Discover Oncology

Analysis

Mendelian randomization analysis of environmental pollution factors and head and neck cancer risk: a causal inference study integrating autophagy-related genes

Xuejin Xu[1] · **Zhen Wang**[1]

Received: 5 August 2024 / Accepted: 21 February 2025
Published online: 11 March 2025

Abstract

Background Head and neck squamous cell carcinoma (HNSC) is a significant global health challenge. While traditional risk factors are well-established, the role of environmental pollutants in HNSC development remains unclear.

Objective To investigate the causal relationship between environmental pollution factors and HNSC risk using Mendelian Randomization (MR) analysis.

Methods Two-sample MR analysis was performed using genome-wide association study data from the IEU OpenGWAS project and HNSC RNA-seq data from TCGA. Environmental pollution-associated genes (MRGs) were identified and analyzed along with autophagy-related genes (ATGs) in HNSC samples. Cox proportional hazards models were used to develop a clinical prediction model.

Results MR analysis revealed significant causal relationships between nitrogen dioxide air pollution, nitrogen oxides air pollution, PM2.5, and increased HNSC risk. Nine MRGs were identified, with four (IRF4, LINGO1, PTHLH, RSRC1) differentially expressed in HNSC. A six-factor clinical prediction model (IRF4, LINGO1, PTHLH, RSRC1, Age, USP10) showed good predictive performance for HNSC survival (C-index = 0.63, 10-year AUC = 0.761). Tumor mutation burden and immune cell infiltration analyses provided further insights into HNSC biology.

Conclusion This study provides evidence for causal relationships between specific air pollutants and HNSC risk, and identifies potential gene targets for further investigation. The developed clinical prediction model may aid in HNSC prognosis and personalized treatment strategies.

Keywords Head and neck squamous cell carcinoma · Environmental pollution · Mendelian Randomization · Autophagy · Clinical prediction model · Tumor mutation burden · Immune infiltration

1 Introduction

Head and neck squamous cell carcinoma (HNSC) represents a significant global health challenge, ranking as the sixth most common cancer worldwide. With an estimated annual incidence of over 900,000 cases and 450,000 deaths, HNSC poses a substantial burden on healthcare systems and patient quality of life. Despite advancements in diagnostic techniques and treatment modalities, the five-year survival rate for HNSC patients remains unsatisfactory, hovering around 50–60%. This grim prognosis underscores the urgent need for a deeper understanding of HNSC etiology and risk factors. While traditional

✉ Zhen Wang, wang1121325@wmu.edu.cn; Xuejin Xu, xuxj1124@wmu.edu.cn | [1]Department of Stomatology, The Quzhou Affiliated Hospital of Wenzhou Medical University (Quzhou People's Hospital), Kecheng District, Minjiang Avenue No. 100, Quzhou 332400, Zhejiang Province, China.

risk factors such as tobacco use and alcohol consumption are well-established, the role of environmental pollutants in HNSC development and progression has gained increasing attention in recent years.

There is substantial evidence linking environmental pollutants to the development of head and neck cancer (HNC). Occupational exposure to diesel exhaust (DE) has been specifically associated with an increased risk of HNC, particularly laryngeal cancer, as demonstrated by a systematic review and meta-analysis which found a summary relative risk (RR) of 1.08 for HNC overall and 1.13 for laryngeal cancer [1]. Additionally, pollutants such as heavy metals, pesticides, and other toxicants have been implicated in cancer progression through mechanisms involving cytokine level alterations and cell damage [2]. The role of cadmium (Cd) as a carcinogen is particularly noteworthy, with studies indicating significantly higher cadmium levels in nasopharyngeal and pharyngeal cancer patients compared to controls, although this was not observed for laryngeal cancer [3]. Environmental pollutants, including bisphenol A (BPA), benzo[a]pyrene (BaP), and persistent organic pollutants (POPs), have also been shown to reduce the efficacy of chemotherapeutic drugs, potentially complicating cancer treatment [4]. The interaction between genetic factors and environmental components further exacerbates the risk, as pollutants can modulate the activity of biological molecules and contribute to tumorigenesis, particularly in industrialized regions where exposure is more prevalent [5, 6]. Moreover, the synergistic interactions between different pollutants, even at low concentrations, can lead to significant health issues, including cancer, by activating key biological pathways such as reactive oxygen species production and cytochrome P450 activation [7]. The persistence and bioaccumulation of metal(loid)s and radionuclides, which are nonbiodegradable, pose additional risks, as these substances can be biomagnified along the food chain, leading to severe health problems, including cancer [8]. Overall, the evidence underscores the critical need for controlling environmental pollutant exposure to mitigate the risk of HNC and improve public health outcomes.

Although previous studies have explored the association between environmental pollution and various diseases, there remains a significant research gap regarding the causal relationship between environmental factors and HNSC. Traditional epidemiological studies are often limited by confounding factors and reverse causality, making it challenging to establish reliable causal inferences. To overcome these limitations, this study innovatively employs the Mendelian Randomization (MR) method to investigate the causal relationship between environmental pollution factors and HNSC. The MR method uses genetic variants as instrumental variables, effectively controlling for potential confounding factors and reverse causality, thereby providing more reliable causal inferences. This approach not only reveals potential causal links between environmental pollution and HNSC risk but also quantifies the strength of these relationships. Furthermore, by integrating GWAS data and gene expression information from the TCGA database, we are able to identify key genes associated with environmental pollution (MRGs) and analyze their roles in the development and progression of HNSC in depth, providing important clues for understanding disease mechanisms and developing new therapeutic strategies. This multidimensional research approach represents a cutting-edge integration of environmental epidemiology and molecular biology, with the potential to bring new breakthroughs in the prevention and treatment of HNSC.

This study aims to investigate the causal relationship between environmental pollution factors and HNSC risk using the MR method. By analyzing genome-wide association study (GWAS) data from the IEU OpenGWAS project and HNSC RNA-seq datasets from TCGA-GDC, we will systematically evaluate the impact of various environmental pollutants on HNSC occurrence. Additionally, we will identify key genes associated with environmental pollution and analyze their expression patterns and interrelationships in HNSC. This research not only helps to elucidate the association between environmental pollution and HNSC but may also provide new insights and targets for HNSC prevention, early diagnosis, and personalized treatment.

2 Materials and methods

2.1 Data sources

Data from the GWAS study of environmental contamination and HNSC were obtained from the IEU OpenGWAS program website (https://gwas.mrcieu.ac.uk/). A total of 12 datasets were obtained, including 1 HNSC dataset and 11 environmental pollution datasets, and the testers in the 12 data sets were from Europe. The basic information statistics of the 12 data sets are shown in Table 1.

The RNA-seq dataset of HNSC was downloaded from TCGA-GDC (https://portal.gdc.cancer.gov/), and the type of expression values in the samples was selected as Transcripts Per Million (TPM). The HNSC expression matrix consists of 515 tumor samples and 44 normal samples. After integrating the sample expression matrix, we downloaded the relevant clinical information dataset and SNV (simple nucleotide variation) dataset of HNSC from TCGA, in which the data type of SNV dataset was Masked Somatic Mutation. The 515 tumor samples and HNSC clinical data were combined to obtain a sample expression matrix of 509 Tumors with complete Overall survival (OS). The 509 Tumor samples was characterized (Table 2).

After integrating the dataset, we processed the HNSC expression profiles (509 tumor samples and 44 normal samples). The genes with GeneSymbol duplicates were first de-weighted for mean value. Then we found that there were large differences in gene expression values in the same sample, so we performed log2 scaling on the dataset. We also found that some genes in the dataset had expression values of 0 in many samples, and the significance of the existence of these genes was not high. Therefore, if the number of samples belonging to a gene for which the expression value is 0 was more than 50% of the total number of samples, this gene was removed from the dataset. Finally, the samples co-occurring in the expression profile and clinical information dataset were extracted, and the final expression profile dataset was obtained based on these samples. Autophagy related genes (ATGs) were downloaded from the Human Autophagy Database (http://www.autophagy.lu/) and 222 ATGs were obtained.

2.2 M&M section I-Mendelian Randomization (MR) analysis

2.2.1 Two-sample MR analysis process

Mendelian Randomization (MR) is an experimental design method that is widely used in the field of epidemiology. The causal relationship between exposure factors and outcomes is analyzed by introducing an intermediate variable called an instrumental variable, which was Single Nucleotide Polymorphism (SNP) in this study.

In order to ensure the validity of the results, MR studies require three assumptions: 1: The assumption of association, that there is a strong correlation between the SNP (Single Nucleotide Polymorphism) and the exposure factor. 2: The assumption of independence, that the SNP is independent of the any confounding factors. 3: The assumption of exclusivity, that the SNP can only have an effect on the outcome through the exposure factor.

To investigate the causal relationship between environmental pollution factors and HNSC, we used the GWAS dataset from the IEU database using two-sample Mendelian randomization. Two-sample MR was used to explore whether environmental pollution factors influence HNSC occurrence, where SNPs highly correlated with environmental pollution factors were defined as instrumental variables for exposure factors, and information about these SNPs in HNSC was defined as outcome variables.

2.2.2 Parameters of genetic instrument variables

In this study, the screening criteria for instrumental variables of exposure were as follows: for the GWAS data of each environmental pollution factor-related trait, we chose $P < 1 \times 10^{-6}$ of SNPs in the data as the primary screening condition. If a small number of SNPs were obtained, the threshold of P-value could be relaxed to $P < 5 \times 10^{-6}$. SNPs in linkage disequilibrium (SNPs with $r^2 < 0.001$ and physical distance between every two genes > 10,000 kb) were excluded. For MR

Table 1 Basic information statistics of HNSC and environmental pollution factors

Trait	GWAS ID	Case	Control	Sample size	Number of SNPs
Head and neck cancer (HNSC)	ieu-b-4912	1106	372,016	373,122	9,655,080
Particulate matter air pollution (pm2.5)	ukb-b-10817	NA	NA	423,796	9,851,867
Particulate matter air pollution (pm10)	ukb-b-18469	NA	NA	423,796	9,851,867
Nitrogen oxides air pollution	ukb-b-12417	NA	NA	456,380	9,851,867
Nitrogen dioxide air pollution	ukb-b-9942	NA	NA	456,380	9,851,867
Workplace very dusty: Often	ukb-d-22609_2	9561	80,070	89,631	13,188,866
Workplace full of chemical or other fumes: Often	ukb-d-22610_2	5872	82,863	88,735	12,324,484
Workplace had a lot of cigarette smoke from other people smoking: Often	ukb-d-22611_2	14,941	74,862	89,803	13,566,864
Worked with materials containing asbestos: Often	ukb-d-22612_2	1262	75,195	76,457	9,680,277
Worked with paints, thinners or glues: Often	ukb-d-22613_2	3723	84,995	88,718	11,525,313
Worked with pesticides: Sometimes	ukb-d-22614_1	3385	84,756	88,141	11,359,743
Workplace had a lot of diesel exhaust: Often	ukb-d-22615_2	3483	85,621	89,104	11,409,239

Table 2 Statistics of clinical characteristics of HNSC

Trait	Group	Sample size	Total
Overall survival (OS)	Between 0 and 3 years	425 (83.7%)	509
	Between 0 and 5 years	473 (92.9%)	
	Between 0 and 10 years	499 (98.04%)	
Event	Alive	344 (67.6%)	
	Dead	165 (32.4%)	
Gender	Female	135 (26.5%)	
	Male	374 (73.5%)	
Age	Old(Age ≥ 60)	282 (55.4%)	
	Young (Age < 60)	227 (44.6%)	
Stages	Stage I	20 (3.9%)	
	Stage II	96 (18.9%)	
	Stage III	104 (20.4%)	
	Stage IV	276 (54.2%)	
	Other	13 (2.6%)	
	Total (Stage I–Stage IV)	496 (97.4%)	
Stages_T	T1	35 (6.9%)	
	T2	149 (29.3%)	
	T3	135 (26.5%)	
	T4	179 (35.2%)	
	Other	11 (2.2%)	
	Total (T1–T4)	498 (97.8%)	
Stages_M	M0	484 (95.1%)	
	M1	5 (1%)	
	Other	20 (3.9%)	
	Total (M0–M1)	489 (96.1%)	
Stages_N	N0	243 (47.7%)	
	N1	83 (16.3%)	
	N2	158 (31%)	
	N3	7 (1.4%)	
	Other	18 (3.5%)	
	Total (N0–N3)	498 (97.8%)	
Grade	G1	62 (12.2%)	
	G2	296 (58.2%)	
	G3	122 (24%)	
	G4	7 (1.4%)	
	Other	22 (4.3%)	
	Total (G1–G4)	487 (95.7%)	

analysis, it was necessary to harmonize the effects of SNPs on outcome and exposure factors to be relative to the same allele. The F statistic was then calculated for SNPs obtained from the harmonizaiton analysis to assess weak instrumental variable bias. When $F < 10$, it suggests that the genetic variant used is a weak instrumental variable that may bias the results to some extent, and it is then excluded to avoid affecting the results. The F statistic was calculated using the following formula:

$$F = (N - k - 1)/k \times \frac{R^2}{1 - R^2}$$

where n is the sample size, k is the number of instrumental variables used, and R^2 reflects the extent to which the instrumental variables explain the exposure. $R^2 = 2 \times (1\text{-MAF}) \times \text{MAF} \times \beta 2$, where MAF is the minimum allele frequency and β is the allele effect value. Finally, SNPs with F > 10 were subjected to MR-PRESSO analysis, and SNPs that did not appear as outliers in the MR-PRESSO analysis were included in the subsequent analysis.

2.2.3 MR causal effect estimates

We assessed the causal effect of the exposure instrumental variables on the outcome using a variety of two-sample Mendelian randomization methods, including: inverse-variance weighted (IVW), MR-Egger (Mendelian randomization-Egger), weighted median (Weighted Median), simple mode (simple mode), and weighted mode (weighted mode). It has been shown that the IVW method is slightly stronger than the others under certain conditions; it is characterized by regression that does not take into account the presence of an intercept term and is fitted with the inverse of the outcome variance as weights. Therefore, in the absence of pleiotropy, regardless of the presence of heterogeneity, the IVW method was used in this study as the primary MR analysis method, with the other four methods as supplements.

2.2.4 Sensitivity analysis

The sensitivity analysis of the analytical results was performed by using various methods such as heterogeneity test, multiple validity test and leave-one-out test as follows:

(1) Heterogeneity test

Heterogeneity between individual SNP estimates was assessed using the Cochran Q test, and significant heterogeneity in the analysis was demonstrated if the Cochran Q test was statistically significant. Heterogeneity was assessed by IVW and MR-Egger tests, and a p-value < 0.05 indicated that there was heterogeneity in the study, which could be caused by the large differences between the p values of the SNPs in the dataset of the outcome variables. Therefore, it is more appropriate for the analysis of the results to find no heterogeneity in the data (p value > 0.05).

(2) Steiger test: When analyzing the causal relationship between exposure factors and outcomes, SNPs should correlate more with exposure factors than with outcomes, otherwise reverse causality will result, i.e., SNPs cannot be used as instrumental variables for exposure. Due to measurement error in GWAS data, some SNP measurements may not be what they appear to be in GWAS data. Therefore, it is possible that after the GWAS data are screened and processed, the phenomenon that SNPs are less correlated with exposure factors than with outcomes is present, which leads to the hypothesis that exposure leads to outcomes being invalid. Using the Steiger test, we can statistically test the validity of the hypothesis that exposure leads to outcome. If the P value of the Steiger test is < 0.05, this indicates that our analysis is in the right direction.

(3) Pleiotropy test

Two methods, MR-Egger and MR-PRESSO, were used to test the instrumental variables for pleiotropy. For the MR-Egger method, if the p-value of the MR-Egger intercept is < 0.05, it indicates that there is significant horizontal pleiotropy for the genetic variables. For the MR-PRESSO method, if the P value of the global test is < 0.05, it indicates that there is significant horizontal pleiotropy in the genetic variables. If there is no pleiotropy in both methods (P value > 0.05), the analysis is more reliable. If outliers appear in MR-PRESSO analysis, it is necessary to extract the data after removing these outliers, and then re-perform the MR correlation analysis.

(4) Leave-one-out test

Using IVW analysis, MR results for the remaining instrumental variables were calculated by excluding individual SNPs one at a time to assess whether individual SNPs influenced the association between environmental pollution factors and HNSC. If the difference between the MR effect estimates and the aggregate effect estimates after excluding an instrumental variable is large, it means that the MR effect estimates are sensitive to that SNP.

2.2.5 MRG access

If the analysis results showed that there was a causal effect between the exposure factors and the outcome, we extracted the significant SNPs among them, and annotated these SNPs from Ensemble VEP (https://grch37.ensembl.org/Multi/Tools/VEP?db=core) to obtain their corresponding genes and near-neighbor genes, where the Upstream/Downstream distance (bp) parameter used the default value set by the database.

After obtaining the corresponding genes of SNPs through VEP annotation, for each exposure factor that had a causal effect with HNSC, we extracted the expression matrix of these exposure factor-related genes in HNSC separately. The results of the leave-one-out test were combined with the genes present in the expression matrix to screen the genes, and the selected genes were labeled as Mendelian Randomization genes (MRGs) for inclusion in the subsequent analysis.

2.2.6 Statistical analyses of MR

All data calculations and statistical analyses in this study were carried out using R programming (https://www.r-project.org/, version 4.2.2). During MR analysis, MR was carried out using the TwoSampleMR package, the MR-PRESSO package was used for the pleiotropy test, the heatmap (Fig. 1A) was drawn using the circlize and ComplexHeatmap packages, and the forest plot (Fig. 1B) was generated using the forestploter package.

All statistical P values were from two-sided tests, SNP loci generated by the GWAS study were considered statistically significant at $P < 5 \times 10^{-6}$, and other statistical tests were considered statistically significant at $P < 0.05$.

2.3 M&M section II-study of MRGs and ATGs based on the HNSC expression matrix

2.3.1 MRG expression level analysis and relationship prediction

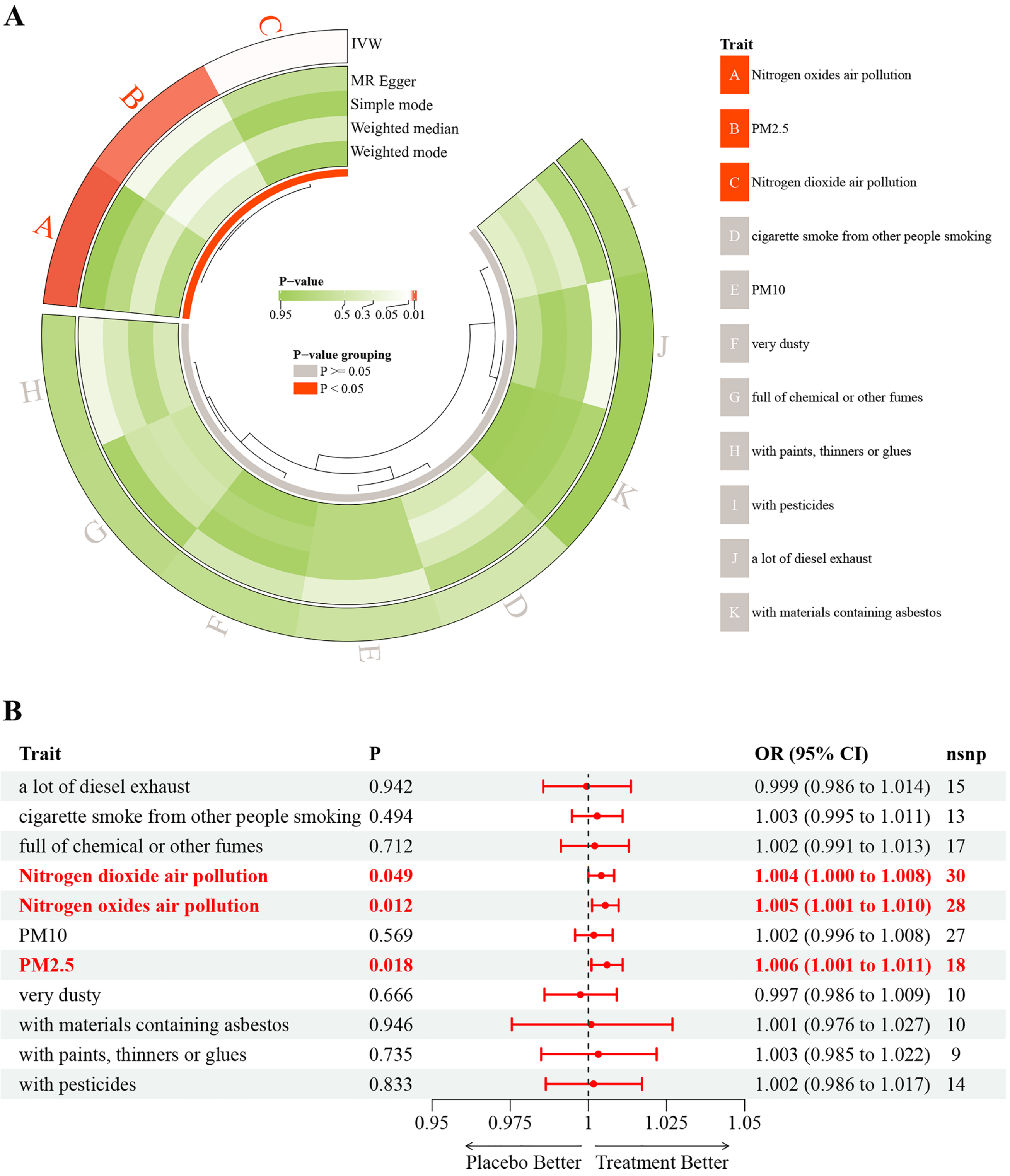
Heatmap was used to demonstrate the expression levels of these MRGs in HNSC. Box plots were used to demonstrate the expression distribution of these MRGs in HNSC using the Wilcoxon test to access MRG variability between tumor samples and normal samples. The Tumor sample expression matrix was extracted from the HNSC expression matrix and the correlation between MRGs was caculated using the Pearson correlation coefficient.

2.3.2 Differential expression (DE) analysis, marker MR-DEG, AT-DEG screening

Based on the HNSC expression matrices of tumor and normal samples, the "limma" package was used to perform differential expression analysis of the HNSC expression matrix, where the comparison was tumor versus normal. Differentially expressed genes (DEGs) were screened based on the results of the limma analysis. The DEGs were selected based on the P-value (or P.adjust) and | log2(fold change)| indicators in the differential analysis results, which were typically selected according to empirical values (the commonly used selection criteria were P-value < 0.05, | log2(FC)|> 1), which could be adjusted according to the actual analysis requirements. Therefore, we chose genes with P value < 0.05, |log2(FC)|> 0.5 as DEGs. The intersections of MRGs and ATGs with DEGs were determined separately, where the intersection of MRGs and DEGs was labeled as Marker MR-DEG, and the intersection of ATGs and DEGs was labeled as AT-DEG.

2.3.3 Marker MR-DEG and AT-DEG enrichment analysis

Integration of Marker MR-DEG and AT-DEG related genes, GO Biological process and KEGG pathway analyses of these genes were carried out using the clusterProfiler package in R, to observe the biological processes affected by the target genes. GO Biological process and KEGG pathways with p.adjust < 0.05 were considered significantly associated with the target genes.



Trait	P	OR (95% CI)	nsnp
a lot of diesel exhaust	0.942	0.999 (0.986 to 1.014)	15
cigarette smoke from other people smoking	0.494	1.003 (0.995 to 1.011)	13
full of chemical or other fumes	0.712	1.002 (0.991 to 1.013)	17
Nitrogen dioxide air pollution	**0.049**	**1.004 (1.000 to 1.008)**	**30**
Nitrogen oxides air pollution	**0.012**	**1.005 (1.001 to 1.010)**	**28**
PM10	0.569	1.002 (0.996 to 1.008)	27
PM2.5	**0.018**	**1.006 (1.001 to 1.011)**	**18**
very dusty	0.666	0.997 (0.986 to 1.009)	10
with materials containing asbestos	0.946	1.001 (0.976 to 1.027)	10
with paints, thinners or glues	0.735	1.003 (0.985 to 1.022)	9
with pesticides	0.833	1.002 (0.986 to 1.017)	14

Fig. 1 Presentation of MR analysis results of traits related to environmental pollution factors. **A** MR analysis results of environmental pollution factors; **B** HR forest plot presentation for environmental pollution factors

2.3.4 Marker AT-DEG screening and expression level analysis

Based on the expression matrix of tumor samples in HNSC, the correlation between Marker MR-DEG and AT-DEG was calculated using the Pearson correlation coefficient, and relationship pairs with correlation > 0.6 were selected as highly correlated Marker MR-DEG&AT-DEG relationship pairs. These highly correlated AT-DEGs with Marker MR-DEG were extracted for inclusion in the subsequent analysis, and these AT-DEGs were labeled as Marker AT-DEGs. Heatmaps demonstrated the expression levels of these Marker AT-DEGs in HNSC. Box plots were used to demonstrate the expression distribution of these Marker AT-DEG in HNSC using the Wilcoxon test to access MRG AT-DEG variability between tumor samples and normal samples.

2.3.5 Prediction of marker MR-DEG and marker AT-DEG relationships

The correlation between Marker AT-DEG was calculated using the Pearson correlation coefficient based on the expression matrix of tumor samples in HNSC. We demonstrated the relationship between Marker AT-DEG using the R language ggcor package, and also demonstrated the relationship between Marker MR-DEG and Marker AT-DEG.

We extracted the pathways that were significantly related to Marker AT-DEG from the results of KEGG pathway enrichment analysis, obtained the related genes under all pathways from the KEGG database (https://www.kegg.jp/), screened the DEGs corresponding to the significantly related pathway, and labeled these DEGs as bridge DEGs. We searched for the associations existing between Marker MR-DEGs, pathways, and bridge DEGs. Then, we extracted the pathways corresponding to the bridge DEGs, extracted the pathways that were significantly correlated with Marker AT-DEG based on KEGG pathway enrichment analysis, and searched for the associations between Marker MR-DEG, pathways, bridge DEG, and Marker AT-DEGs. We observed the potential connection between Marker MR-DEG and Marker AT-DEG based on pathways.

2.4 M&M section III-Cox-PH clinical predictive modeling and survival analysis

2.4.1 Selection of features relevant to clinical prediction models

First, since MR analysis was the core analysis of this study, we included all Marker MR-DEGs as traits for clinical prediction models in the subsequent analysis. Then, for Marker AT-DEG that were significantly correlated with Marker MR-DEG, we built a univariate Cox proportional hazards regression model (Cox-PH) for each Marker AT-DEG characterization factor based on the clinical characteristics OS and Survival Status using the survival package in the R language. The characteristic factors with P-value < 0.05 in the univariate analysis results were selected to be included in the subsequent analysis as traits for the clinical prediction model. Finally, for the clinical characterization data of HNSC, a univariate Cox-PH model was also constructed and the characteristic factors with P-value < 0.05 in the analysis results were selected to be included in the subsequent analysis as traits for the clinical prediction model.

2.4.2 Multivariate analysis

After obtaining the required features for the model, the Cox-PH model was used for multivariate analysis. The multivariate Cox-PH model could predict the Risk Score of the tumor patients based on the overall expression level of all the features. HNSC patients were classified into a High risk group and a Low risk group based on the median of their risk scores. group). Based on different Risk Score groupings, box plots were used to show the distribution of expression of these characteristics in HNSC, and the Wilcoxon test was used to analyze the variability of each trait between the High risk group and the Low risk group. Survival analysis was performed for patient risk scores within 4 time points: 3 years, 5 years, 10 years, and all samples, to examine the variability in survival between the High risk group and the Low risk group within different time points.

2.4.3 Establishment of a clinical prediction model nomogram

Sample risk scores were obtained by multivariate analysis. Survival analysis found that risk scores and survival were significantly correlated, and the risk scores were included as a new characteristic factor among the characteristic factors that were significantly correlated with survival to build the final Cox-PH clinical prediction model. The R language rms package was used to construct a nomogram plot to show the effect of the characteristic factors on survival, while the R language nomogramFormula package was used to obtain all sample Total Points in the nomogram plot.

2.4.4 Robustness assessment of clinical prediction model

In order to verify the prediction effect of the Cox-PH clinical prediction model, the calibration curves of the Cox-PH model for 3 years, 5 years and 10 years were firstly calculated using the calibrate method in the rms package of R. Then the C-index value of the model was examined, and the method of calculating C-index was C-index = 1-C, where the C value was obtained using the rms package "predict" method to predict the model and obtain the sample scores, using the R language Hmisc package "rcorrcens" method on the model scores to obtain the C value. We finally used the R language timeROC package "timeROC" method to calculate the ROC of the model at 3 years, 5 years and 10 years and drew the time-dependent ROC curves to verify the model effect.

2.4.5 Decision Curve Analysis (DCA)

Clinical Decision Curve Analysis (DCA) is a simple method for evaluating clinical prediction models, diagnostic tests, and molecular markers. The trait, Risk Score, and Total Points used for integrating clinical prediction models are incorporated into DCA, through which the relationship between High Risk Threshold and Net Benefit can be found for each characteristic factor. The Clinical Impact Curve (CIC) was also used to demonstrate the feature factors with good Net Benefit performance, providing better model selection for clinical decision making.

2.5 M&M section IV-the other analysis of MRGs

2.5.1 Tumor Mutation Burden (TMB) analysis

The tumor mutational burden (TMB) is the total number of somatic mutations with substitutions, insertions/deletions per Mb of bases in the coding region of exons on the genome in a tumor sample. The TMB is calculated as the total number of somatic mutations (including non-synonymous point mutations, insertions, and deletions in the coding region of the exon) divided by the total length of the exon in mutations/ Mb. The more mutations in the tumor tissue, the more likely it is that more abnormal proteins will be produced, and the more likely the tumor will have a systemic impact. The more mutated genes there are in the tumor tissue, the more likely it is that more abnormal proteins will be produced, and the greater impact the tumor may have on the whole body. At the same time, the more easily these abnormal proteins can be recognized by the immune system, thus activating the body's anti-cancer immune response, and therefore the better the efficacy of immunotherapy for tumors.

The SNV dataset in HNSC was analyzed for mutations and TMB scores were calculated for samples in HNSC. Based on the survival analysis High risk group and Low risk group, box plots were used to show the expression distribution of TMB in HNSC, and the Wilcoxon test was used to show the variability between different risk groups. The Pearson correlation coefficient was used to assess the relationship between survival related traits and TMB.

2.5.2 CIBERSORT immuno-infiltration analysis

CIBERSORTx (https://cibersortx.stanford.edu/) uses gene expression data to infer cell type abundance in mixed cell populations. The method analyzes gene expression profiles based on a known reference dataset, the official gene expression signature set for 22 immune cell subtypes: LM22. CIBERSORT analysis was performed based on the HNSC expression matrix and the LM22 dataset to obtain the cell abundance score expression matrix for each of the 22 immune cell subtypes in each sample, and additionally the significance and correlation between each sample and the corresponding sample of the expression matrix (P value and Correlation) were included in the results. Samples with P value < 0.05 were selected as samples with significant correlation between the immune cell feature set matrix and the PI expression matrix.

Based on the CIBERSORT Cell Abundance Score Expression Matrix, heatmaps were first used to view the scores of 22 immune cells in HNSC samples, and violin plots were used to show the distribution of scores of immune cells between tumor and normal groups, and the Wilcoxon test was used to identify immune cells that differed between sample types. Then we extracted immune cells with significantly different Wilcoxon test results in the tumor group for abundance scores and used the Pearson correlation coefficient to assess the relationship between these immune cells and all immune cells, as well as to assess the relationship between these immune cells and survival-related traits. Finally we analyzed the High risk group and the Low risk group based on survival and used cloud rain plots to observe the distribution of abundance scores of these significantly different immune cells in HNSC, using the Wilcoxon test to assess variability between different risk groups.

3 Results

3.1 MR assessment of the impact of environmental pollution factors on the risk of developing HNSC

3.1.1 MR analysis results

The GWAS data related to environmental pollutants were screened according to the screening criteria of the instrumental variables in this study. For each environmental pollutant-associated trait, SNP exposure instrumental variables were obtained by removing SNPs with linkage disequilibrium, and these exposure instrumental variables were matched with the HNSC GWAS data to obtain the corresponding HNSC-associated SNP endpoint instrumental variables, and the exposure and endpoints were aligned to the same allele according to the SNPs in exposure and endpoints through harmonization to obtain SNPs for MR analysis. F-statistics were performed for each environmental pollution factor-related trait, and calculations revealed that the F-test statistic for the instrumental variables of these indicators was greater than 10, indicating that the SNPs screened by the coordinated alleles were mostly strong-effect instrumental variables, and the possible bias caused by the weak instrumental variables was limited. Then the SNPs with $F > 10$ were analyzed by MR-PRESSO analysis, and the outliers that appeared in MR-PRESSO analysis were removed, and the final statistics of the number of SNPs obtained are shown in Table 3.

Five models, MR Egger, Weighted median, Inverse variance weighted, Simple mode, and Weighted mode, were used for the analysis, respectively. Among them, the IVW model was used as the main reference model, and the other four models were used as supplementary data. The results showed that $P_{IVW} < 0.05$ was found between traits and HNSC associated with three groups of environmental pollution factors, which were Nitrogen dioxide air pollution, Nitrogen oxides air pollution and PM2.5. There was a significant causal relationship between the three groups of environmental factors and HNSC (Fig. 1A), and traits associated with all three groups of environmental pollution factors increased the risk of HNSC, including $OR_{IVW} = 1.004$, $P_{IVW} = 0.049$ for Nitrogen dioxide air pollution, $OR_{IVW} = 1.005$, $P_{IVW} = 0.012$, and $OR_{IVW} = 1.006$, $P_{IVW} = 0.018$ for PM2.5 (Fig. 1B).

Table 3 Screening of data P value thresholds for head and neck cancer and environmental pollution

Trait	P value filter	SNPs Number (P value filter)	SNPs Number (linkage disequilibrium)	SNPs Number (harmonized allele, F-statistics > 10, MR-PRESSO filter)
Nitrogen oxides air pollution	1.00E-06	516	32	28
Nitrogen dioxide air pollution	1.00E-06	715	33	30
Particulate matter air pollution (pm2.5)	1.00E-06	442	23	18
Particulate matter air pollution (pm10)	5.00E-06	95	30	27
Workplace very dusty: Often	5.00E-06	96	16	10
Workplace full of chemical or other fumes: Often	5.00E-06	125	21	17
Workplace had a lot of cigarette smoke from other people smoking: Often	5.00E-06	103	18	13
Worked with materials containing asbestos: Often	5.00E-06	72	12	10
Worked with paints, thinners or glues: Often	5.00E-06	60	14	9
Worked with pesticides: Sometimes	5.00E-06	74	17	14
Workplace had a lot of diesel exhaust: Often	5.00E-06	172	22	15

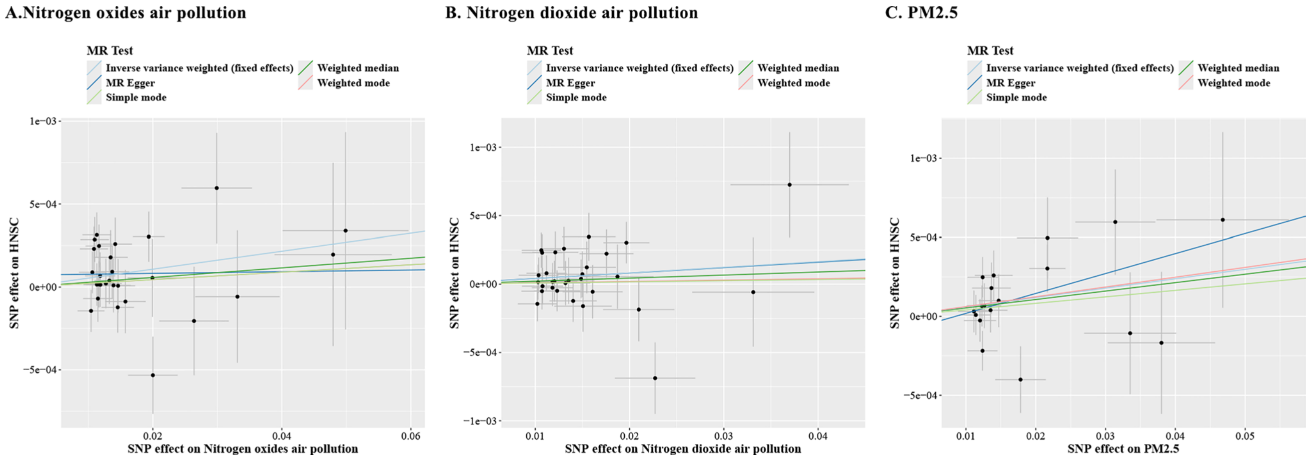


Fig. 2. Linear analysis of 3 environmental pollution factors significantly associated with HNSC traits. The x-axis represents the effect of SNP on exposure and the y-axis represents the effect of SNP on outcome. A slope of less than 0 means that the exposure factor is a favorable factor for the outcome and vice versa. Each point represents an instrumental variable SNP, and each line actually represents the 95% confidence interval. The horizontal coordinate is the effect of SNP on the exposure factor (environmental pollution factor); the vertical coordinate is the effect of SNP on the outcome factor (HNSC). Their ratio represents the effect of the exposure on the outcome, which is the slope of the colored lines in the above figure, and the lines of different colors indicate the different algorithms. The results show that the lines for the different algorithms are generally sloping upwards, which implies that the risk of HNSC increases as environmental pollution factors are elevated

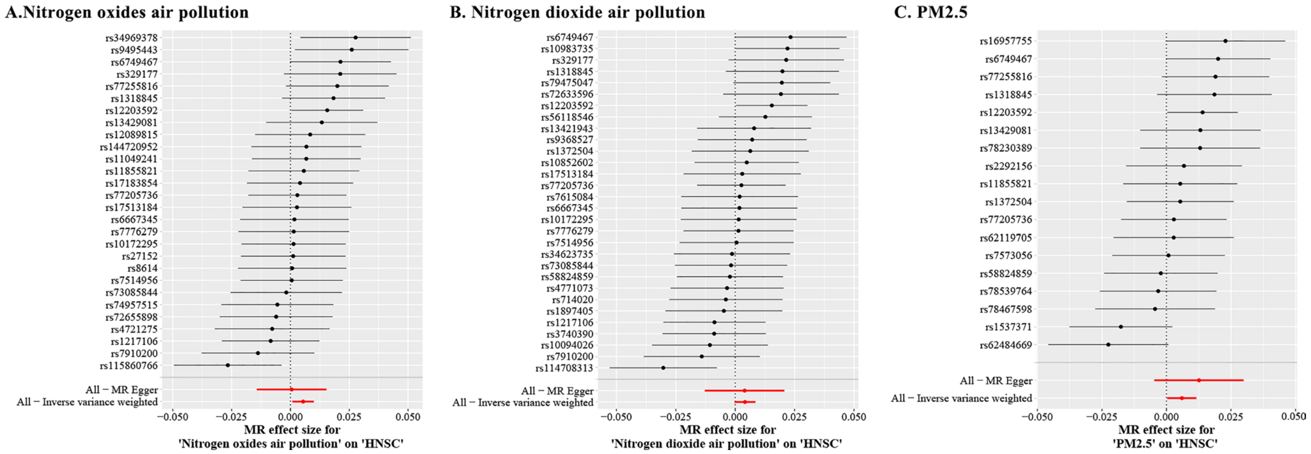


Fig. 3 Forest plot of effect values for the effects of Nitrogen dioxide air pollution (**A**), Nitrogen oxides air pollution (**B**) and PM2.5 (**C**) on outcome. This Figure is a forest plot of the instrumental variables, or SNPs. Each horizontal solid line represents a single SNP, estimated using the Wald ratio method. If the solid line is entirely to the left of 0, it means that for this SNP, increased environmental pollution factors reduce the risk of developing HNSC (outcome). If the solid line is entirely to the right of 0, it means that for this SNP, elevated environmental pollution factors increase the risk of developing HNSC disease. Results crossing 0 indicate that it is not significant, but the results for a single SNP are not robust because the environmental pollution factors are influenced by multiple SNPs, so you need to look at the results together, specifically the red line at the very bottom. This red line indicates that elevated environmental pollution factors increase the risk of developing HNSC disease

We show the linear relationship of the 3 environmental pollution factors traits significantly associated with HNSC on the development of HNSC in 5 models (Fig. 2A–C), where the slope indicates the magnitude of the β-value, i.e., the strength of the linear dependence of the instrumental variable on the exposure effect versus the effect on the outcome. Forest plots of the effect value of each instrumental variable in each trait on the effect of outcome are shown in Fig. 3A–C.

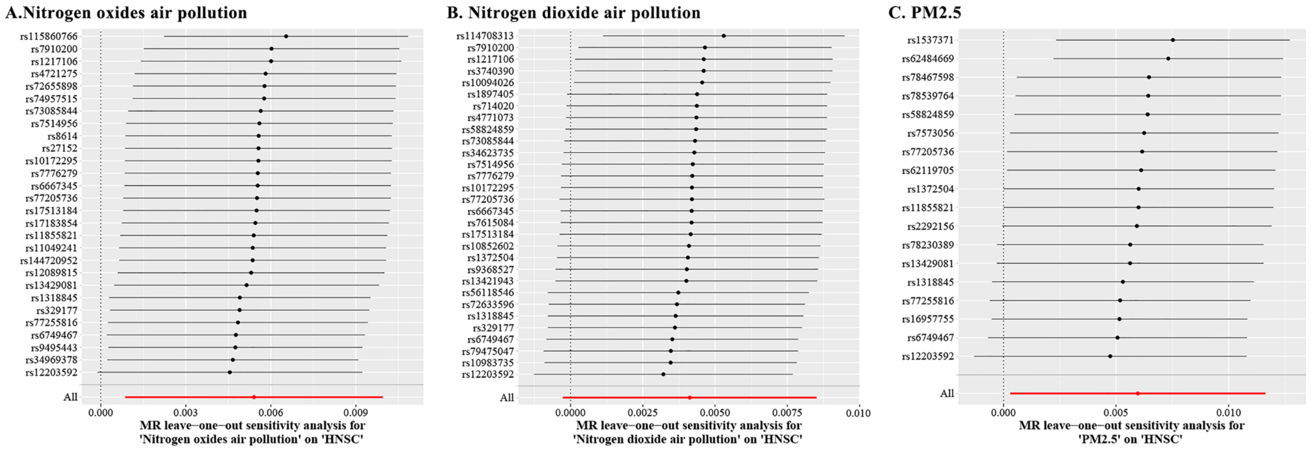


Fig. 4 Estimated causal relationship between environmental pollution factors and HNSC after excluding each SNP

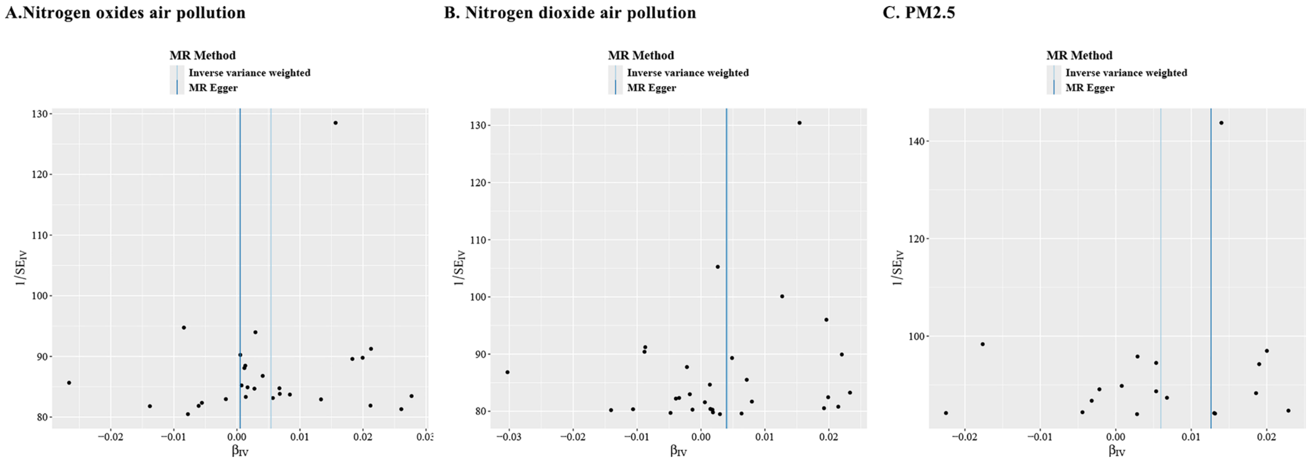


Fig. 5 Funnel plot of Mendelian randomization analysis of heterogeneity test of significant environmental pollution factors with HNSC

3.1.2 Results of sensitivity analysis

The Cochran Q test was used to test the heterogeneity of the MR Egger and IVW model results, in which the P value of Nitrogen dioxide air pollution was 0.23 and 0.271, the P value of Nitrogen oxides air pollution was 0.242 and 0.264, and the P value of PM2.5 was 0.167 and 0.179, respectively. The P-value for all three sets of traits were all greater than 0.05, which indicated that there was no significant heterogeneity in the three sets of data. The P-value of traits for all 3 groups was less than 0.05, indicating that the hypothesis of predicting that exposure may affect outcomes based on the instrumental variables (SNPs) associated with the 3 groups of exposure is valid and that there is no reverse causality between exposure and outcome. The leave-one-out analysis was used to analyze the effect of the instrumental variables on the causal effect of environmental pollution factors and HNSC by removing each instrumental variable locus one by one (Fig. 4). There was no significant shift in the total effect of the instrumental variables. The funnel plot showed a symmetry of the scatter points (Fig. 5), which suggests that there is no potential bias in the results.

After analyzing the MR results, sensitivity analysis was used to test the reliability and stability of the analysis results, and the one-by-one exclusion test is one of them. One by one exclusion test refers to the gradual exclusion of each SNP, calculating the meta effect of the remaining SNPs, and observing whether the results change after the exclusion of each SNP. If the results change significantly after the exclusion of a SNP, it means that there is a SNP that has a great impact on the results, and there is an outlier that needs to be excluded. As shown in the Figure above, the overall error line does not change much after excluding each SNP, meaning that all the error lines are to the right of 0, which indicates that the results are reliable.

The funnel plot can look at the heterogeneity of SNPs, focusing mainly on whether the points on the left and right sides of the IVW line are roughly symmetrical. If there are any particularly outlying points that indicate outliers, they can be removed and analyzed again by MR. Using MR-Egger intercept test for multivariate testing. The P-value of Nitrogen dioxide air pollution = 0.985, the P-value of Nitrogen oxides air pollution = 0.5, and the P-value of PM2.5 = 0.437, and the P-value of all 3 groups of trait were greater than 0.05. The multivariate validity test using -MR-PRESSO found that the P-value of Nitrogen dioxide air pollution = 0.28, Nitrogen oxides air pollution = 0.247, and PM2.5 = 0.985, Nitrogen oxides air pollution = 0.5, and PM2.5 = 0.437, respectively. The P-value of traits for all 3 groups was also greater than 0.05. This indicates that when the effect of instrumental variable (SNP) on environmental pollution factors is 0, the effect on the development of HNSC is also not significant, i.e., there is no horizontal pleiotropy, which means that the results of the analysis in this study are reliable.

3.1.3 Access to MRG

The results of MR analysis showed that Nitrogen dioxide air pollution, Nitrogen oxides air pollution and PM2.5 increased the risk of developing HNSC. Therefore, we extracted 30 SNPs with Nitrogen dioxide air pollution as an exposure factor, 28 SNPs with Nitrogen oxides air pollution as an exposure factor, and 18 SNPs with PM2.5 as an exposure factor, and obtained a total of 54 exposure-associated SNPs after removing duplicates. The VEP annotation tool was used to obtain the corresponding genes and neighboring genes of these SNPs. 16 SNP-associated genes were found in Nitrogen dioxide air pollution, 15 SNP-associated genes were found in Nitrogen oxides air pollution and 11 SNP-associated genes were found in PM2.5. A total of 27 SNP-associated genes were obtained after removing duplicates.

The expression matrix of 27 genes in HNSC was extracted, among which 15 Nitrogen dioxide air pollution related genes were present in the HNSC expression matrix, 15 Nitrogen oxides air pollution related genes were present in the HNSC expression matrix, and 10 PM2.5 related genes were present in the HNSC expression matrix. We screened the genes among them based on the results of the leave-one-out test. When the beta value of the SNP in the leave-one-out method result is closer to 0, it means that the loss of this SNP pair will make the beta value smaller, which implies that the loss of this snp will have a greater impact on the outcome. Therefore, we selected the TOP 5 SNPs closest to 0 for inclusion in the subsequent analysis. If there were not enough genes corresponding to the SNPs in the TOP 5 SNPs to be 5, or that a gene among the 5 genes did not exist in the HNSC expression matrix, we selected the SNPs and the genes in cis-forward order until we selected the 5 genes, and labeled the finally obtained genes as MRG. The screening process is shown in Fig. 6. By following the Fig. 6 process we finally obtained a total of 9 MRGs (IRF4, CDKAL1, UNC5D, ZNF423, ARL15, RSRC1, PTHLH, LINGO1, C19orf35).

3.2 Study of MRGs and ATGs based on the HNSC expression matrix

3.2.1 MRG expression level analysis and relationship prediction

A heatmap was used to show the expression levels of 9 MRGs in HNSC (Fig. 7A), and a boxplot was used to show the expression distribution of these MRGs in HNSC (Fig. 7B). Figure 7AB found that the expression level of ZNF423 was lower in the tumor group than in the normal group, the expression level of UNC5D was lower, and the expression level of the remaining 7 MRGs was higher in the tumor group than in the normal group. The Wilcoxon test found that all 9 MRGs were significantly different between tumor and normal groups. Based on the expression matrix of HNSC tumor group samples, the correlation between MRGs was calculated using the Pearson correlation coefficient (Fig. 7C), in which RSRC1 and CDKAL1 were highly positively correlated (0.62) and PTHLH and IRF4 were highly negatively correlated (−0.32).

3.2.2 Differential expression (DE) analysis, marker MR-DEG, AT-DEG screening

The "limma" package was used to analyze the differential expression of the HNSC expression matrix. Genes with P value < 0.05, |log2(FC)|> 0.5 were selected as DEGs, where genes with P value < 0.05, log2(FC) > 0.5 were up-regulated DEGs, and genes with P value < 0.05, log2(FC) < −0.5 were down-regulated DEGs (Fig. 8A). A total of 7185 DEGs were obtained from differential expression analysis. MRGs and ATGs were each intersected with DEGs, respectively (Fig. 8B), and 4 Marker MR-DEGs (IRF4, LINGO1, PTHLH, RSRC1) and 84 AT-DEGs were found.

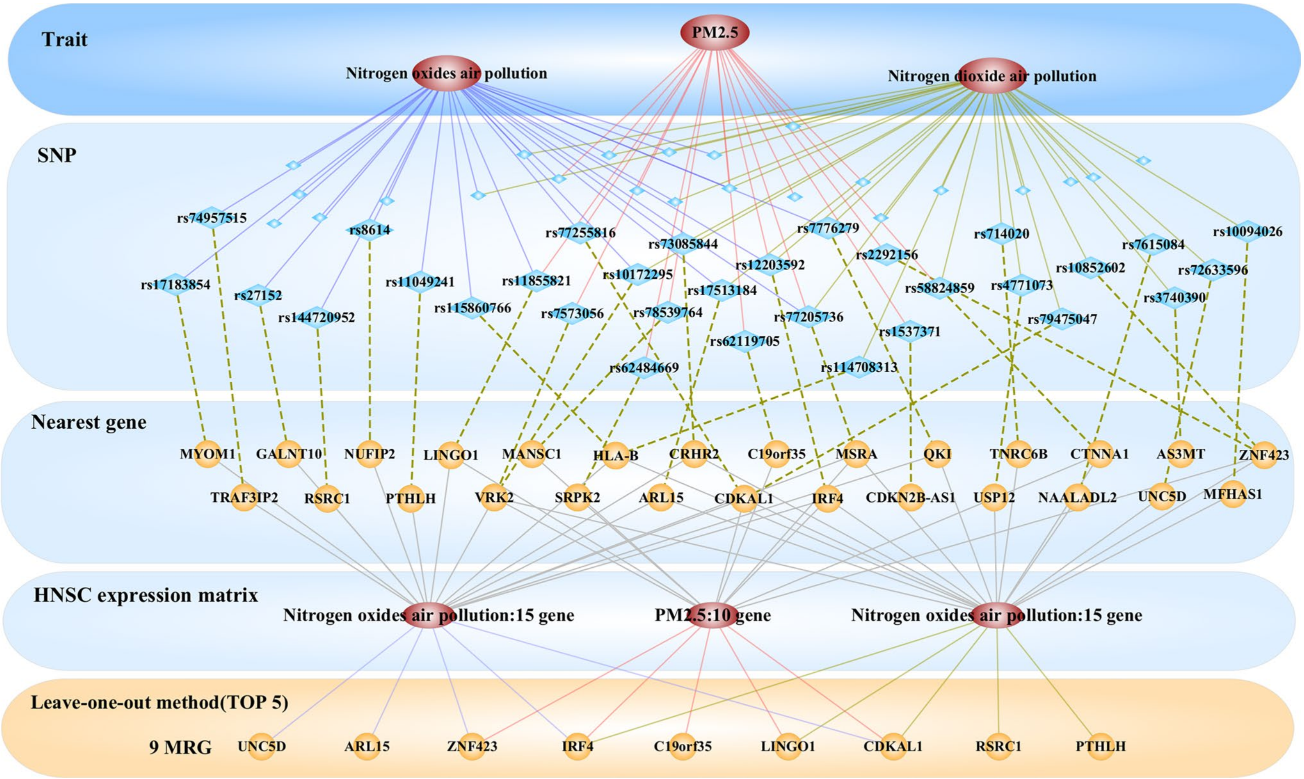


Fig. 6 MRG screening process

3.2.3 Marker MR-DEG and AT-DEG enrichment analysis

Integration of 4 Marker MR-DEG (IRF4, LINGO1, PTHLH, RSRC1) and 84 AT-DEG totaling 88 genes comprised the target gene set. GO Biological process and KEGG pathway analyses of these 88 genes were performed using the clusterProfiler package of R. P value < 0.05 was selected as the significant pathway, and the top 15 significantly enriched results of biological process and KEGG pathway were selected for display. As a result, 88 target gene sets were obtained that mainly regulate biological processes and pathways such as regulation of autophagy, macroautophagy, neuron death and cellular response to external stimulus Fig. 8C, D).

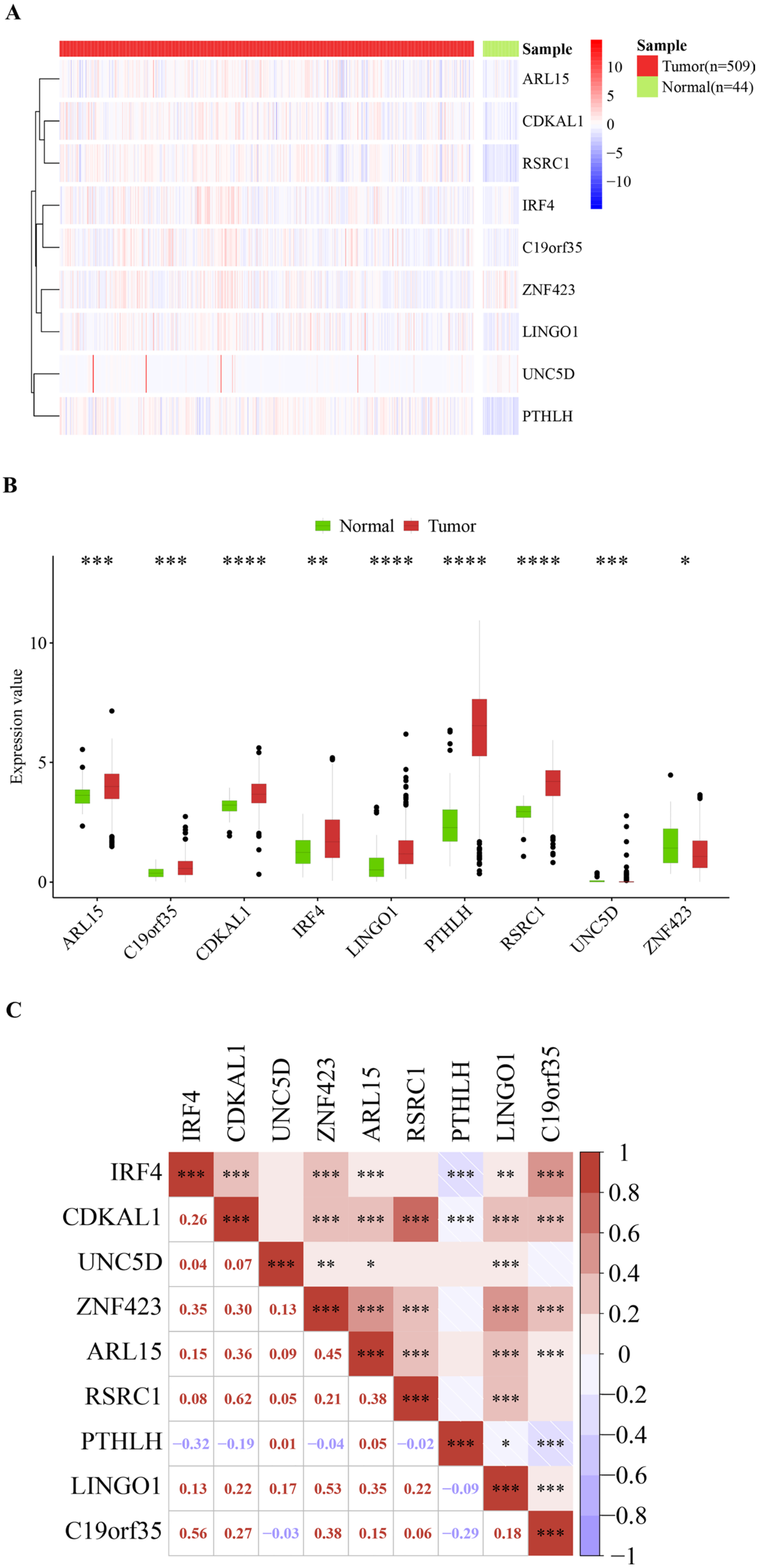
3.2.4 Marker AT-DEG screening and expression level analysis

The Pearson correlation coefficient was used to calculate the correlation between 4 Marker MR-DEG and 84 AT-DEG. Among them, there were 10 Marker MR-DEG&AT-DEG relationship pairs with correlation > 0.6, and 10 Marker AT-DEGs highly correlated with Marker MR-DEG were identified (PIK3R4, KLHL24, PARP1, USP10, ATG3, CXCR4. EIF4G1, TBK1, RB1CC1, RAB7A). A heatmap demonstrated the expression level of 10 Marker AT-DEG in HNSC (Fig. 9A). A box plot demonstrates the expression distribution of these Marker AT-DEG in HNSC (Fig. 9B). As a result, the expression levels of all 10 Marker AT-DEG were found to be higher in the tumor group than in the normal group, and the Wilcoxon test found that there was a significant difference between the tumor group and the normal group for all 10 Marker AT-DEG.

3.2.5 Prediction of marker MR-DEG and marker AT-DEG relationships

We calculated the correlation between Marker AT-DEG using the Pearson correlation coefficient. The relationship between Marker AT-DEG was demonstrated using the R language ggcor package, along with the relationship between Marker MR-DEG and Marker AT-DEG (Fig. 9C). The results showed a significant negative correlation between CXCR4 and PTHLH (cor = – 0.3261); a significant positive correlation between KLHL24 and RSRC1 (cor = 0.7231); and a significant positive correlation between PIK3R4 and RSRC1 (cor = 0.7927).

Fig. 7 MRG expression levels. **A** Heatmap of MRG expression in HNSC; **B** Differential expression of MRG in disease and normal samples of HNSC; **C** Correlation between MRG in HNSC



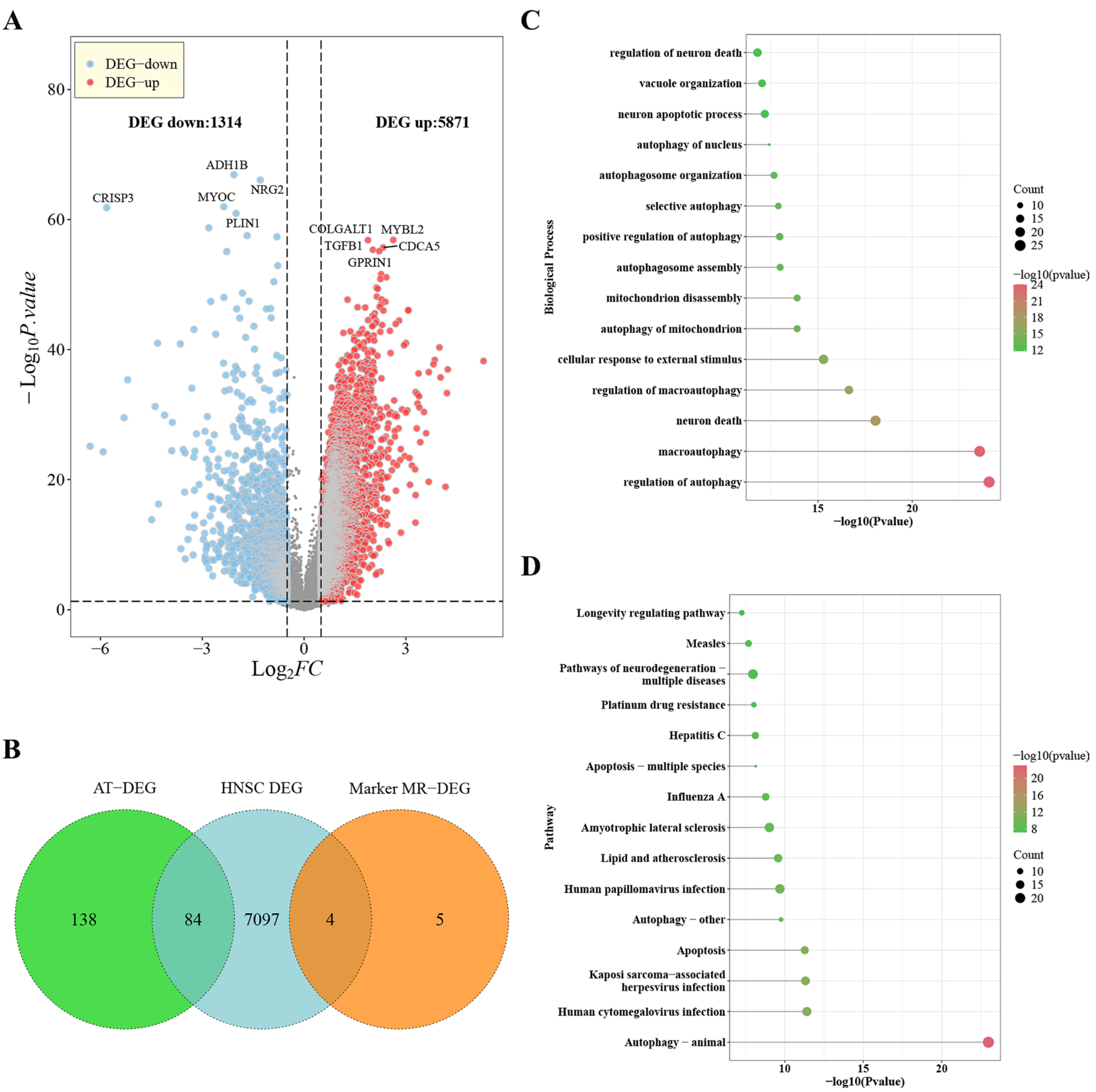


Fig. 8 Marker MR-DEG and AT-DEG screening and enrichment analysis. **A** Volcano plot of HNSC gene differential expression analysis; **B** Venn diagram showing overlap between AT-DEG\HNSC-DEG and Maker MR-DEG; **C** GO biological process enrichment analysis of 88 target genes; **D** KEGG pathway enrichment analysis of 88 target genes

The results of Marker MR-DEG and Marker AT-DEG enrichment analysis were combined with the Pathway dataset obtained from the KEGG database. 94 combinations were found between the 3 Marker MR-DEG, pathway, and bridge DEG. 34,336 combinations were found between Marker MR-DEG, Pathway, bridge DEG, and Marker AT-DEG. Cytoscape software was utilized to demonstrate these combinations (Fig. 9D), and Fig. 9D mainly demonstrates the potential connection between Pathway-based, Marker MR-DEG and Marker AT-DEG. In the enrichment analysis results IRF4 (Marker MR-DEG) and Th17 cell differentiation (Pathway) were significantly correlated, Th17 cell differentiation includes the gene FOS (bridge DEG), FOS is present in Apoptosis (Pathway), and Apoptosis is significantly correlated with PARP1 (Marker AT-DEG) in the enrichment results. We therefore suggest that there may be some association between IRF4-Th17 cell differentiation-FOS-Apoptosis- PARP1.

3.3 Results of the Cox-PH clinical prediction model and survival analysis

3.3.1 Feature selection related to clinical prediction models

First, four Marker MR-DEGs (IRF4, LINGO1, PTHLH, RSRC1) were directly included in the subsequent analysis as traits of the clinical prediction model. Then, a Cox-PH model was established based on clinical characteristics OS and Survival Status, and univariate analysis was performed on 10 Marker AT-DEG significantly correlated with Marker MR-DEG, and univariate analysis was also performed on the clinical characteristics data of HNSC (Fig. 10A). It was found that Age and USP10 were associated with survival significantly correlated (P value $<$ 0.05). Combining the results of Marker MR-DEG and univariate analysis, six factors were finally obtained: IRF4, LINGO1, PTHLH, RSRC1, Age, and USP10.

3.3.2 Multivariate analysis

Multivariate analysis of the six traits (IRF4, LINGO1, PTHLH, RSRC1, Age, USP10) was conducted using the Cox-PH model to predict the overall effect of the six traits on HNSC survival (Fig. 10B).Fig. 10B found significant correlations between IRF4, Age, USP10 and survival. Multivariate Cox-PH modeling was used to obtain risk scores for the samples, and the samples were categorized into a High risk group and a Low risk group based on the median. Observation of the distribution of the samples in different risk groups revealed that the number of deaths in the High risk group was higher than in the Low risk group, and the number of survivors in the High risk group was lower than in the Low risk group (Fig. 10C).

Based on the different risk score subgroups, box plots were used to show the expression distribution of these traits in HNSC (Fig. 10D), and the Wilcoxon test found that each trait was significantly different between the High risk group and Low risk group. All samples and the risk scores of the samples at 3, 5, and 10 years were analyzed separately for survival to see if there was a difference in survival between the High risk group and the Low risk group at different time periods (Fig. 10E–H). Figure 10E–H found that there was a significant difference in survival between the High risk group and the Low risk group. The survival rate of the High risk group was significantly lower than that of the Low risk group, which indicates that risk score is an important factor affecting survival.

3.3.3 Establishment of clinical prediction model nomogram

The final Cox-PH clinical prediction model was built by incorporating the sample risk score as the 7th feature into the survival significantly correlated feature factors. A nomogram plot was used to predict sample Total Points and show sample survival (Fig. 11) at 3-, 5-, and 10-year time nodes. Figure 11 found that sample survival decreased as sample Total Points increased, with 4 traits having a greater impact on Total Points (IRF4, RSRC1, Age, USP10).

3.3.4 Clinical prediction model robustness assessment

Calibration curves (Fig. 12A–C) were calculated for the Cox-PH model over 3 years, 5 years and 10 years to validate the predictive effect of the multivariate Cox-PH model. Calibration curves found that the model predicted probability was close to the actual probability. The model C-index value was then calculated to be 0.63.ROC analysis was performed to assess the predictive effectiveness of the model at 3 years, 5 years and 10 years (Fig. 12D).ROC analysis found that the model had the highest AUC value (AUC = 0.761) at the 10-year time node.

3.3.5 Results of DCA analysis

To find the diagnostic method with the greatest patient benefit, we used DCA to look at 6 traits (IRF4, LINGO1, PTHLH, RSRC1, Age, USP10) and two groups of model predictive scores (riskScore, Total Points) for actual clinical utility (Fig. 13A), Fig. 12A found that as High Risk Threshold increased, riskScore and Total Points Net Benefit outperformed 6 traits. When High Risk Threshold was greater than 0.6, Total Points showed that Net Benefit has converged to 0 and riskScore showed that there is still a Net Benefit. Using Clinical Impact Curve (CIC) to simulate the predicted effects of 1000 people using Total Points and riskScore respectively (Fig. 13B, C), Fig. 12B found that when the High Risk Threshold is close to 0.5, the number of people whose Total Points are judged by the model to be high risk (red line) is very close to the number of

Fig. 9 Expression level analysis of maker genes. (**A**) Heatmap of the expression levels of 10 Marker AT-DEG in HNSC; (**B**) Differential expression of 10 Marker AT-DEG in disease and normal groups of HNSC; (**C**) Marker AT-DEG, Marker MR-DEG and Marker AT-DEG correlation among them; (**D**) Pathway-gene composite network ▸

people who do have an outcome event (blue line). Figure 12c found that when High Risk Threshold is close to 0.55, the number of people whose riskScore is judged by the model to be high risk (red line) is very close to the number of people who do have an outcome event (blue line). When the High Risk Threshold is the same, the number of people judged to be at high risk by Total Points is slightly higher than by riskScore, suggesting that using Total Points to predict survival is more conservative than using riskScore.

3.4 MRG extended analysis results

3.4.1 Tumor Mutation Burden (TMB) analysis results

The SNV dataset was organized to obtain 510 samples, and the R language maftools package was used to draw waterfall plots to show the mutations of 5 survival-related traits (IRF4, LINGO1, PTHLH, RSRC1, USP10) (Fig. 14A), followed by calculating the TMB scores for the SNV dataset. Integrating TMB and survival-related traits (IRF4, LINGO1, PTHLH, RSRC1, Age, USP10, riskScore, Total Points), a total of 509 valid samples were obtained. Based on the survival analysis High risk group and Low risk group, a box plot was used to show the expression distribution of these TMBs in HNSC (Fig. 14B), and the Wilcoxon test found that the TMBs were significantly different between different risk groups. Then the Pearson correlation coefficient was used to view the relationship between hub genes and TMB (Fig. 14C–J), Fig. 14C–J found that Age, RSRC1, riskScore, Total Points and TMB were significantly correlated, but the correlation was not high.

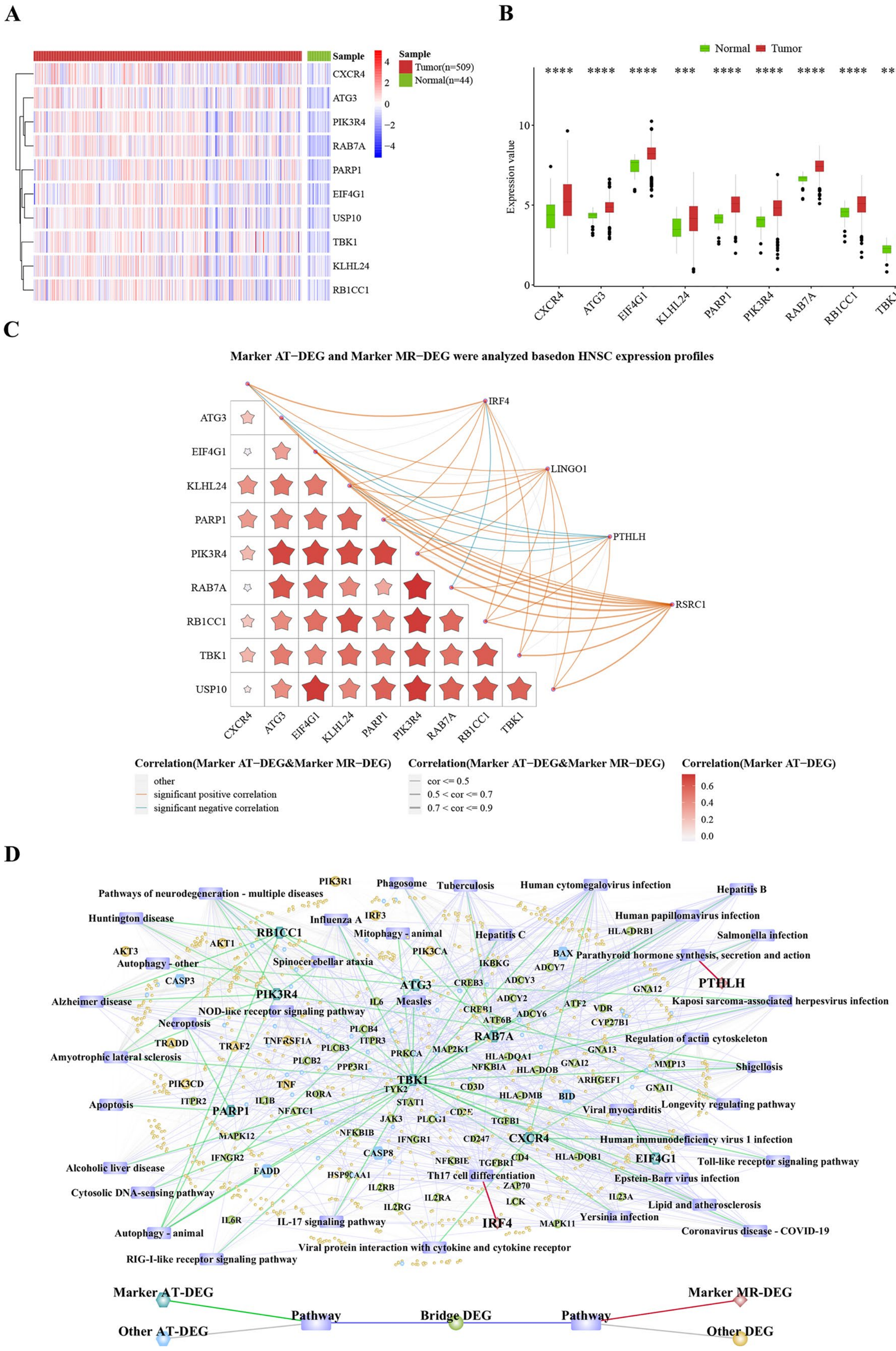
3.4.2 CIBERSORT immune infiltration analysis results

For CIBERSORT analysis we obtained 22 immune cells in HNSC analysis results. P value < 0.05 samples were selected as significantly relevant samples, and a total of 439 tumor samples and 10 normal samples were obtained, and the heatmap demonstrated the fraction of these samples in immune cells (Fig. 15A). The Wilcoxon test was used to analyze the difference of 22 immune cells between tumor samples and normal samples (Fig. 15B), which resulted in finding that B cells naïve, T cells CD8, T cells CD4 memory activated, NK cells resting, Macrophages M0 and Mast cells resting were significantly different between samples. Based on 439 tumor samples, Pearson correlation coefficients were used to predict the relationship between these 6 significantly different immune cells and all immune cells, as well as the relationship between the 6 immune cells and survival-related traits. The results showed a significant positive correlation between T cells CD8 and T cells CD4 memory activated (cor = 0.5970); and a significant negative correlation between T cells CD8 and T cells CD4 memory resting (cor = – 0.5104); a significant negative correlation between T cells CD8 and Macrophages M0 (cor = – 0.5403) (Fig. 15C).B cells naïve was significantly positively correlated with IRF4 (cor = 0.3801);T cells CD8 was significantly negatively correlated with PTHLH (cor = – 0.3855) (Fig. 15D).

Based on survival analysis, we used a cloud rain plot to observe the distribution of abundance scores of these significantly different immune cells in HNSC using the Wilcoxon test to access differences between different risk groups (Fig. 16A–F). The results showed that all 6 immune cells were significantly different between the High risk group and the Low risk group.

4 Discussion

In this comprehensive study, we employed MR analysis to investigate the causal relationships between environmental pollution factors and the risk of HNSC. Our findings reveal significant causal associations between three air pollutants—nitrogen dioxide, nitrogen oxides, and particulate matter 2.5 (PM2.5)—and an increased risk of HNSC. We identified nine genes associated with these environmental pollution factors (MRGs), of which four (IRF4, LINGO1, PTHLH, RSRC1) were found to be differentially expressed in HNSC tissues. Furthermore, we developed a robust six-factor clinical prediction model incorporating four MRGs (IRF4, LINGO1, PTHLH, RSRC1), age, and the autophagy-related gene USP10, which demonstrated good predictive performance for HNSC survival. Our analyses


A
Sample
CXCR4
ATG3
PIK3R4
RAB7A
PARP1
EIF4G1
USP10
TBK1
KLHL24
RB1CC1
4
2
0
−2
−4
Sample
Tumor(n=509)
Normal(n=44)
B
Normal
Tumor
****
***
Expression value
10
5
0
CXCR4
ATG3
EIF4G1
KLHL24
PARP1
PIK3R4
RAB7A
RB1CC1
TBK1
USP10
C
Marker AT−DEG and Marker MR−DEG were analyzed basedon HNSC expression profiles
IRF4
LINGO1
PTHLH
RSRC1
Correlation(Marker AT−DEG&Marker MR−DEG)
other
significant positive correlation
significant negative correlation
cor <= 0.5
0.5 < cor <= 0.7
0.7 < cor <= 0.9
Correlation(Marker AT−DEG)
0.6
0.4
0.2
0.0
D
Pathways of neurodegeneration - multiple diseases
Huntington disease
Autophagy - other
Alzheimer disease
Amyotrophic lateral sclerosis
Apoptosis
Alcoholic liver disease
Cytosolic DNA-sensing pathway
Autophagy - animal
RIG-I-like receptor signaling pathway
Phagosome
Tuberculosis
Human cytomegalovirus infection
Hepatitis B
Human papillomavirus infection
Salmonella infection
Parathyroid hormone synthesis, secretion and action
Kaposi sarcoma-associated herpesvirus infection
Regulation of actin cytoskeleton
Shigellosis
Longevity regulating pathway
Viral myocarditis
Human immunodeficiency virus 1 infection
Toll-like receptor signaling pathway
Epstein-Barr virus infection
Lipid and atherosclerosis
Yersinia infection
Coronavirus disease - COVID-19
Influenza A
Mitophagy - animal
Hepatitis C
Spinocerebellar ataxia
Measles
NOD-like receptor signaling pathway
Necroptosis
Th17 cell differentiation
IL-17 signaling pathway
Viral protein interaction with cytokine and cytokine receptor
RB1CC1
PIK3R4
ATG3
RAB7A
TBK1
PARP1
CXCR4
EIF4G1
PTHLH
IRF4
PIK3R1
IRF3
AKT1
AKT3
CASP3
PIK3CA
IKBKG
ADCY7
BAX
HLA-DRB1
ADCY3
CREB3
IL6
ADCY2
CREB1
ADCY6
ATF6B
ATF2
VDR
CYP27B1
GNA12
TRADD
TRAF2
TNFRSF1A
PLCB4
PLCB3
ITPR3
PLCB2
PRKCA
MAP2K1
HLA-DQA1
NFKBIA
HLA-DOB
GNAI2
GNA13
MMP13
ARHGEF1
PPP3R1
PIK3CD
ITPR2
TNF
RORA
TYK2
CD3D
HLA-DMB
BID
GNAI1
IL1B
NFATC1
STAT1
CD3E
JAK3
PLCG1
TGFB1
MAPK12
NFKBIB
IFNGR1
CD247
IFNGR2
CASP8
NFKBIE
TGFBR1
CD4
HLA-DQB1
FADD
HSP90AA1
IL2RB
ZAP70
IL2RA
IL23A
LCK
IL2RG
IL6R
MAPK11
Marker AT-DEG
Other AT-DEG
Pathway
Bridge DEG
Pathway
Marker MR-DEG
Other DEG

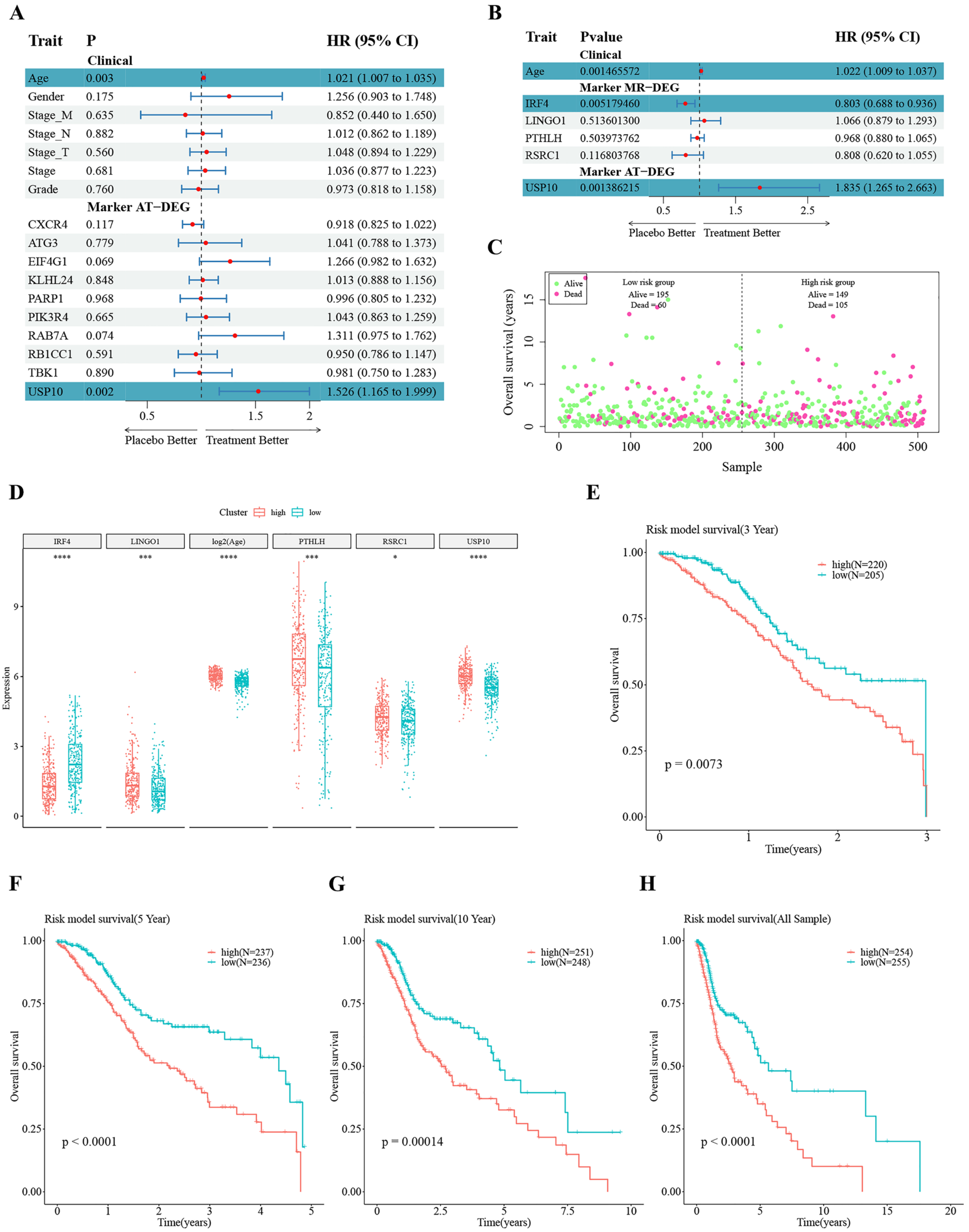


Fig. 10 Results of univariate and multivariate analysis. **A** HR forest plot constructed using traits obtained from univariate cox regression analysis; **B** HR forest plot constructed using traits obtained from multivariate cox regression analysis; **C** Distribution of death and survival samples in high and low risk groups at different time periods. Number of samples on the horizontal axis and survival time on the vertical axis; **D** Difference in risk scores for different traits in the high- and low-risk groups; **E**–**H** KM survival analysis of samples at 3, 5, and 10 years and all years

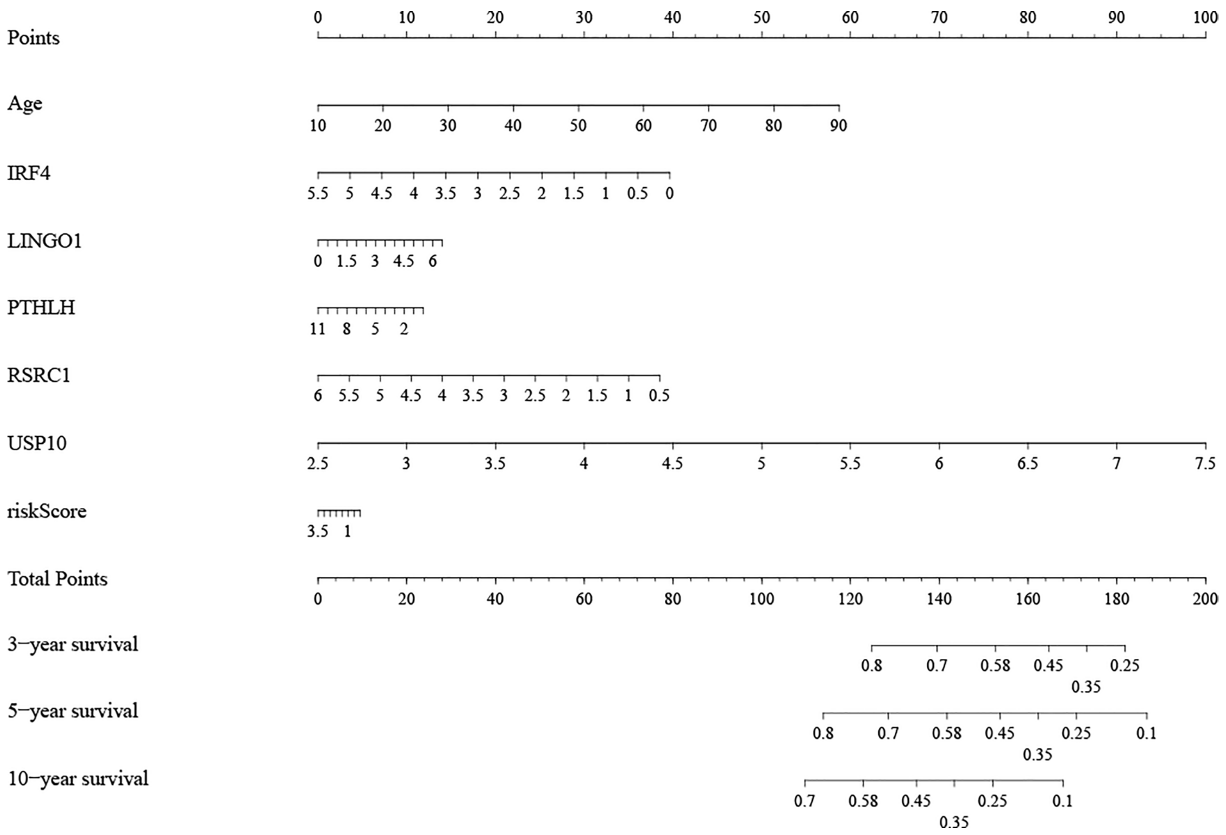


Fig. 11 Nomogram

also uncovered potential connections between these genes, autophagy processes, and immune cell infiltration in HNSC, providing new insights into the complex interplay between environmental factors, genetic susceptibility, and cancer development.

Nitrogen dioxide (NO_2) is a highly reactive gas and a significant air pollutant that can contribute to the development of head and neck cancer through several mechanisms. NO_2 is known to initiate the autoxidation of unsaturated fatty acids, leading to membrane damage and potential cell death when it reacts with polyunsaturated fatty acids in pulmonary lipids via a hydrogen-abstraction mechanism, forming nitrous acid directly in the cell membrane [9]. This oxidative stress can cause cellular damage and inflammation, which are critical factors in carcinogenesis. Additionally, NO_2 exposure is linked to the formation of nitrosamines, such as N-nitrosodimethylamine (NDMA), which are potent carcinogens. Studies have shown that inhalation of NO_2 can increase the biosynthesis of NDMA, which can cause DNA damage and contribute to cancer development [10]. Furthermore, nitrogen oxides, including NO_2, are associated with various human diseases, including cancer, due to their ability to react with cellular components and generate reactive nitrogen species (RNS) that can cause mutations and promote tumor growth [11]. In the context of head and neck squamous cell carcinoma (HNSCC), nitric oxide (NO), a related nitrogen oxide, has been implicated in the pathogenesis of the disease. Increased levels of nitrotyrosine, a marker of NO-induced protein nitrosylation, have been observed in reactive, dysplastic, and HNSCC tissues, suggesting that NO and its byproducts contribute to the immunosuppression and mutagenesis observed in these cancers [12]. High levels of NO can enhance cell cycle progression and apoptosis while inhibiting cell invasion in HNSCC cells, indicating a complex role in cancer progression [13]. Environmental tobacco smoke (ETS), a significant source of NO_2, further exacerbates the risk, as it contains high levels of nitrogen oxides that can irritate the respiratory system and damage lung cells and mucous membranes, leading to increased cancer risk [14]. Epidemiological studies have also linked NO_2 exposure to respiratory illnesses and lung function impairment, which can indirectly contribute to cancer development by creating a pro-inflammatory environment conducive to tumor growth [15]. The multifaceted role of NO_2 in cancer biology underscores the importance of understanding its mechanisms of action, including oxidative stress, nitrosamine formation, and immune modulation, to develop effective prevention and therapeutic strategies [16]. Therefore, NO_2 contributes to head and neck cancer through a combination of direct cellular damage, promotion

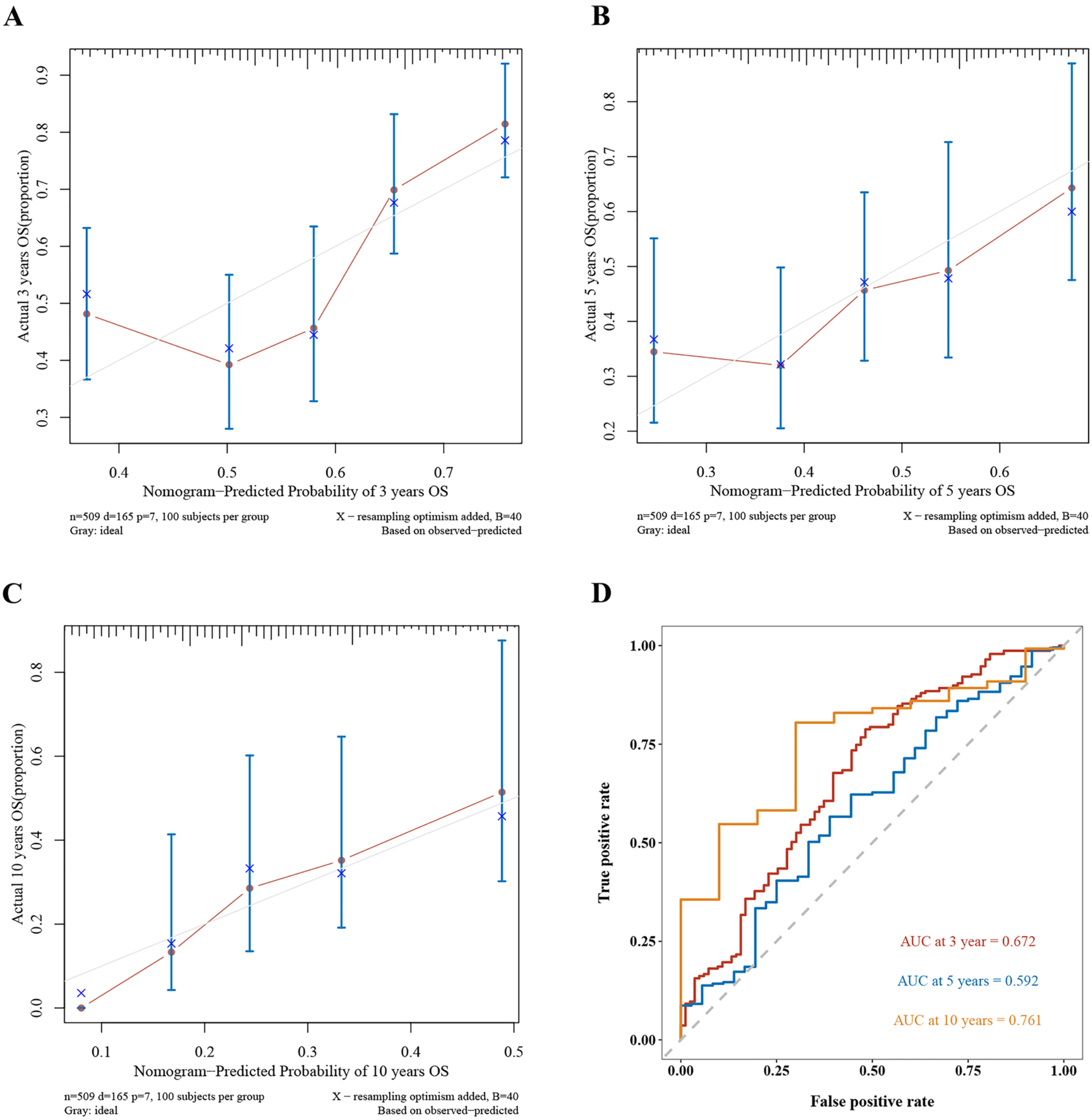


Fig. 12 Assessment of Nomogram robustness. **A**–**C** Calibration curves over 3 years, 5 years, and 10 years. The horizontal axis represents the survival probability predicted by the model, and the vertical axis represents the actual observed survival probability; **D** ROC analyses over 3 years, 5 years, and 10 years

of carcinogenic nitrosamines, and modulation of the immune response, highlighting the need for stringent control of NO_2 emissions to mitigate cancer risk.

PM2.5, or fine particulate matter with an aerodynamic diameter of less than 2.5 µm, has been implicated in the development of various cancers, including head and neck cancer, through several complex mechanisms. One of the primary pathways involves the induction of epithelial-mesenchymal transition (EMT), a process where epithelial cells acquire mesenchymal characteristics, enhancing their migratory and invasive capabilities. PM2.5 particles, along with reactive oxygen species (ROS) and components such as ions and polyaromatic hydrocarbons (PAHs), activate multiple signaling pathways, including TGF-β/SMADs, NF-κB, ERK, PI3K/Akt, Wnt/β-catenin, Notch, Hedgehog, HMGB1-RAGE, and AHR, which are crucial in regulating EMT-related gene expression and cytoskeletal rearrangement [17]. Specifically, PM2.5 exposure has been shown to activate the Notch1 signaling pathway, which is critical in EMT induction. This activation is mediated by the downregulation of miR-139-5p, a microRNA that normally inhibits Notch1. Overexpression

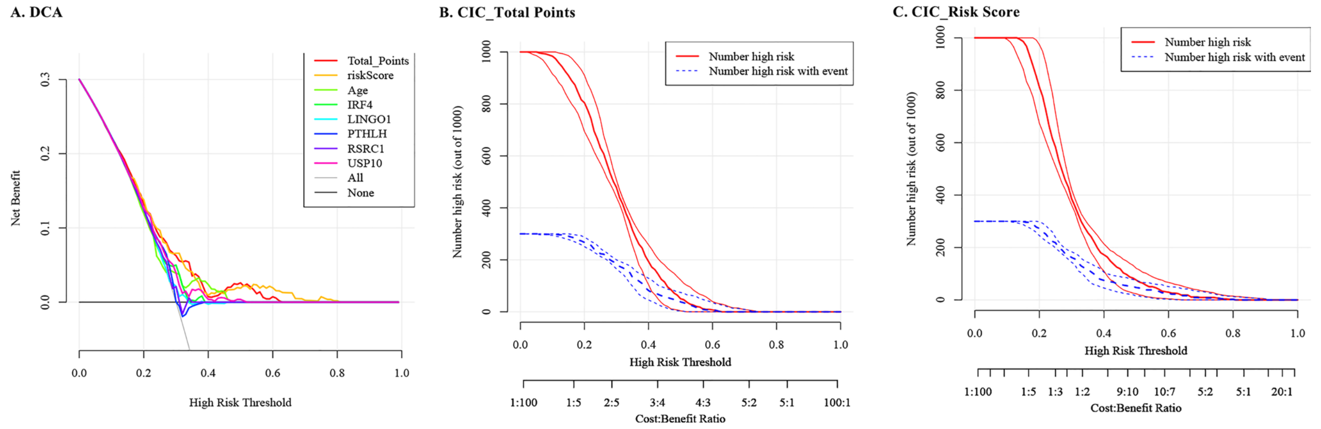


Fig. 13 Comparison of DCA and CIC for different characteristic factors. **A** DCA curves. The horizontal line represents the case when all people do not receive treatment, when the net benefit is 0 regardless of the High Risk Threshold. The dashed line represents the case when all people receive treatment, which corresponds to a change in net benefit as the High Risk Threshold changes. These two lines represent 2 extreme cases. The greater the net benefit at a given probability threshold, the better, so generally the curve should be as far away from the two particular lines as possible. **B**, **C** Clinical Impact Curve (CIC) simulation results. The horizontal coordinate is still the probability threshold and the vertical coordinate is the number of people. The red line indicates the number of people judged by the model to be at high risk at different probability thresholds; the blue line indicates the number of people judged by the model to be at high risk at different probability thresholds and who do have an outcome event. A loss:gain ratio has been added at the bottom, indicating the ratio of losses to gains at different probability thresholds

of miR-139-5p can block Notch1 pathway activation and reverse EMT, suggesting a potential therapeutic target for PM2.5-induced malignancies [18]. Additionally, PM2.5 exposure leads to a shift towards glycolysis in cancer cells, enhancing their malignancy. This metabolic reprogramming is driven by the upregulation of glycolytic genes such as DLAT, facilitated by the activation of eIF4E and transcription factor Sp1. This shift not only promotes glycolysis but also suppresses acetyl-CoA production, contributing to tumor growth and poor prognosis in cancers like non-small cell lung cancer (NSCLC), which shares similar pathogenic mechanisms with head and neck cancers [19]. Furthermore, PM2.5 exposure can induce systemic inflammation and biased hematopoiesis towards the myeloid lineage, generating excessive inflammatory immune cells such as neutrophils and monocytes. This inflammatory milieu can exacerbate pulmonary fibrosis and systemic inflammation, creating a pro-tumorigenic environment that can facilitate the development and progression of head and neck cancers [20]. The involvement of key genes and pathways, such as the Fos proto-oncogene (FOS) and extracellular matrix components, further underscores the multifaceted impact of PM2.5 on various biological processes and molecular functions, including those related to cancer development [21]. Collectively, these findings highlight the intricate mechanisms by which PM2.5 contributes to head and neck cancer, involving EMT induction, metabolic reprogramming, and systemic inflammation, thereby providing potential targets for therapeutic intervention and prevention strategies.

Nitrogen oxides (NOx), including nitric oxide (NO) and nitrogen dioxide (NO_2), play a significant role in the pathogenesis of head and neck cancer (HNC) through various mechanisms. These reactive nitrogen species (RNS) are generated from both environmental sources, such as tobacco smoke and fossil fuel combustion, and endogenous sources, including inflammatory cells and metabolic processes [11, 14]. The high levels of NOx in tobacco smoke, particularly NO_2, can irritate the respiratory system and damage lung cells and mucous membranes, contributing to carcinogenesis [14]. NOx can initiate autoxidation of unsaturated fatty acids in cell membranes, leading to membrane damage and cell death, which is a precursor to cancer development [9]. In the context of inflammation, NO is synthesized by activated inflammatory cells such as macrophages and neutrophils, which release a plethora of inflammatory mediators, including NO, that contribute to the tumor microenvironment [22]. NO has a dual role in cancer biology, acting as both a pro-tumorigenic and anti-tumorigenic agent depending on its concentration and the cellular context. At low concentrations, NO promotes tumor growth by inhibiting apoptosis and supporting angiogenesis, while at high concentrations, it induces cytotoxicity and apoptosis in tumor cells [22, 23]. The epigenetic regulation of nitric oxide synthase (NOS) expression by non-coding RNAs (ncRNAs) and microRNAs (miRNAs) further modulates NO production, impacting tumor proliferation and the tumor microenvironment [24]. NO and its reactive derivatives can cause direct DNA damage through base deamination, adduct formation, and strand breaks, as well as indirect damage by generating secondary radicals that alter gene expression and

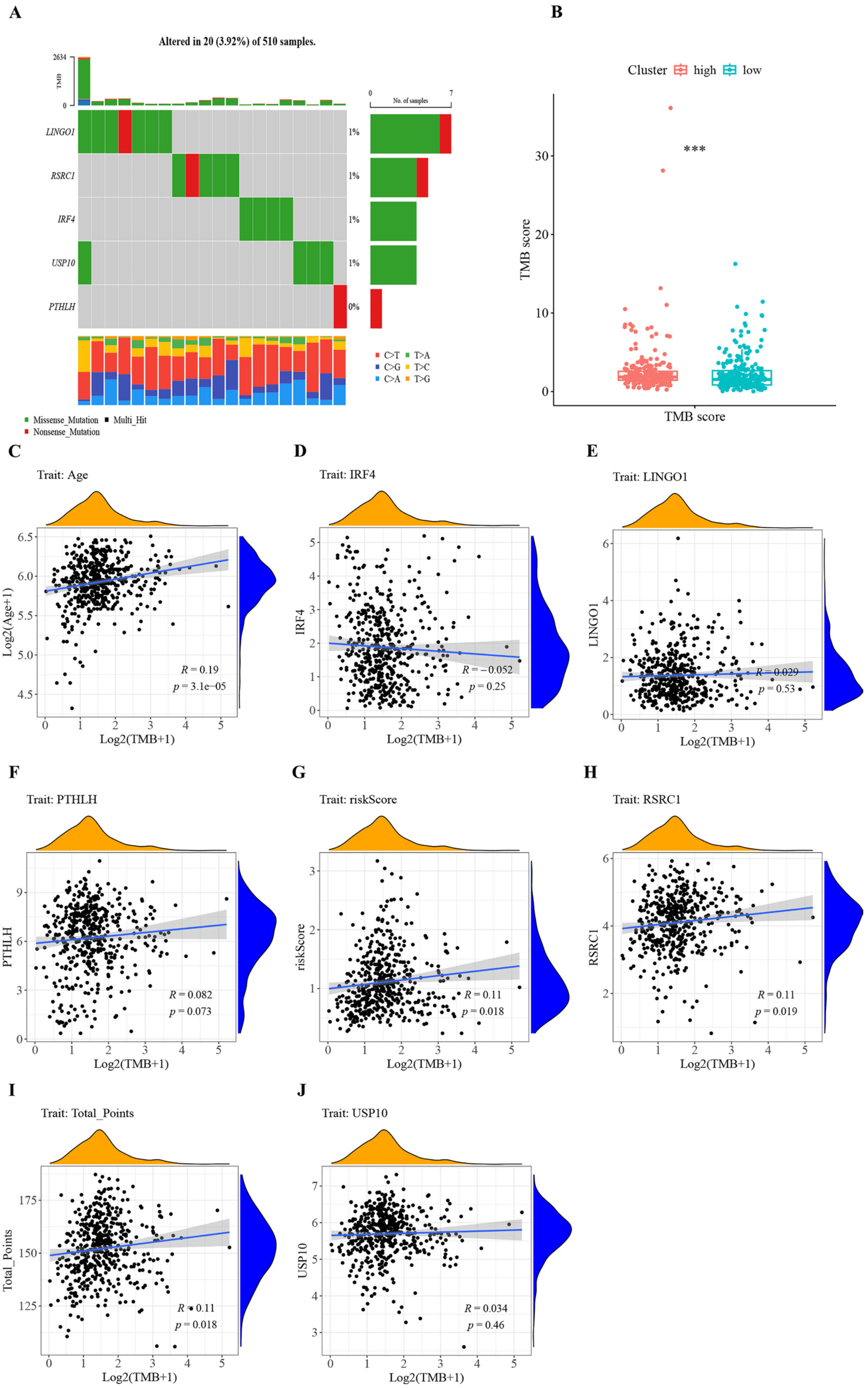
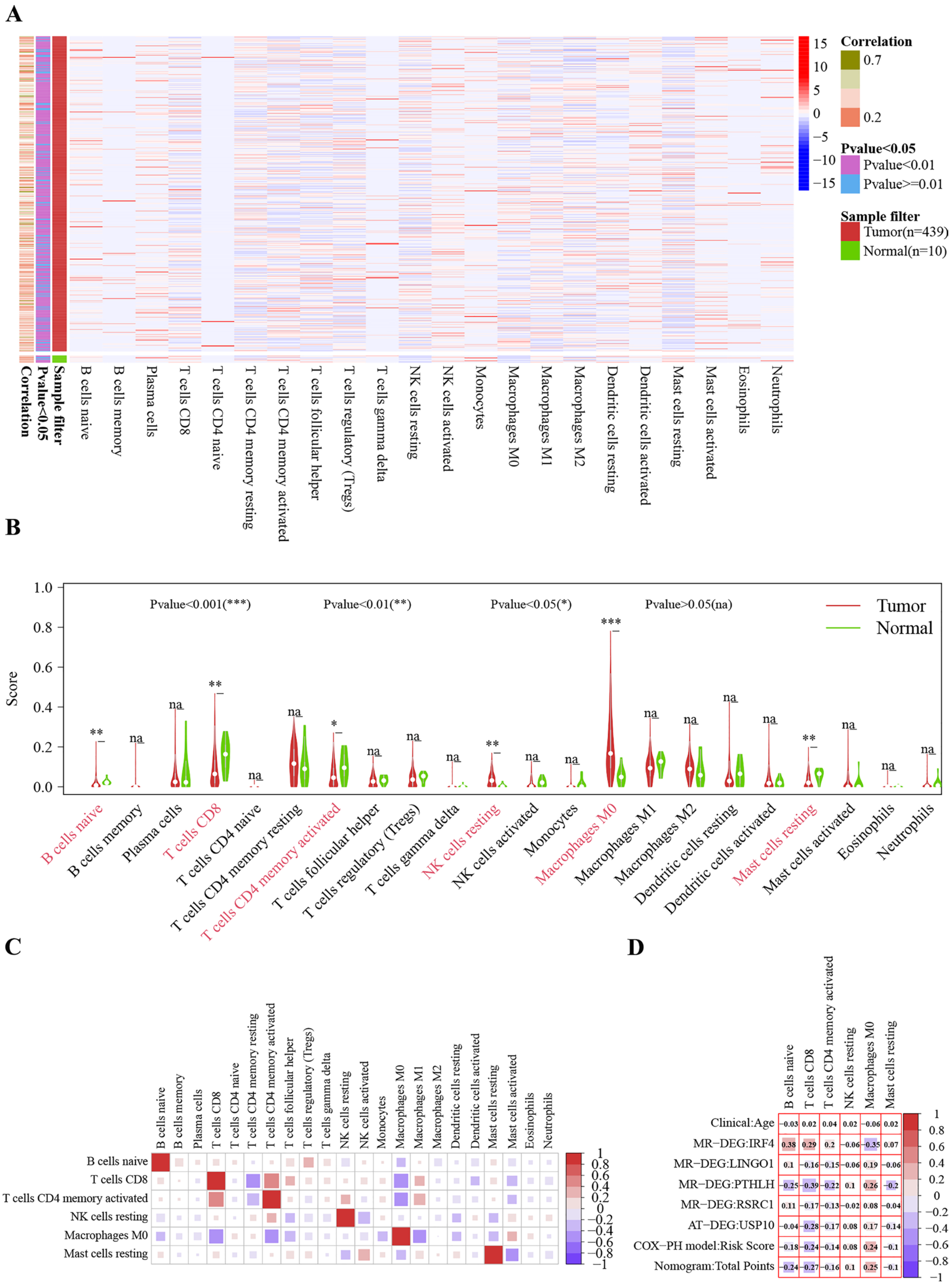


Fig. 14 TMB analysis. **A** Waterfall plot of TMB distribution of the samples; **B** TMB distribution of the samples in the high and low risk groups; **C**–**J** correlation between factors significantly associated with survival and TMB. Panel **A** detail: The horizontal axis represents samples, the vertical axis represents genes, the scale value on the right side of the picture refers to the percentage of samples with mutated genes in the total number of samples. The grey squares in the picture mean that the samples have not mutated, while other colors represent mutated samples, according to the mutation type shown in the legend. The top bar of the image counts the number of mutated genes and the type of mutation in each sample, and the right bar of the image counts the number of mutated samples in the current gene. The bottom rows of the image show the number of base changes in each sample

signal transduction pathways [25]. In head and neck squamous cell carcinoma (HNSCC), sustained inflammation upregulates inducible nitric oxide synthase (iNOS) and cyclooxygenase-2 (COX-2), leading to increased production of reactive species (RS) that activate NF-κB and promote the expression of proinflammatory and proangiogenic proteins such as vascular endothelial growth factor (VEGF) and interleukin-8 (IL-8) [26]. This inflammatory milieu



	B cells naive	T cells CD8	T cells CD4 memory activated	NK cells resting	Macrophages M0	Mast cells resting
Clinical:Age	−0.03	0.02	0.04	0.02	−0.06	0.02
MR−DEG:IRF4	0.38	0.29	0.2	−0.06	−0.35	0.07
MR−DEG:LINGO1	0.1	−0.16	−0.15	−0.06	0.19	−0.06
MR−DEG:PTHLH	−0.25	−0.39	−0.22	0.1	0.26	−0.2
MR−DEG:RSRC1	0.11	−0.17	−0.13	−0.02	0.08	−0.04
AT−DEG:USP10	−0.04	−0.28	−0.17	0.08	0.17	−0.14
COX−PH model:Risk Score	−0.18	−0.24	−0.14	0.08	0.24	−0.1
Nomogram:Total Points	−0.24	−0.27	−0.16	0.1	0.25	−0.1

Fig. 15 Immune cell analysis. **A** Heatmap of abundance of immune cells in HNSC tumor and normal samples; **B** Differential abundance of immune cells in HNSC tumor and normal samples; **C** Correlation between the 6 significantly different immune cells and all immune cells; **D** Correlation between the 6 significantly different immune cells and survival-related traits

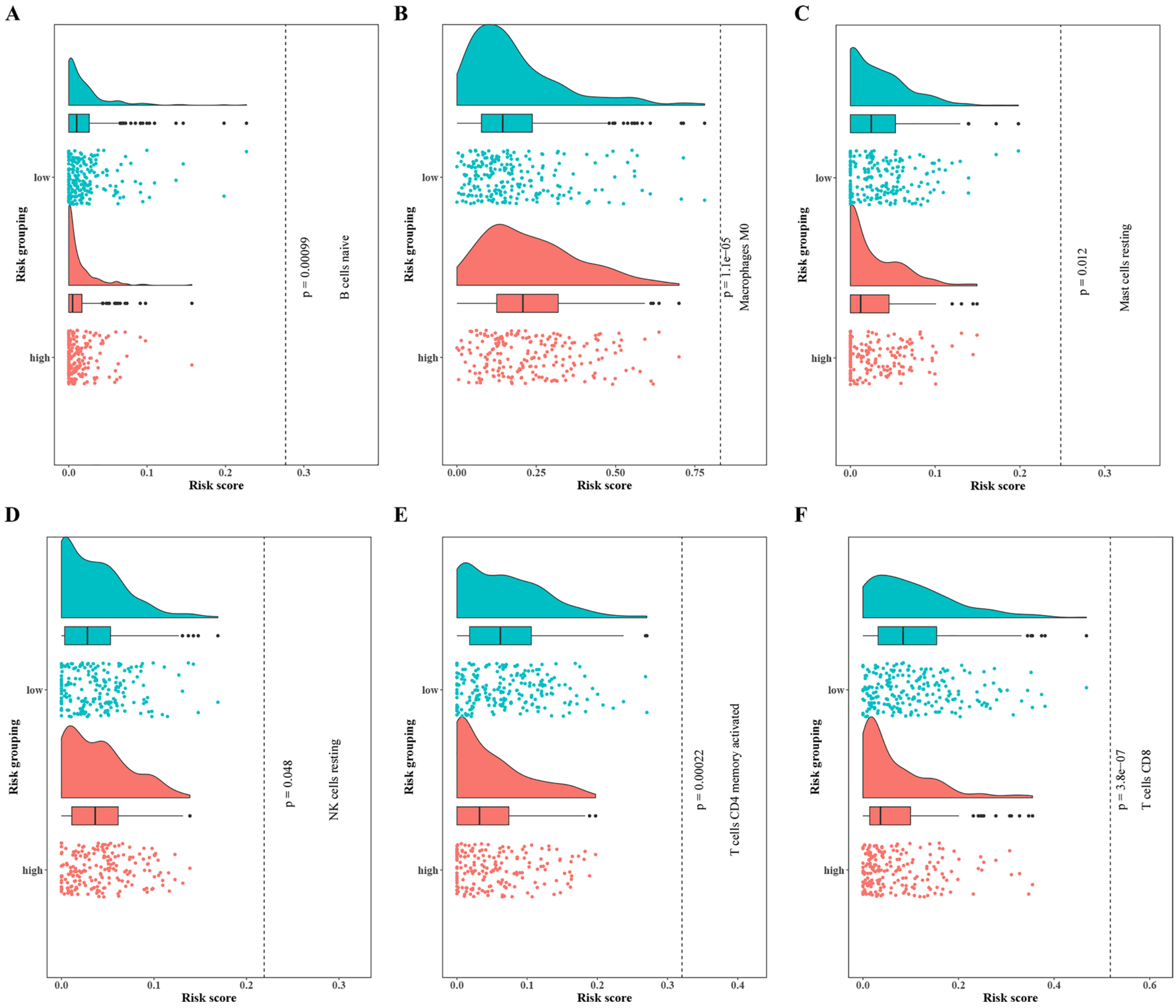


Fig. 16 Distribution of 6 immune cell abundances based on survival analysis subgroups

fosters a tumorigenic phenotype by enhancing cellular proliferation, survival, and angiogenesis. The complex interplay between NOx, inflammation, and genetic and epigenetic factors underscores the multifaceted role of nitrogen oxides in the etiology of head and neck cancer.

This study has several limitations that should be considered when interpreting the results. Firstly, while Mendelian Randomization (MR) analysis can provide evidence for causal relationships, it relies on certain assumptions that may not always hold true in complex biological systems. Secondly, our analysis was based on GWAS data primarily from European populations, which may limit the generalizability of our findings to other ethnic groups. Thirdly, the environmental pollution data used in this study may not fully capture individual-level exposure, potentially leading to exposure misclassification. Fourthly, we were unable to account for potential gene-environment interactions, which could play a significant role in HNSC development. Fifthly, the identified MRGs require further functional validation to confirm their roles in the causal pathway between environmental pollution and HNSC. Sixthly, our study focused on a limited number of environmental pollutants, and other unmeasured pollutants or combinations of pollutants may also contribute to HNSC risk. Lastly, while we analyzed gene expression patterns in HNSC samples, we did not investigate potential epigenetic modifications that could mediate the effects of environmental pollutants on gene expression and HNSC risk.

5 Conclusion

This study provides robust evidence for a causal relationship between specific air pollutants (nitrogen dioxide, nitrogen oxides, and PM2.5) and increased risk of head and neck squamous cell carcinoma (HNSC) using Mendelian Randomization analysis. Our identification of nine MR-associated genes offers insights into potential molecular mechanisms underlying this relationship. These findings enhance our understanding of HNSC etiology and highlight the significant role of environmental factors in cancer development. The results may inform public health policies aimed at reducing air pollution and guide the development of novel prevention strategies and targeted therapies for HNSC.

Author contributions XX conceptualized the study and designed the methods. ZW implemented the analyses, compiled the results, and prepared the preliminary draft. Both XX and ZW edited and prepared the final draft including tables and figures.

Funding This study was supported by the Science and Technology Key Project of Quzhou City (Grant No.: 2022K48).

Data availability All datasets utilized in the study are publicly available. The codes during the current study are available from the corresponding author on reasonable request.

Code availability The codes during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate Not applicable.

Competing interests The authors declare no competing interests.

References

1. Seyyedsalehi MS, Collatuzzo G, Teglia F, Boffetta P. Occupational exposure to diesel exhaust and head and neck cancer: a systematic review and meta-analysis of cohort studies. Eur J Cancer Prev. 2024;33:425–32.
2. Mukherjee S, Ghosh P, Dey C, Paul S. Alteration in immunological profile during malignancy: role of environmental toxicants. Am J Appl Bio-Technol Res. 2022;3(3):38–54.
3. Rezapour M, Rezapour HA, Chegeni M, Khanjani N. Exposure to cadmium and head and neck cancers: a meta-analysis of observational studies. Rev Environ Health. 2021;36(4):577–84. https://doi.org/10.1515/reveh-2020-0109.
4. Lagunas-Rangel FA, Liu W, Schiöth HB. Can exposure to environmental pollutants be associated with less effective chemotherapy in cancer patients? Int J Environ Res Public Health. 2022;19(4):2064.
5. Pagano C, et al. Impacts of environmental pollution on brain tumorigenesis. Int J Mol Sci. 2023;24(5):5045.
6. Miranda-Galvis M, Loveless R, Kowalski LP, Teng Y. Impacts of environmental factors on head and neck cancer pathogenesis and progression. Cells. 2021;10(2):389.
7. Lagunas-Rangel FA, Linnea-Niemi JV, Kudłak B, Williams MJ, Jönsson J, Schiöth HB. Role of the synergistic interactions of environmental pollutants in the development of cancer. GeoHealth. 2022;6(4):e2021GH000552. https://doi.org/10.1029/2021GH000552.
8. Munzeiwa WA, Ruziwa DT, Chaukura N. Environmental pollutants: metal(loid)s and radionuclides. In: Selvasembian R, Van Hullebusch ED, Mal J, editors. biotechnology for environmental protection. Singapore: Springer Nature Singapore; 2022. p. 1–23. https://doi.org/10.1007/978-981-19-4937-1_1.
9. Gallon AA. The mechanism of low levels of nitrogen dioxide reaction with unsaturated fatty acid esters. Louisiana State University and Agricultural & Mechanical College, 1990. Accessed: Aug. 05, 2024. [Online]. Available: https://search.proquest.com/openview/f9729ba1e5d0382662ce1b87171e62f8/1?pq-origsite=gscholar&cbl=18750&diss=y.
10. Rubenchik BL, Glavin AA, Galenko PM, Kilkichko AA, Oleinick IO, Artemov KV. Gaseous nitrogen dioxide increases the endogenous synthesis of carcinogenic N-nitrosodimethylamine in animals. J Environ Pathol Toxicol Oncol Off Organ Int Soc Environ Toxicol Cancer. 1995;14(2):111–5.

11. Wang Y-T, Thomas DD. Nitrogen oxides and their roles in cancer etiology. Curr Pharmacol Rep. 2017;3(4):151–61. https://doi.org/10.1007/s40495-017-0092-3.
12. Bentz BG, Haines GK, Radosevich JA. Increased protein nitrosylation in head and neck squamous cell carcinogenesis. Head Neck. 2000;22(1):64–70. https://doi.org/10.1002/(SICI)1097-0347(200001)22:1%3c64::AID-HED10%3e3.0.CO;2-J.
13. Utispan K, Koontongkaew S. High nitric oxide adaptation in isogenic primary and metastatic head and neck cancer cells. Anticancer Res. 2020;40(5):2657–65.
14. Shamansouri M, Shahabi K, Rostami R. Air concentrations of nitrogen oxide related to tobacco smoke: a systematic review. Tob Health. 2023;2(2):79–82.
15. Magnussen H. Experimental exposures to nitrogen dioxide. Eur Respir J. 1992;5(9): 1040–1042. Accessed: Aug. 05, 2024. [Online]. Available: https://erj.ersjournals.com/content/5/9/1040.short.
16. Pandey SK, Singh J. Nitrogen dioxide: risk assessment, environmental, and health hazard. in Hazardous Gases, Elsevier, 2021, pp. 273–288. Accessed: Aug. 05, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323898577000013.
17. Xu Z, Ding W, Deng X. PM2. 5, fine particulate matter: a novel player in the epithelial-mesenchymal transition? Front Physiol. 2019;10:1404.
18 Wang Y, Zhong Y, Zhang C, Liao J, Wang G. PM2.5 downregulates MicroRNA-139–5p and induces EMT in bronchiolar epithelium cells by targeting Notch1. J Cancer. 2020;11(19):5758.
19. Chen Q, et al. PM2.5 promotes NSCLC carcinogenesis through translationally and transcriptionally activating DLAT-mediated glycolysis reprograming. J Exp Clin Cancer Res. 2022;41(1):229. https://doi.org/10.1186/s13046-022-02437-8.
20. Wang Y, et al. PM $_{2.5}$ increases systemic inflammatory cells and associated disease risks by inducing NRF2-dependent myeloid-biased hematopoiesis in adult male mice. Environ Sci Technol. 2023;57(21):7924–37. https://doi.org/10.1021/acs.est.2c09024.
21. Zhang S, et al. Disease types and pathogenic mechanisms induced by PM2. 5 in five human systems: an analysis using omics and human disease databases. Environ Int. 2024;190: 108863.
22. Aouf S, Laaribi E, Harizi H. Nitric oxide synthases in cancer genetics; focus on nasopharyngeal carcinoma. Oral Health. 2019;4:1–5.
23. Ramírez-Patiño R, et al. Influence of nitric oxide signaling mechanisms in cancer. Int J Immunopathol Pharmacol. 2022;36:039463202211354. https://doi.org/10.1177/03946320221135454.
24. de la Cruz-Ojeda P, Flores-Campos R, Dios-Barbeito S, Navarro-Villarán E, Muntané J. Role of nitric oxide in gene expression regulation during cancer: epigenetic modifications and non-coding RNAs. Int J Mol Sci. 2021;22(12):6264.
25. Kilinc K, Kilinc A. Mutagenic actions of nitrogen oxides. Indoor Built Environ. 2005;14(6):503–12. https://doi.org/10.1177/1420326X05060046.
26. Bradburn JE, et al. The effects of reactive species on the tumorigenic phenotype of human head and neck squamous cell carcinoma (HNSCC) cells. Anticancer Res. 2007;27(6B):3819–27.