Identifying Genetic Variants Associated With Noise-induced Hearing Loss Based on a Novel Strategy for Evaluating Individual Susceptibility

Zhuang Jiang
Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiaotong University
Botao Fa
Shanghai Jiaotong University
Xunmiao Zhang
Tongji University Affiliated Shanghai Pulmonary Hospital
Yanmei Feng
Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiaotong University
Jiping Wang
Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiaotong University
Haibo Shi
Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiaotong University School
Daoyuan Sun
Tongji University Affiliated Shanghai Pulmonary Hospital
Hui Wang (wangh2005@alumni.sjtu.edu.cn)
Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiaotong University School https://orcid.org/0000-0003-4196-997X
Shankai Yin
Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiaotong University

Research

Keywords: noise-induced hearing loss, individual susceptibility, machine learning, susceptible gene, genetic variant

Posted Date: December 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-122195/v1

Version of Record: A version of this preprint was published on June 6th, 2021. See the published version at https://doi.org/10.1016/j.heares.2021.108281.

Abstract

Background: The overall genetic profile for noise-induced hearing loss (NIHL) remains to be explored. Here we used a novel machine learning (ML) strategy to evaluate individual susceptibility to NIHL and identify the underlying genetic variants based on a subsample of participants with extreme phenotype.

Methods: Demographic and audiometric data of 5,539 shipbuilding workers from large cross-sectional surveys were included in four ML algorithms to predict the hearing level. The area under the curve (AUC) and prediction accuracy were used to assess the performance of the classification models. We screened 300 participants that were misclassified by all of the four ML models, with extreme phenotypes implying they were either highly susceptible or resistant to NIHL and used whole-exome sequencing (WES) to identify the underlying variants associated with NIHL risk among the NIHL-susceptible and NIHL-resistant individuals. Subsequently, candidate risk loci were validated in a large independent noise-exposed cohort, followed by a meta-analysis.

Results: With 10-fold cross-validation, the performances of the four ML models were robust and similar, with average AUC and accuracy ranging from 0.783 to 0.798 and 73.7% to 73.8%, respectively. The phenotypes of the NIHL-susceptible group and NIHL-resistant group were significantly different (all $p<0.001$). After WES analysis and filtering, 12 novel variants contributing to NIHL susceptibility were identified and replicated. The meta-analyses shown that the rs41281334 A allele of *CDH23* (OR=1.506, 95% CI=1.106-2.051) and the rs12339210 C allele of *WHRN* (OR=3.06, 95% CI=1.398-6.700) were significantly associated with increased risk of NIHL after adjustment for conventional risk factors.

Conclusions: This study determined two novel genetic variants in *CDH23* rs41281334 and *WHRN* rs12339210 associated with NIHL risk, based on a potential approach for evaluating individual susceptibility using ML models.

Trial registration: Chinese Clinical Trial Registry (registration number: ChiCTR-RPC-17012580)

Background

Noise is one of the most common pollutions in industrial settings and communities. Long-term exposure to hazardous noise can result in noise-induced hearing loss (NIHL), which has become the second most frequent form of sensorineural hearing loss, after age-related hearing impairment[1]. It is estimated that 16% of disabling hearing loss in adults is attributed to occupational noise worldwide, ranging from 7–21% in various subregions[2, 3], posing a huge burden to health care[4].

NIHL is a heterogeneous disease induced by interactions between genetic and environmental factors. It is known that the risk of noise-induced damage to the auditory system depends on the noise intensity and duration of exposure[5]. However, there are undeniable large variations in a person's susceptibility to noise exposure, and such differences are likely regulated by genetic background[6, 7]. Twin studies estimated the heritability for noise-induced hearing loss (NIHL) to be approximately 36%[8], and strain-specific variation in sensitivity has been demonstrated in several heterozygous and homozygous knockout mice, including *SOD1, CDH23, and MYH14*[9–11] which have been shown to be more sensitive to noise than their wild-type littermates. To date, many case-control studies screening for tag single nucleotide polymorphisms (SNPs) have been adopted in noise-exposed human populations, involving putative susceptibility genes that are known to play functional roles in the inner ear[12]. However, this method is inherently biased and limited in its scope for the discovery of novel mutations[13], and the overall genetic profile for NIHL remains to be explored.

To proceed with the study of NIHL susceptibility genes, an appropriate quantification and selection of workers who are highly susceptible or resistant to noise is crucial. High-throughput sequencing technology such as whole exome sequencing (WES)[14] is a highly effective approach for discovering genes underlying multifactorial diseases, and the extreme phenotype design will increase the power of a genome-wide search for susceptible or protective genetic variants[15, 16]. Yet, scientific consensus is still lacking regarding screening protocols for identifying individuals who are at the two extreme ends of the NIHL distribution. Previously, susceptible individuals were selected with hearing loss at 3, 4, or 6 kHz after matching with controls in the aforementioned tag SNPs studies. Other studies have estimated the individual susceptibility to NIHL based on acoustic reflex[17], or the regression model between hearing loss and noise exposure dose[18]. The sample size of these studies was relatively small, and they failed to take other risk factors into consideration. For example, aging[19], and the use of alcohol and tobacco[20] and other harmful chemicals[21] will aggravate the progress of NIHL, while female sex is a protective factor[22]. Another study explored the

classification of noise-susceptible and noise-resistant workers based on the ISO 1999:1990 standard[23]; however, the results depended on the statistical distribution of hearing threshold levels of a specific noise-exposed population, and have become arbitrary. In addition, while the notch audiogram at 3–6 kHz was considered as the marker for the differential diagnosis of NIHL from other types of sensorineural hearing loss[24, 25], extended high-frequency (EHF) audiometry has been advocated in recent years and has shown a great advantage in the early identification of hearing loss due to various reasons, especially noise exposure[26, 27]. Thus, quantifying NIHL only at the traditional frequencies may introduce bias.

Machine learning (ML) techniques have been successfully used for building medical predictive classification models[28, 29]. One major advantage of ML is its ability to identify nonlinear relationships between multiple variables and a targeted outcome (in our case, the contribution of various factors to NIHL), which makes it possible to assess the individual susceptibility by integrating multiple factors. One recent study used ML to predict NIHL with limited variables including age, exposure duration, and statistical metric of noise kurtosis[30]. The main focus of this study was just on the prediction accuracy of algorithms, which ranged from 76.6–83%, while the reasons for the prediction errors were not further analyzed. The participants who were misclassified by ML that deviated significantly from the predicted results primarily had abnormal phenotypes. It is clear that some unknown mechanisms or factors must be responsible for them. In the case of NIHL, these mechanisms could be attributed to genetic variants.

In this study, we aimed to evaluate the individual susceptibility of NIHL in a large-scale noise-exposed population and investigate the potential genetic variants using the misclassified participants with extreme phenotypes screened by ML models. We hypothesized that WES analysis might identify variants associated with extreme susceptibility or resistance to NIHL. To our knowledge, this is the first study to discover variants using exome sequencing and an extreme phenotype study design for NIHL.

Methods

Study participants

The study was conducted in a large shipyard in Shanghai and comprised several phases. It was initiated in June 2015, after which the employees underwent occupational health examinations annually. The most recent auditory and demographic data were collected from a total of 6,840 Chinese noise-exposed workers, primarily comprising grinders, welders, stampers, and cutters, all of whom are exposed to high levels of noise. A structured questionnaire was filled through face-to-face interviews to collect demographic features, occupational history, medical history, smoking and alcohol drinking habits, family history including genetic and drug-related hearing loss, the use of hearing protection devices, and exposure to other harmful chemicals. The exclusion criteria are shown in Fig. 1. This study was approved by the Institutional Ethics Review Board of the Shanghai Sixth People's Hospital affiliated with Shanghai Jiao Tong University (Approval No: 2017 – 136) and was registered in the Chinese Clinical Trial Registry (registration number: ChiCTR-RPC-17012580). Potential consequences and benefits were explained, and written informed consent was obtained from each participant before this study.

Noise exposure estimation

Industrial noise was measured with an ASV5910-R digital recorder (Aihua Instruments; Hangzhou, China) across the work areas of different jobs according to the national standard of China[31]. The long-term equivalent (Leq) noise level was adopted as the primary exposure metric and measured three times at each spot. The mean Leq value in each spot was transformed into 8 h of continuous equivalent A-weighted sound pressure level (Leq8h, shown in Table 1). Cumulative noise exposure (CNE) was calculated using LAeq8h over the years of on-duty time (Leq-total, dBA·year): $CNE = Leq8h + 10*lg(T)$, where T is the duration of employment in years[32].

Table 1
Characteristics of the total samples for machine learning modelling.

Variables	B-HL group (n = 1926)	W-HL group (n = 3613)	*P* value
Age (yrs)	33.6 ± 7.4	42.3 ± 8.3	< 0.001
Sex, n (%)			< 0.001
Male	1518(78.8)	3169(87.7)	
Female	408(21.2)	444(12.3)	
Career length (yrs)	6.7 ± 4.4	8.2 ± 4.7	< 0.001
CNE (dB*year)	90.1 ± 6.4	93.7 ± 7.5	< 0.001
Leq8h dB(A)	85.1 ± 4.2	87.1 ± 5.4	< 0.001
Smoking, n (%)			0.014
Currently	675(35.0)	1387 (38.4)	
Never	1251(65.0)	2226 (61.6)	
Drinking n, (%)			< 0.001
Currently	846(34.0)	1402(38.8)	
Never	1641(66.0)	2211(61.2)	
$PTA_{0.5–2\ kHz}$	14.8 ± 4.6	20.9 ± 7.7	< 0.001
$PTA_{3-6\ kHz}$	15.4 ± 6. 2	39.4 ± 16.0	< 0.001
$PTA_{10-12.5\ kHz}$	19.5 ± 9.9	51.3 ± 17.9	< 0.001
$PTA_{3-6\ \&\ 10–12\ kHz}$	17.0 ± 5.4	44.2 ± 14.0	< 0.001

Data are presented as mean ± standard deviation or numerical (%). The *p* values are the result of the Mann–Whitney test. The B-HL group is the better hearing group, The W-HL group is the worse hearing group. CNE, cumulative noise exposure. $PTA_{0.5–2\ kHz}$ represents the mean of 0.25, 0.5, 1, and 2 kHz for both ears. $PTA_{3-6\ kHz}$ represents the mean of 3, 4, and 6 kHz on both ears. $PTA_{10-12.5\ kHz}$ represents the mean of 10 and 12.5 kHz on both ears. $PTA_{3-6\ \&\ 10–12\ kHz}$ represent the mean of 3, 4, 6 kHz and 10, 12.5 kHz on both ears.

Audiometric evaluation and hearing impairment definition

Audiological evaluations, including tympanometry and pure-tone audiometry (PTA, range from 0.25-16 kHz) were performed on-site by qualified medical assistants in a multifunctional screening vehicle equipped with five soundproof chambers. The tests were performed at least 12 h after the participant's last shift in the noise-exposed job. First, otoscopic inspection and tympanometry were performed on each participant to establish the normal function of their external and middle ears. Subsequently, audiometry tests were carried out to determine hearing thresholds. Tympanograms were measured using a TympStar tympanometer (Grason-Stadler; Eden Prairie, MN, USA). The passing criterion was a type A tympanogram (peak between − 100 and + 100 daPa). Pure-tone air-conduction thresholds were measured from each of the two ears separately using Type 1066 manual audiometers (Natus Hearing & Balance; Taastrup, Denmark) coupled with TDH-39 headphones (Telephonics; Farmingdale, NY, USA) for conventional frequencies (0.25-8 kHz), and Sennheiser (Wedemark, Germany) HDA-300 headphones for the EHFs (10, 12.5, and 16 kHz). All thresholds were calculated in decibel hearing level (dB HL) and audiometers were calibrated annually according to the ISO 389-5-2006 standard. If a participant did not respond to the maximum intensity output of the audiometer for the EHFs, which were 90, 80, and 60 dB HL for 10, 12.5, and 16 kHz, respectively, the threshold was marked as 95, 85, and 65 dB HL, respectively. The bilateral average PTA across 3, 4, 6,10 and 12.5 kHz (or $PTA_{3-6\ \&\ 10–12.5}$, for short) was chosen to quantify the degree of NIHL. The threshold above 12.5 kHz was

not used because of a significant ceiling effect. The participants were classified into two groups according to the average $PTA_{3-6\&10-12.5}$ level: <25 dB HL as the better hearing level (B-HL) group and ≥ 25 dB HL as the worse hearing level (W-HL) group.

Machine learning for individual susceptibility assessment

Predictive modelling and performance evaluation

Different ML algorithms have different advantages. The following four supervised algorithms were used for the classification of hearing impairment: adaptive boosting (AdaBoost)[33], multi-layer perceptron (MLP)[34], random forest (RF)[35] and support vector machine (SVM)[36]. Age, sex, CNE, smoking, and alcohol drinking status were used as inputs for predictive modeling of $PTA_{3-6\&10-12.5}$ dichotomy. To train and validate the models, 10-fold cross-validation was adopted. In short, the entire dataset was randomly divided into 10 datasets using the caret package in R programming language v 3.6.1; nine of them for modeling, and the remaining one for validation. This step was repeated for 10 runs and the parameters of each algorithm were adjusted to ensure that the model had the best classification performance, which was estimated by two indexes: area under the receiver operator characteristic (ROC) curve (AUC) and prediction accuracy. The reported accuracy and AUC are the average over the 10 cycles. The algorithms were implemented using randomForest, adabag, monmlp, and e1071 libraries. When building the AdaBoost and RF models, default parameters were utilized. For the MLP model, we used five nodes in the first hidden layer and 15 ensembles to fit, and set the cut-off value of 0.5 in prediction probability for dichotomous classification. Regarding the SVM algorithm, hyper-parameters including gamma and cost were initially determined by 10-fold cross-validation, and the best of which was applied to train the classifier.

Individual susceptibility assessment and extreme individual selection

Individual susceptibility was assessed based on whether an individual was correctly classified by comparing the predicted label with the actual label. We consider the few participants who were misclassified with abnormal phenotypes to be either highly susceptible or resistant to NIHL. For example, those who were predicted to be in the B-HL group but actually had severe hearing loss were regarded as susceptible to NIHL, and conversely, those who were predicted to be the W-HL group but actually had better hearing were regarded as resistant to NIHL. To avoid errors of the model itself, the selection procedures for extreme individuals were strictly applied to the subsamples selected from the two misclassified groups by all four models for the next step of exome sequencing.

Identification and replication of risk variants

The procedure for genomic DNA preparation and exome sequencing analysis is described in the **supplementary material**. To identify the most likely pathogenic mutations, functional variants were filtered as follows: (1) considering the limited sample size and false negative signals, the SNPs with *p* values of Fisher's exact test for genetic association analysis < 0.05 or marginal significance (0.05 < P < 0.10), were selected; (2) the analysis was restricted to non-synonymous (missense), stop-gain/loss (nonsense), and splicing because changes in amino acids may affect biological functions; (3) minor allele frequencies of the mutation less than 0.1 in one of the 1000 genomic data (1000g_all), gnomAD data (gnomAD_ALL and gnomAD_EAS), and ExAC public database: and (4) mutation loci within candidate genes which have been shown to be involved in several crucial pathways including oxidative stress, potassium ion circulation, heat shock protein, notch signaling, apoptosis signaling, and monogenic gene of hereditary hearing loss[6, 12, 37] (see **supplementary Table S2** for gene list). From the filtered results, a truly pathogenic rare mutation can be obtained by removing the diversity locus between individuals. Additional two groups of noise-exposed participants were selected from the total sample for replication: 1,077 individuals with an average hearing threshold < 25 dB HL but as high as possible in terms of age and noise exposure dose were classified as the low-risk group, and 1,031 individuals with an average hearing threshold ≥ 25 dB HL but as low as possible in terms of age and noise exposure dose were classified as high-risk group. The demographic characteristics of the two groups are summarized in **Supplementary Table S1**. Candidate SNPs were genotyped using the ligation detection reaction and SNaPshot assay. Ten percent of the samples were randomly selected and genotyped repeatedly for quality control, and the concordance was > 99.9%.

Statistics

Continuous data are presented as mean ± standard deviation (SD) and were compared using the Mann–Whitney test between groups given their skewed distribution. Categorical data are expressed as number (%) and were compared using Pearson's $\chi 2$ test. The Hardy–Weinberg equilibrium test was performed before association analysis. The allelic frequencies between the NIHL-

susceptible and NIHL-resistant groups were compared using Fisher's exact test, and logistic regression was used to compare the difference in genotype distributions between the two independent validation groups under an additive model (AA = 0, Aa = 1, aa = 2; a is the minor allele), and odds ratios (ORs) with 95% confidence intervals (CIs) are presented. Statistical analyses were performed using SPSS 24.0 (IBM, Armonk, NY, USA) or PLINK v1.9[38]. Combined ORs from two stages were calculated using a Comprehensive Meta-Analysis (Biostat, Englewood, NJ, USA) with a fixed- or random-effect model after testing for heterogeneity. Differences were considered significant when $p < 0.05$.

Results

Demographic characteristics

Figure 1 outlines the study procedures and criteria for participant exclusion. A total of 5,539 individuals with the most recent hearing data were included for classification model construction, and their demographic characteristics are summarized in Table 1. The whole study samples were divided into two groups based on their hearing impairment level. Compared with the B-HL group, the W-HL group had an older age, longer career length, larger average cumulative noise exposure dose, and a higher proportion of male workers, smokers, and drinkers. Notably, the hearing impairment at different frequency ranges in the W-HL group were significantly worse than those in the B-HL group. All differences were statistically significant at $p < 0.01$.

Machine learning for individual susceptibility

The prediction performances of the four algorithms were robust and similar, with no significant difference in terms of accuracy and AUC. The MLP algorithm achieved the highest accuracy of 73.8% with an AUC of 0.797 for prediction, followed by Adaboost (accuracy of 73.7%; AUC of 0.796), SVM (accuracy of 73.7%; AUC of 0.798), and RF (accuracy of 73.7%; AUC of 0.783). The average accuracy and the average AUC values of the cross-validation for all models are shown in Fig. 2A and B. The average ROC curves were plotted in Fig. 2C, and the AUC values were almost 0.8, indicating a high predictive power (AUC = 1 indicates a perfectly discriminating test). Figure 3 depicts the confusion matrix of the predicted results of the four models with the whole samples. More than 4,000 participants were correctly classified in each model. Regarding the two misclassified groups with incompatible labels, there were 589, 582, 530, and 566 susceptible individuals (predicted label: B-HL, actual label: W-HL) and 797, 845, 936, and 858 resistant individuals (predicted label: W-HL, actual label: B-HL) in the Adaboost, MLP, RF, and SVM models, respectively. We took the intersection of the misclassified groups in all four algorithms, and there were 454 susceptible and 740 resistant individuals. Finally, according to the top prediction probability of MLP, we selected 150 participants from each of the two misclassified groups as the NIHL-susceptible group and NIHL-resistant group; the differences in the characteristics of these two groups are shown in Fig. 4. When compared to the NIHL-resistant group, the NIHL-susceptible group had a lower average age (25.8 ± 4.0 versus 46.7 ± 3.4 years old; $p < 0.001$), a shorter exposure duration (3.3 ± 2.1 versus 9.5 ± 4.9 years; $p < 0.001$), and a lower CNE level (86.9 ± 5.4 versus 94.4 ± 7.5 dBA*year; $p < 0.001$); however, their hearing loss was significantly more serious than the NIHL-resistant group (37.5 ± 11.1 versus 20.2 ± 4.2 dB HL; $p < 0.001$). In addition, there were significantly more female workers in the NIHL-susceptible group compared to the NIHL-resistance group (33 [22.0%] versus 4 [2.7%]; $p < 0.001$). There were 90 and 129 workers with a habit of smoking or drinking in the NIHL-susceptible group and the NIHL-resistant group, respectively (60% versus 86%, $p < 0.001$).

Discovery and validation of variants in the two extreme groups

After quality control for WES data, we discovered 993,409 SNPs and 207,683 short indels. Following the filtering criteria described above, we screened 1,104 non-synonymous low-frequency mutations located in the exons. All SNPs were in Hardy–Weinberg equilibrium ($p > 0.05$). The presence of NIHL-associated genes in our cohort was further investigated. We intersected the remaining loci that were enriched in the NIHL-susceptible group but were not present or observed at low frequencies in the NIHL-resistant group. Finally, with a threshold *p* value of 0.1, 12 novel variants were identified, detailed information and allele frequencies are displayed in Table 2. Given the sample size and the low mutation frequency, 10 variants in *NPC1*, *GJB2*, *EPHA2*, *TCIRG1*, *CDH23*, *KITLG*, *PTPRQ*, *WHRN*, *OTOGL*, and *ADGRV1* genes were significantly ($P < 0.05$) associated with the risk of NIHL, whereas other two mutations in *PTPRQ* and *KARS* genes were marginally ($0.05 < P < 0.1$) associated with the risk of NIHL. To validate the effects of these SNPs, we further sought to replicate them in the independent samples. The rs10862089 could not be genotyped due to unqualified primer amplification, the remaining loci were successfully replicated in both 300 WES samples and 2,108 independent validation samples. Since the rs199632510 variant was not detected in the NIHL-resistance group, it was excluded in the following

statistical analysis. Table 3 shows the association effects of the 10 SNPs with NIHL risk under additive model. After adjustment for age, sex, CNE, smoking, and drinking status, the rs1805084 of *NPC1*, the rs72474224 of *GJB2*, the rs41281334 of *CDH23*, the rs147541734 of *PTPRQ*, and the rs117041419 of *KARS* were significant in the 300 WES samples, while only the rs41281334 of *CDH23* and the rs12339210 of *WHRN* were significant in the independent validation samples. We also performed a meta-analysis for these SNPs combining the results of the two cohorts. As shown in Table 3, the risk allele A of rs41281334 (OR = 1.506, 95% CI = 1.106–2.051) and risk allele C of rs12339210 (OR = 3.06, 95% CI = 1.398-6.700) conferred a higher risk of NIHL.

Table 2
Association between the candidate SNPs and the risk for NIHL susceptibility in the two extreme groups.

Chr. position (hg19)	Gene	SNP	AA change	Major/minor allele	Risk allele frequency		OR* (95% CI)	*P* for allele*
					Susceptible group (n = 150)	Resistant group (n = 150)		
18:21112206	NPC1	rs1805084	p.R1266Q	C/T	0.3121	0.1967	1.853 (1.273–2.698)	0.001378
13:20763612	GJB2	rs72474224	p.V37I	C/T	0.07	0.02333	3.151 (1.319–7.527)	0.01061
1:16461581	EPHA2	rs55747232	p.T511M	G/A	0.08784	0.03667	2.53 (1.226–5.22)	0.01067
11:67810471	TCIRG1	rs199632510	p.G159E	G/A	0.02333	0	-	0.01508
10:73558128	CDH23	rs41281334	p.V43I	G/A	0.08	0.03333	2.522 (1.184–5.37)	0.02046
12:88926250	KITLG	rs3741457	p.T54A	T/C	0.03667	0.006667	5.671 (1.246–25.81)	0.02108
12:80878317	PTPRQ	rs57971665	p.R259Q	G/A	0.08667	0.04	2.277 (1.127–4.603)	0.02814
9:117170241	WHRN	rs12339210	p.P211A	G/C	0.02667	0.003333	8.192 (1.018–65.91)	0.03766
12:80699475	OTOGL	rs10862089	p.Q1102H	G/T	0.08667	0.04333	2.095 (1.055–4.16)	0.04567
5:90445889	ADGRV1	rs77469944	p.M6159V	A/G	0.04333	0.01333	3.352 (1.08–10.4)	0.04591
12:80899857	PTPRQ	rs147541734	p.I604T	T/C	0.02333	0.003333	7.143 (0.874–58.42)	0.06849
16:75675514	KARS	rs117041419	p.N85S	T/C	0.05	0.02	2.579 (0.9868-6.74)	0.07318

Chr., chromosome; AA, amino acid. *P* value is the results of Fisher's exact test, the odds ratio (OR) with 95% confidence interval (CI) shown is for the minor allele.

Table 3
Association between SNPs and NIHL risk under additive genetic models in the WES and the independent validation samples.

SNP	Risk allele	WES samples (n = 300)				Validation samples (n = 2108)				Meta-analysis	
		OR (95% CI)	P value	OR (95% CI) *	P value *	OR (95% CI)	P value	OR (95% CI) *	P value *	OR (95% CI)	P value
rs1805084	T	1.82 (1.246–2.658)	0.002	2.257 (1.231–4.137)	0.008	1.075 (0.934–1.238)	0.311	1.008 (0.853–1.19)	0.929	1.067 (0.908–1.252)	0.431
rs72474224	T	2.921 (1.237–6.898)	0.014	4.321 (1.178–15.845)	0.027	1.098 (0.81–1.489)	0.547	1.011 (0.709–1.443)	0.95	1.118 (0.794–1.576)	0.522
rs55747232	A	2.421 (1.177–4.981)	0.016	2.169 (0.841–5.594)	0.109	1.326 (1-1.757)	0.05	1.292 (0.928-1.8)	0.129	1.367 (1-1.869)	0.050
rs41281334	A	2.537 (1.183–5.443)	0.017	3.157 (1.166–8.551)	0.024	1.283 (0.982–1.676)	0.068	1.392 (1.006–1.927)	0.046	1.506 (1.106–2.051)	**0.0094**
rs3741457	C	5.005 (1.114–22.484)	0.036	3.246 (0.381–27.632)	0.281	1.244 (0.825–1.875)	0.298	1.06 (0.668–1.682)	0.806	1.114 (0.709–1.749)	0.6392
rs57971665	A	2.411 (1.167–4.982)	0.017	2.453 (0.831–7.243)	0.104	1.13 (0.884–1.445)	0.327	0.998 (0.748–1.331)	0.988	1.059 (0.802–1.399)	0.686
rs12339210	C	8.394 (1.037–67.971)	0.046	3.015 (0.307–29.579)	0.343	3.603 (1.783–7.283)	0.0004	3.066 (1.331–7.061)	0.008	3.06 (1.398-6.7)	**0.005**
rs77469944	G	3.464 (1.103–10.88)	0.033	3.234 (0.678–15.411)	0.141	1.267 (0.89–1.804)	0.189	0.985 (0.654–1.482)	0.941	1.063 (0.716–1.579)	0.762
rs14754174	C	7.294 (0.886–60.029)	0.065	4.42 (0.139-140.628)	0.400	1.748 (0.996–3.066)	0.052	1.591 (0.82–3.085)	0.169	1.65 (0.861–3.162)	0.132
rs117041419	C	2.667 (1.005–7.073)	0.049	5.631 (1.322–23.995)	0.019	0.989 (0.738–1.326)	0.943	0.978 (0.692–1.382)	0.898	1.075 (0.768–1.505)	0.675

* calculated with logistic regression adjusted by age, sex, CNE, smoking, and drinking status

Discussion

The present study proposes the utilization of ML algorithms for the assessment of individual susceptibility to NIHL based on prediction error, and investigates potential genetic variants through misclassified individuals with extreme phenotypes. In the first phase, all of the ML models exhibited robust and similar performances. Most importantly, we focused on the individuals that were misclassified by all of the four ML models, with large deviations from the predicted results implying that genetic variation may be involved. In the second phase, we used WES to explore the underlying variants associated with NIHL risk and validated in a large independent cohort. With this novel two-stage approach, we screened two subgroups of individuals with opposite phenotypes, and determined two novel variants of rs41281334 and rs12339210 that significantly increased the risk of NIHL for the first time in the Chinese Han population.

Since hearing restoration is not currently available, computational methods to identify susceptible individuals and novel candidate genes for early identification and prevention of high-risk individuals from working in noisy environments are necessary. ML algorithms are superior to traditional statistical methods owing to their flexibility and evolving performance for exploring the complex associations between risk factors and the development of NIHL. Although we expect the prediction accuracy of ML models

to be as high as possible, it is restricted when only using phenotypic data due to genetic factors. All four models performed well in our study, achieving similar accuracy and AUC, with a prediction accuracy of approximately 74%, while those who were misclassified by all four ML models had extreme phenotypes, namely, the degree of their hearing loss did not conform with their age, noise exposure level, and other factors, indicating that this novel strategy could be applied to distinguish high-risk individuals in noise-exposed populations.

Several studies employing an extreme phenotype design have been successful in identifying rare variants[39, 40]. In this study, we retrieved a wide range of candidate genes related to NIHL as previously reported for discovering novel variants. Other unknown genes may also contribute to the variation in susceptibility, but we have not listed them for a lack of definite evidence. It is recognized that replication of findings in independent populations is much more important than obtaining highly significant *p* values[41]. Thus, to avoid false-positive results, replication in independent sample sets and combined meta-analyses were performed. Despite significant phenotypic differences in our samples, most of the candidate loci showed negative results after adjustment. Notably, we identified the variant rs72474224 of *GJB2*, which is consistent with one recent study that the knock-in homozygous mice based on human p.V37I variant (c.109G > A) manifested as more vulnerable to noise damage[42], however, we failed to replicate this association in the validation samples. The *CDH23* gene has been reported to be linked to NIHL susceptibility[43, 44]. *CDH23* encodes cadherin 23, which is expressed in a variety of structures within the inner ear. In our study, the A allele of rs41281334 improved NIHL risk by almost 1.5-fold. Although the *WHRN* variants have not been reported to be associated with NIHL, the risk allele C of rs12339210 significantly increased NIHL risk by 3-fold even after adjustment in our study. The *WHRN* mRNA transcripts are expressed in the organ of Corti, vestibular, and retinal tissues, where mutations disturb the aid in assembling large multiprotein complexes, which play a role in maintaining stereocilia length[45], leading to autosomal recessive non-syndromic deafness type 31 (DFNB31) or Usher syndrome[46]. We speculate that the nonsynonymous mutations of rs41281334 and rs12339210 located in exons might influence the expression level of the genes but heterozygous carriers do not have profound congenital deafness initially due to the weak genetic efficacy. However, their inner ear may be more vulnerable to noise exposure and they may experience hearing loss earlier than a normal person.

Several limitations of this study should be mentioned. First, the data in this study was derived from cross-sectional data without long-term follow-up, which is critical to confirm the susceptibility of these workers. Second, the small sample size of the WES made it difficult to achieve genome-wide significance. In addition, all participants were from the Chinese Han population, causing conceivable inherent bias due to the impact of ethnic differences on genetic polymorphism, and the results should be validated among multiple ethnic populations.

Conclusions

The present study provides a novel strategy to evaluate the individual susceptibility of NIHL based on the prediction error of ML models and highlights an application for prescreening high-risk individuals from noise-exposed populations before gene sequencing. We also expanded the mutation spectrum of NIHL susceptibility genes and validated the association between *CDH23* rs41281334 and *WHRN* rs12339210 variants and NIHL susceptibility for the first time with human genetic evidence; this should be followed up in larger cohorts and verified by functional studies.

Abbreviations

NIHL, noise-induced hearing loss; SNP, single nucleotide polymorphism; ML, machine learning; CNE, cumulative noise exposure; PTA, pure-tone audiometry; EHF, extended high frequencies; WES, whole exome sequencing; AUC, area under the curve; ROC, receiver operator characteristic; HL, hearing level.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Ethics Review Board of the Shanghai Sixth People's Hospital affiliated with Shanghai Jiao Tong University (Approval No: 2017-136) and was registered in the Chinese Clinical Trial Registry (registration number: ChiCTR-

RPC-17012580). Potential consequences and benefits were explained, and written informed consent was obtained from each subject before this study.

Consent for publication

Not applicable.

Availability of data and materials

The data of this study are not publicly available due to the issue of intellectual property but are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Funding

This study was supported by the State Key Program of the National Natural Science Foundation of China (Grant No.81530029), the International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No.81720108010), National Natural Science Foundation of China (Grant No. 82071041/H1304), and Shanghai Municipal Commission of Science and Technology (Grant No.18DZ2260200).

Authors' contributions

Study conception and design: SKY and HW; Acquisition of data: ZJ, HW, DYS, XMZ, YMF, JPW; Analysis and interpretation of data: ZJ, BTF, HBS; Drafting of manuscript: ZJ, WH; Critical revision: SKY and HW. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge all the participants and institutions in this research.

References

1. Royster JD: Preventing Noise-Induced Hearing Loss. N C Med J 2017, 78(2):113-117.
2. Nelson DI, Nelson RY, Concha-Barrientos M, Fingerhut M: The global burden of occupational noise-induced hearing loss. Am J Ind Med 2005, 48(6):446-458.
3. Fuente A, Hickson L: Noise-induced hearing loss in Asia. Int J Audiol 2011, 50 Suppl 1:S3-10.
4. Sayler SK, Roberts BJ, Manning MA, Sun K, Neitzel RL: Patterns and trends in OSHA occupational noise exposure measurements from 1979 to 2013. Occup Environ Med 2018.
5. Hong O, Kerr MJ, Poling GL, Dhar S: Understanding and preventing noise-induced hearing loss. Dis Mon 2013, 59(4):110-118.
6. Annelies Konings LVL, and Guy Van Camp: Genetic Studies on Noise-Induced Hearing Loss A Review. Ear & Hearing 2009, 2009;30;151–159).
7. Sliwinska-Kowalska M, Pawelczyk M: Contribution of genetic factors to noise-induced hearing loss: A human studies review. Mutation Research/Reviews in Mutation Research 2013, 752(1):61-65.
8. Heinonen-Guzejev M, Vuorinen HS, Mussalo-Rauhamaa H, Heikkila K, Koskenvuo M, Kaprio J: Genetic component of noise sensitivity. Twin Res Hum Genet 2005, 8(3):245-249.
9. McFadden SL, Ohlemiller KK, Ding D, Shero M, Salvi RJ: The Influence of Superoxide Dismutase and Glutathione Peroxidase Deficiencies on Noise-Induced Hearing Loss in Mice. Noise Health 2001, 3(11):49-64.
10. Holme RH, Steel KP: Progressive hearing loss and increased susceptibility to noise-induced hearing loss in mice carrying a Cdh23 but not a Myo7a mutation. J Assoc Res Otolaryngol 2004, 5(1):66-79.
11. Fu X, Zhang L, Jin Y, Sun X, Zhang A, Wen Z, Zhou Y, Xia M, Gao J: Loss of Myh14 Increases Susceptibility to Noise-Induced Hearing Loss in CBA/CaJ Mice. Neural Plast 2016, 2016:6720420.

12. Miao L, Ji J, Wan L, Zhang J, Yin L, Pu Y: An overview of research trends and genetic polymorphisms for noise-induced hearing loss from 2009 to 2018. Environ Sci Pollut Res Int 2019, 26(34):34754-34774.
13. Churko JM, Mantalas GL, Snyder MP, Wu JC: Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. Circ Res 2013, 112(12):1613-1623.
14. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011, 12(11):745-755.
15. Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, Wright FA, Rieder MJ, Tabor HK, Nickerson DA *et al*: Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat Genet 2012, 44(8):886-889.
16. Shtir C, Aldahmesh MA, Al-Dahmash S, Abboud E, Alkuraya H, Abouammoh MA, Nowailaty SR, Al-Thubaiti G, Naim EA, B AL *et al*: Exome-based case-control association study using extreme phenotype design reveals novel candidates with protective effect in diabetic retinopathy. Hum Genet 2016, 135(2):193-200.
17. Miyakita T, Miura H, Yamamoto T: Evaluation of noise susceptibility: effects of noise exposure on acoustic reflex. Int Arch Occup Environ Health 1983, 52(3):231-242.
18. Lu J, Cheng X, Li Y, Zeng L, Zhao Y: Evaluation of individual susceptibility to noise-induced hearing loss in textile workers in China. Arch Environ Occup Health 2005, 60(6):287-294.
19. Rosenhall U: The influence of ageing on noise-induced hearing loss. Noise Health 2003, 5(20):47-53.
20. Wang D, Wang Z, Zhou M, Li W, He M, Zhang X, Guo H, Yuan J, Zhan Y, Zhang K *et al*: The combined effect of cigarette smoking and occupational noise exposure on hearing loss: evidence from the Dongfeng-Tongji Cohort Study. Sci Rep 2017, 7(1):11142.
21. Yang HY, Shie RH, Chen PC: Hearing loss in workers exposed to epoxy adhesives and noise: a cross-sectional study. BMJ Open 2016, 6(2):e010533.
22. Kim S, Lim EJ, Kim HS, Park JH, Jarng SS, Lee SH: Sex Differences in a Cross Sectional Study of Age-related Hearing Loss in Korean. Clin Exp Otorhinolaryngol 2010, 3(1):27-31.
23. MARIOLA ŚLIWIŃSKA-KOWALSKA AD, PIOTR KOTYŁO, EWA ZAMYSŁOWSKA-SZMYTKE, MAŁGORZATA PAWLACZYK-ŁUSZCZYŃSKA, and ANNA GAJDA-SZADKOWSKA: Individual susceptibility to noise-induced hearing loss choosing an optimal method of retrospective classification of workers in to noise-susceptible and noise-resistant groups. International Journal of Occupational Medicine and Environmental Health 2006, 19(4):235 – 45.
24. Coles RR, Lutman ME, Buffin JT: Guidelines on the diagnosis of noise-induced hearing loss for medicolegal purposes. Clin Otolaryngol Allied Sci 2000, 25(4):264-273.
25. Yang Q, Yu S, He L: Comparision of diagnostic criteria of occupational noiseinduced hearing loss in China and abroad. Zhonghua Lao Dong Wei Sheng Zhi Ye Bing Za Zhi 2015, 33(12):944-946.
26. Mehrparvar AH, Mirmohammadi SJ, Ghoreyshi A, Mollasadeghi A, Loukzadeh Z: High-frequency audiometry: a means for early diagnosis of noise-induced hearing loss. Noise Health 2011, 13(55):402-406.
27. Antonioli CA, Momensohn-Santos TM, Benaglia TA: High-frequency Audiometry Hearing on Monitoring of Individuals Exposed to Occupational Noise: A Systematic Review. Int Arch Otorhinolaryngol 2016, 20(3):281-289.
28. Sidey-Gibbons JAM, Sidey-Gibbons CJ: Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019, 19(1):64.
29. Ngiam KY, Khor IW: Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019, 20(5):e262-e273.
30. Zhao Y, Li J, Zhang M, Lu Y, Xie H, Tian Y, Qiu W: Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. Ear Hear 2018.
31. China: GBZ/T 189.8-2007 Measurement of noise in the workplace. 2007.
32. Xie HW, Qiu W, Heyer NJ, Zhang MB, Zhang P, Zhao YM, Hamernik RP: The Use of the Kurtosis-Adjusted Cumulative Noise Exposure Metric in Evaluating the Hearing Loss Risk for Complex Noise. Ear Hear 2016, 37(3):312-323.
33. Gutierrez-Tobal GC, Alvarez D, Del Campo F, Hornero R: Utility of AdaBoost to Detect Sleep Apnea-Hypopnea Syndrome From Single-Channel Airflow. IEEE Trans Biomed Eng 2016, 63(3):636-646.

34. Rossi F, Conan-Guez B: Functional multi-layer perceptron: a non-linear tool for functional data analysis. Neural Netw 2005, 18(1):45-60.
35. Yang F, Wang HZ, Mi H, Lin CD, Cai WW: Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics 2009, 10 Suppl 1:S22.
36. Maltarollo VG, Kronenberger T, Espinoza GZ, Oliveira PR, Honorio KM: Advances with support vector machines for novel drug discovery. Expert Opin Drug Discov 2019, 14(1):23-33.
37. Shearer AE, Hildebrand MS, Smith RJH: Hereditary Hearing Loss and Deafness Overview. In: *GeneReviews((R)).* Edited by Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Stephens K, Amemiya A. Seattle (WA); 1993.
38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007, 81(3):559-575.
39. Ung C, Sanchez AV, Shen L, Davoudi S, Ahmadi T, Navarro-Gomez D, Chen CJ, Hancock H, Penman A, Hoadley S *et al*: Whole exome sequencing identification of novel candidate genes in patients with proliferative diabetic retinopathy. Vision Res 2017, 139:168-176.
40. Emond MJ, Louie T, Emerson J, Chong JX, Mathias RA, Knowles MR, Rieder MJ, Tabor HK, Nickerson DA, Barnes KC *et al*: Exome Sequencing of Phenotypic Extremes Identifies CAV2 and TMC6 as Interacting Modifiers of Chronic Pseudomonas aeruginosa Infection in Cystic Fibrosis. PLoS Genet 2015, 11(6):e1005273.
41. Neale BM, Sham PC: The future of association studies: gene-based analysis and replication. Am J Hum Genet 2004, 75(3):353-362.
42. Lin X, Li G, Zhang Y, Zhao J, Lu J, Gao Y, Liu H, Li GL, Yang T, Song L *et al*: Hearing consequences in Gjb2 knock-in mice: implications for human p.V37I mutation. Aging (Albany NY) 2019, 11(18):7416-7441.
43. Kowalski TJ, Pawelczyk M, Rajkowska E, Dudarewicz A, Sliwinska-Kowalska M: Genetic variants of CDH23 associated with noise-induced hearing loss. Otol Neurotol 2014, 35(2):358-365.
44. Sliwinska-Kowalska M, Noben-Trauth K, Pawelczyk M, Kowalski TJ: Single nucleotide polymorphisms in the cadherin 23 (CDH23) gene in Polish workers exposed to industrial noise. Am J Hum Biol 2008, 20(4):481-483.
45. Mustapha M, Beyer LA, Izumikawa M, Swiderski DL, Dolan DF, Raphael Y, Camper SA: Whirler mutant hair cells have less severe pathology than shaker 2 or double mutants. J Assoc Res Otolaryngol 2007, 8(3):329-337.
46. Mathur PD, Yang J: Usher syndrome and non-syndromic deafness: Functions of different whirlin isoforms in the cochlea, vestibular organs, and retina. Hear Res 2019, 375:14-24.

Figures

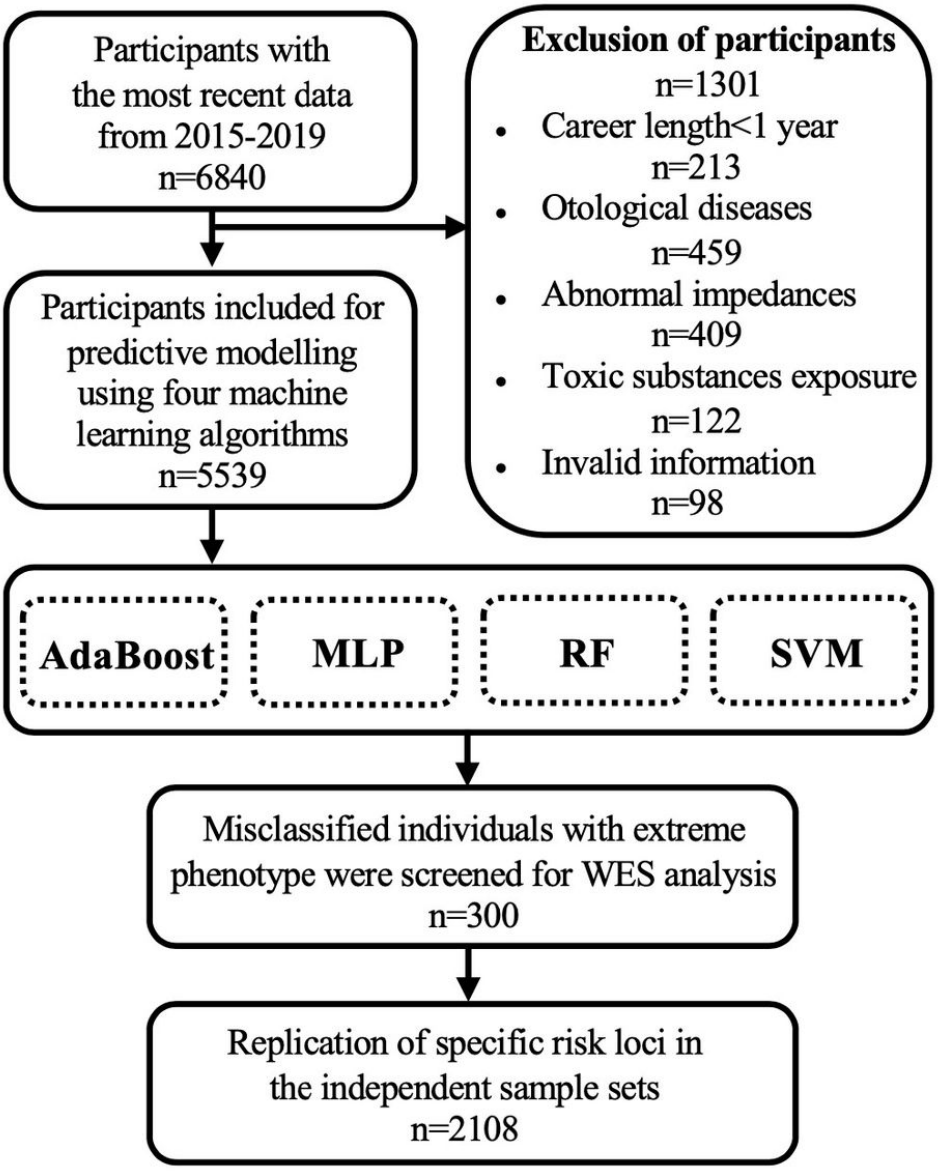

Figure 1

The flow chart of study and exclusion criteria.

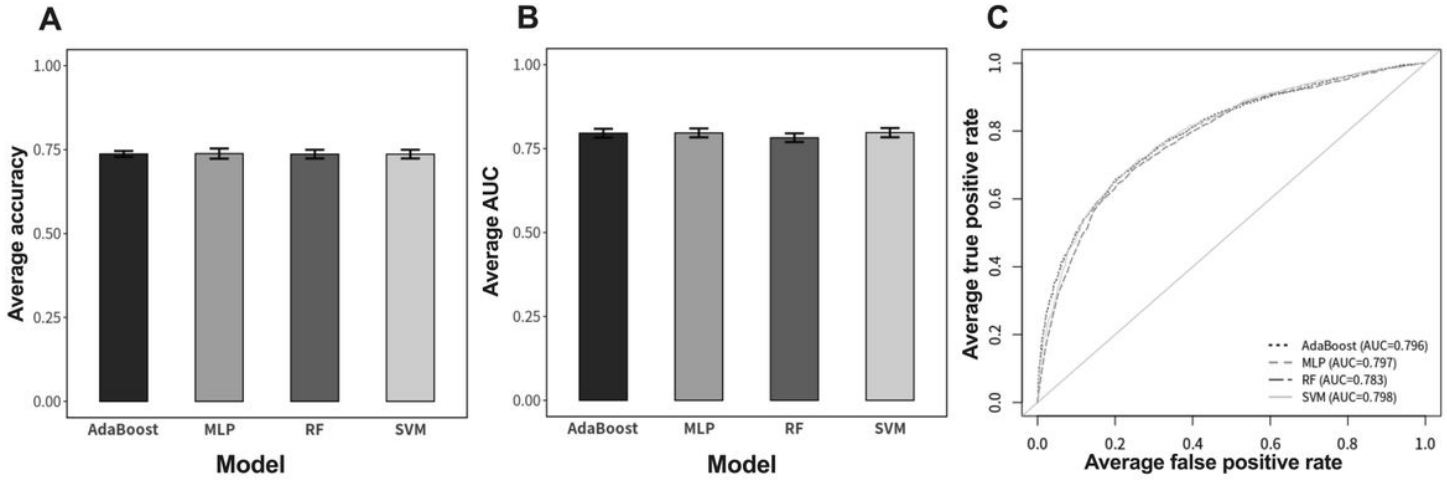

Figure 2

Comparison of the prediction performances of the four machine learning models. A shown the average accuracy and B shown the average AUC value of 10-fold cross-validation. The error bars represent the standard deviations. C shown comparison of area under the ROC curves. Adaboost, adaptive boosting; MLP, multilayer perceptron; RF, random forest; SVM, support vector machine.

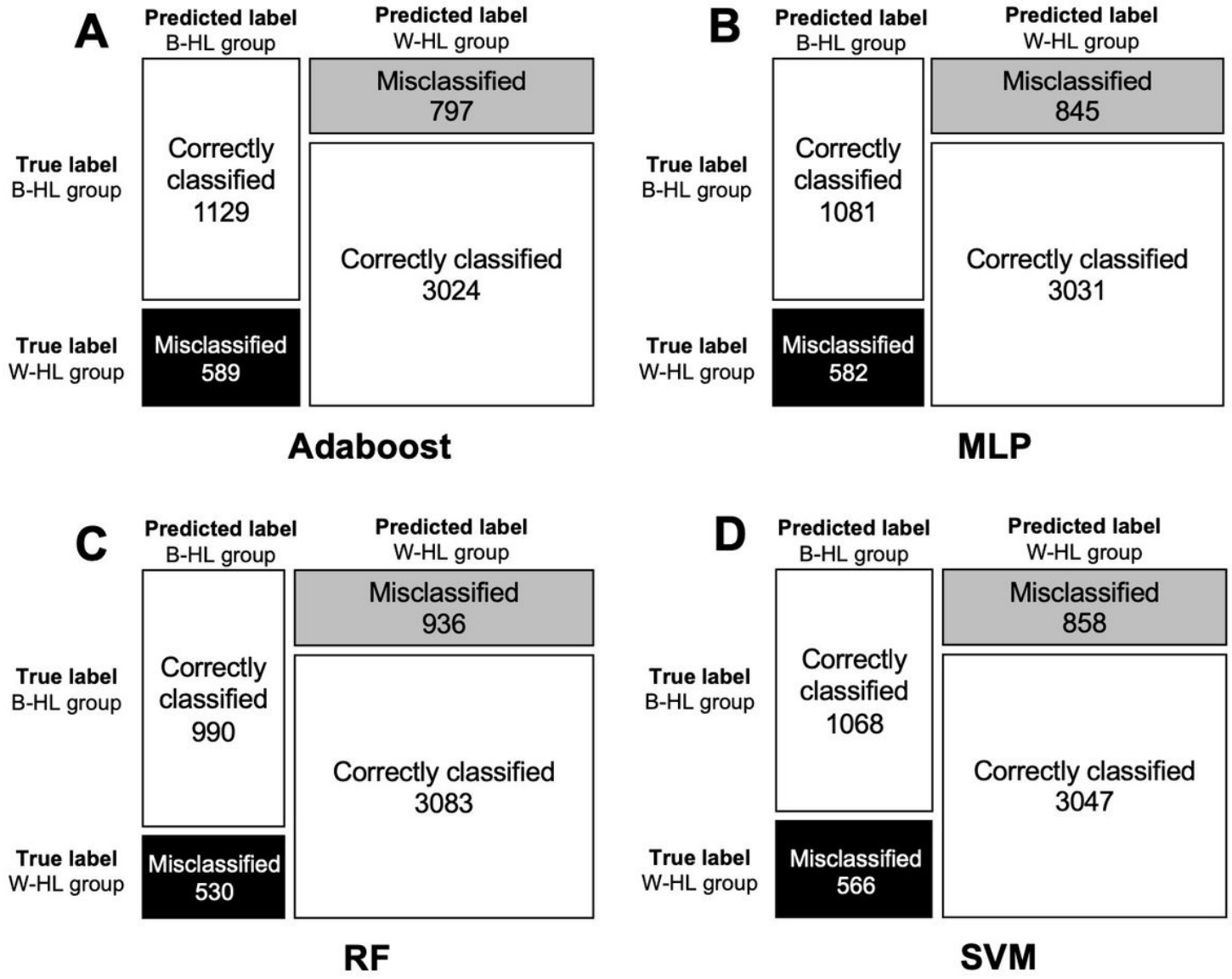

Figure 3

Confusion matrix of the predictions of the 5,539 participants. The number in each cell represents the number of people classified by the model. The black colored lattice represents the susceptibility status, while the gray colored lattice represents the resistance status. Adaboost, adaptive boosting; MLP, multilayer perceptron; RF, random forest; SVM, support vector machine.

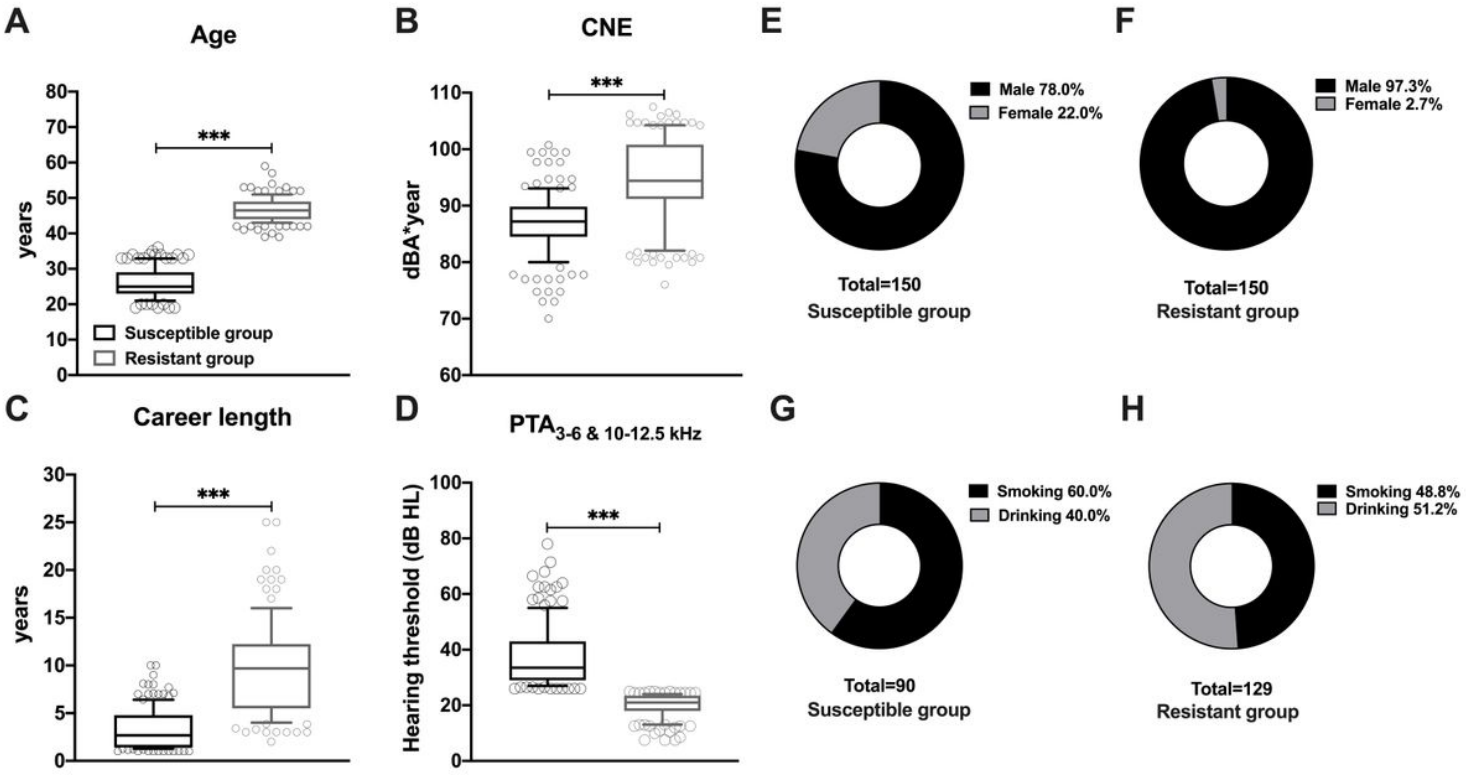

Figure 4

Comparison of the characteristics of the NIHL-susceptible group and NIHL-resistant group. A-D shown the difference of age, CNE, career length, and mean PTA of 3, 4, and 6 kHz plus 10 and 12.5 kHz of the two groups. The lower and upper whiskers of the boxplot are the 10% to 90% percentiles, respectively. The box represents the upper quartile, the median, and the lower quartile. E-H show the percentage of sex and smoking or drinking of the two groups.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx