

## **Supplementary material for**

# **Clustering Analysis of Volatile Organic Compound Biomarkers: to Identify Tobacco Exposure and Association with Cardiovascular Health Outcomes using an Observation Study Cohort**

Juan Zhao, PhD<sup>1</sup>, Haoyun Hong, BA<sup>1</sup>, Joseph Zhai<sup>1</sup>, Remy Poudel, MPH<sup>1</sup>, Sanjay Srivastava, PhD<sup>2</sup>, Andrew C. Stokes, PhD<sup>3</sup>, Pawel Konrad Lorkiewicz, PhD<sup>2</sup>, Tian Jiang, MS<sup>1</sup>, Rose Marie Robertson, MD<sup>1</sup>, Aruni Bhatnagar, PhD<sup>2</sup>, Jennifer L. Hall, PhD<sup>1</sup>, Naomi Hamburg, MD<sup>4</sup>, Rachel J. Keith, PhD<sup>2</sup>

1. American Heart Association, Dallas, TX, USA

2. American Heart Association Tobacco, Regulation and Addiction Center, University of Louisville School of Medicine, Louisville, KY

3. Department of Global Health, Boston University School of Public Health, Boston, MA

4. American Heart Association Tobacco, Regulation and Addiction Center, Vascular Biology Section, Whitaker Cardiovascular Institute, Boston University School of Medicine, Boston, MA.

## **CORRESPONDING AUTHOR:**

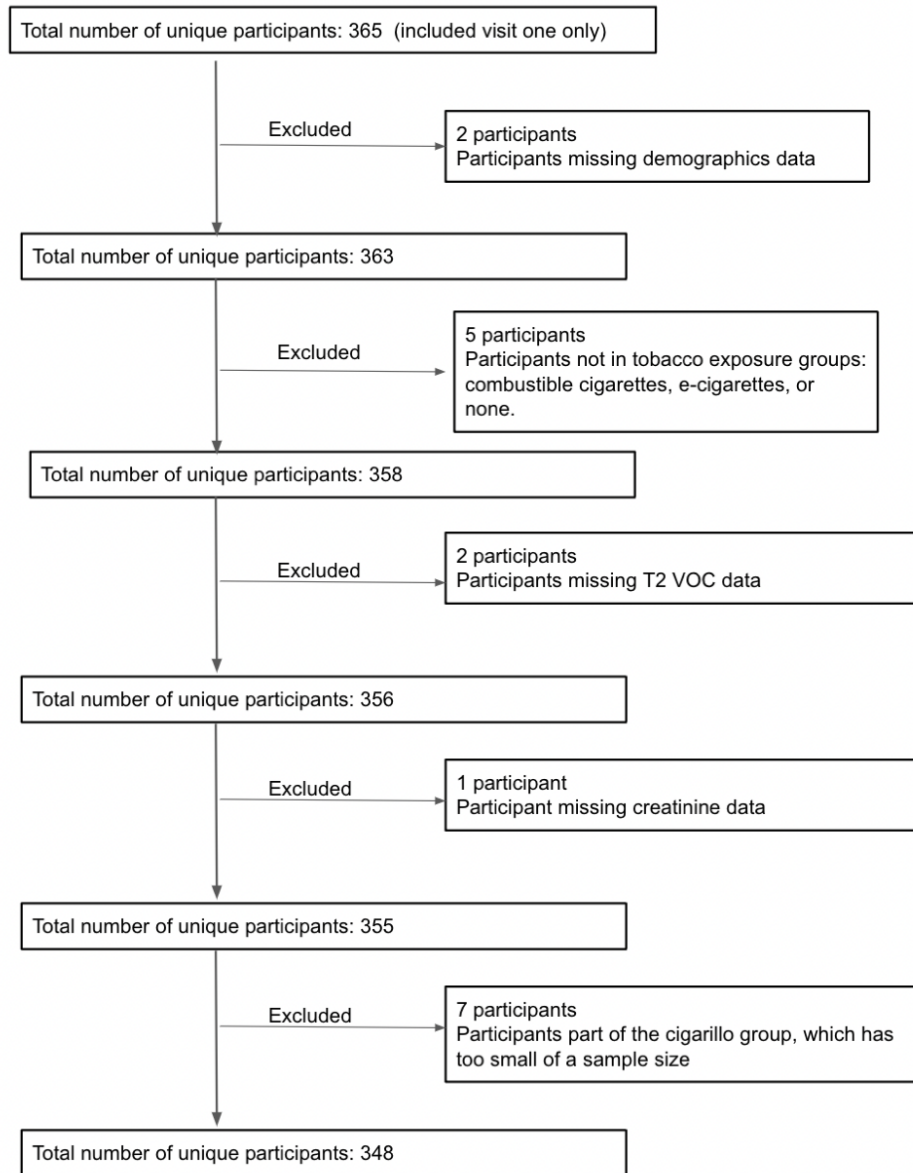
Rachel J. Keith, PhD

American Heart Association Tobacco, Regulation and Addiction Center, University of Louisville School of Medicine, Louisville, KY

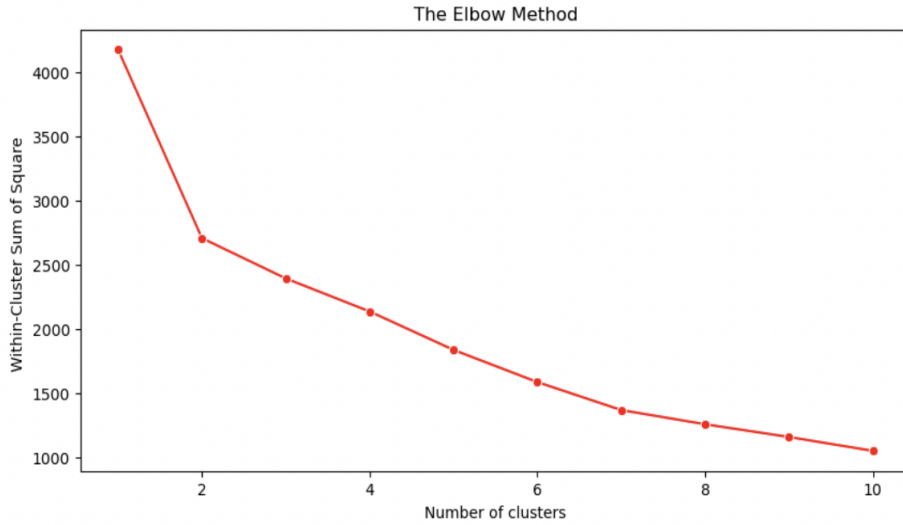
Rachel.keith@louisville.edu

**Table S1.**

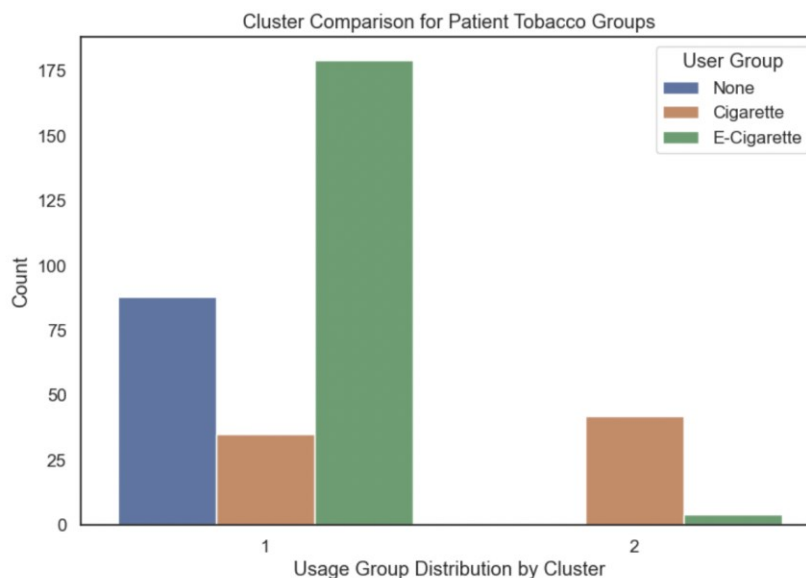
<b>Abbreviation</b>	<b>Compound (Unit)</b>
<b>AAMA</b>	N-Acetyl-S-(2-carbamoylethyl)-l-cysteine (ng/ml)
<b>3HPMA</b>	N-Acetyl-S-(3-hydroxypropyl)-l-cysteine (ng/ml)
<b>2HPMA</b>	N-Acetyl-S-(2-hydroxypropyl)-l-cysteine (ng/ml)
<b>MA</b>	Maleic acid or Malonic acid (ng/ml)
<b>DHBMA</b>	N-Acetyl-S-(3,4-dihydroxybutyl)-l-cysteine (ng/ml)
<b>MHBMA3</b>	N-Acetyl-S-(4-hydroxy-2-buten-1-yl)-l-cysteine (ng/ml)
<b>PGA</b>	Phenylglyoxylic acid (ng/ml)
<b>HPMMA</b>	N-Acetyl-S-(3-hydroxypropyl-1-methyl)-l-cysteine (ng/ml)
<b>2MHA</b>	2-Methylhippuric acid (ng/ml)
<b>34MHA</b>	3-Methylhippuric acid+4-Methylhippuric acid (ng/ml)
<b>BMA</b>	N-Acetyl-S-(benzyl)-l-cysteine (ng/ml)
<b>CYMA</b>	N-Acetyl-S-(2-cyanoethyl)-l-cysteine (ng/ml)



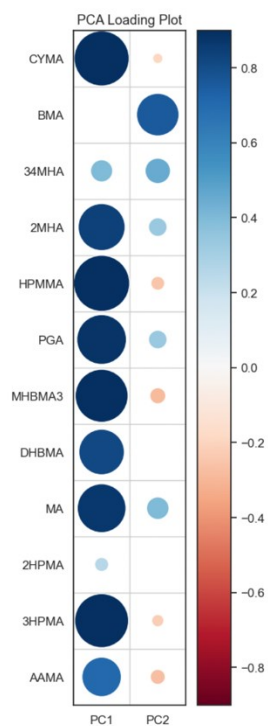
**Figure S1. The flow diagram of study selection based on inclusion and exclusion criteria.**



**Figure S2. Elbow method for determining the number of clusters.** This line chart illustrates the application of the Elbow Method to determine the optimal number of clusters for KMeans clustering. The y-axis represents the within-cluster sum of squares (WCSS), which measures the variance within each cluster, while the x-axis shows the number of clusters. The plot indicates a sharp decline in WCSS as the number of clusters increases from 1 to 2, with the rate of decrease diminishing after 2 clusters, suggesting that 2 clusters would be an optimal number.



**Figure S3. The distribution of patient tobacco usage among clusters.** This bar chart illustrates the distribution of patient tobacco usage across two distinct clusters. Cluster 1 shows a higher prevalence of e-cigarette users (Green), with a small proportion of patients who do not use tobacco (Blue), and few cigarette smokers (Orange). In contrast, Cluster 2 predominantly consists of cigarette users (Orange), with very few patients using e-cigarettes or non-users. The data suggests a potential pattern in tobacco usage behavior among the different clusters.



**Figure S4.** PCA loading plot. The PCA loading plot reveals the VOCs contributing most to the differentiation. Specifically, compounds such as CYMA, HPMMA, MHBMA3, and 3HPMA exhibit high positive loadings on the first principal component (PC1), underscoring their role in differentiating Cluster 1. Conversely, BMA shows a high positive loading on PC1, suggesting its association with Cluster 1. This bimodal distribution underscores distinct biomarker profiles, potentially reflecting different tobacco exposure statuses within the population.