radic Biol Med*. 2000 Feb;28(4):505–13. [https://doi.org/10.1016/S0891-5849\(99\)00264-6](https://doi.org/10.1016/S0891-5849(99)00264-6)
26. Bui TM, Wiesolek HL, Sumagin R. ICAM-1: A master regulator of cellular responses in inflammation, injury resolution, and tumorigenesis. *Journal of Leukocyte Biology*. 2020 Mar 17;108(3):787–99. <https://doi.org/10.1002/JLB.2MR0220-549R>
27. Lee R, Margaritis M, Channon M, Antoniadis K. C. Evaluating Oxidative Stress in Human Cardiovascular Disease: Methodological Aspects and Considerations. *Current Medicinal Chemistry*. 2012 Apr 1;19(16):2504–20. <https://doi.org/10.2174/092986712800493057>
28. Ridker PM, Hennekens CH, Roitman-Johnson B, Stampfer MJ, Allen J. Plasma concentration of soluble intercellular adhesion molecule 1 and risks of future myocardial infarction in apparently healthy men. *The Lancet*. 1998 Jan;351(9096):88–92. [https://doi.org/10.1016/S0140-6736\(97\)09032-6](https://doi.org/10.1016/S0140-6736(97)09032-6)
29. Danesh J, Kaptoge S, Mann AG, Sarwar N, Wood A, Angleman SB et al. Long-Term Interleukin-6 Levels and Subsequent Risk of Coronary Heart Disease: Two New Prospective Studies and a Systematic Review. *Baigent C, editor. PLoS Medicine*. 2008 Apr 8;5(4):e78. <https://doi.org/10.1371/journal.pmed.0050078>
30. Chen Y-WR, Leung JM, Sin DD. A Systematic Review of Diagnostic Biomarkers of COPD Exacerbation. *Eickelberg O, editor. PLOS ONE*. 2016 Jul 19;11(7):e0158843. <https://doi.org/10.1371/journal.pone.0158843>
31. Britton J. E-cigarettes, Public Health England, and common sense. *Lancet*. 2015;386(10000):1238–9. [https://doi.org/10.1016/S0140-6736\(15\)00145-2](https://doi.org/10.1016/S0140-6736(15)00145-2)
32. Stratton K, Kwan LY, Eaton DL, editors. *Public health consequences of E-cigarettes*. Washington DC: The National Academies Press; 2018. <https://doi.org/10.17226/24952>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.