

Smoking-independent DNA methylation markers for lung cancer risk: External validation in a large population-based cohort study

Zitong Zhao^{1,2}  | Megha Bhardwaj^{1,3} | Ziwen Fan¹ | Xianzhe Li^{1,2} |
Petra Schrotz-King^{4,5} | Hermann Brenner^{1,3,4,5}

¹Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

²Medical Faculty Heidelberg, University of Heidelberg, Heidelberg, Germany

³German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

⁴NCT Heidelberg, National Center for Tumor Diseases (NCT), A partnership between DKFZ and University Hospital, Heidelberg, Germany

⁵Division of Preventive Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Correspondence

Hermann Brenner, Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, Heidelberg 69120, Germany.
Email: h.brenner@dkfz-heidelberg.de

Funding information

The Federal Ministry of Family Affairs, Senior Citizens, Women and Youth (Berlin, Germany); the Federal Ministry of Education and Research (Berlin, Germany); the Baden-Württemberg State Ministry of Science; Research and Arts (Stuttgart, Germany); the Saarland state Ministry for Social Affairs, Health, Women and Family Affairs (Saarbrücken, Germany)

Abstract

Smoking-associated epigenetic changes have been linked to lung cancer (LC) risk; however, the role of epigenetic alterations independent of smoking is yet to be fully understood. This study aimed to validate 16 previously reported CpG sites that are independent of smoking yet associated with LC risk within a population-based prospective cohort. Using the Infinium Methylation EPIC BeadChip kit or the Infinium HumanMethylation450K BeadChip Assay, DNA methylation (DNAm) in whole blood was assessed in four subsets ($n=736$, 1027, 997, and 312) of a population-based cohort from Germany. The DNAm levels of the 16 smoking-independent CpG sites were analyzed. Hazard ratios (HRs) and their 95% confidence intervals (95% CIs) were calculated to assess associations of DNAm at the 16 CpG sites with LC risk, adjusting for multiple covariates, including smoking habits and a smoking-associated DNAm score. Over 17 years of follow-up, a total of 199 LCs were observed. Among the 16 CpGs, cg02211449 showed a negative association with LC risk (HR [95% CI] per SD increase, =0.70 [0.63–0.78]), while cg11385536 (1.04 [1.01–1.07]), cg09736286 (1.64 [1.10–2.44]), cg19907023 (1.64 [1.01–2.66]), and cg22032485 (1.52 [1.04–2.21]) displayed positive associations with LC risk. Five of the 16 suggested smoking-independent CpGs could be externally validated as predictors of LC risk. Further research should address their potential contribution to enhanced LC risk stratification.

KEYWORDS

DNA methylation, epigenome-wide association study (EWAS), external validation, lung cancer risk

Abbreviations: AUC, area under the receiver-operating characteristic curve; CI, confidence interval; DNAm, DNA methylation; EWAS, epigenome-wide association study; ESTHER, Epidemiologische Studie zu Chancen der Verhütung, Früherkennung und optimierten Therapie chronischer Erkrankungen in der älteren Bevölkerung; GP, general practitioner; HR, hazard ratio; LC, lung cancer; OR, odds ratio; SmPEGS, smoking methylation poly-epigenetic score.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Cancer Science* published by John Wiley & Sons Australia, Ltd on behalf of Japanese Cancer Association.

1 | INTRODUCTION

Lung cancer (LC) is the most frequently diagnosed cancer and the leading cause of cancer death worldwide, accounting for approximately 2.5 million new cancer cases and 1.8 million deaths in 2022.¹ Smoking is a major risk factor for LC.² Smoking-associated epigenetic alterations, such as hypomethylation at cg05575921 in the AHRR gene and cg03636183 in the F2RL3 gene, have been reported to be associated with LC risk and partially mediate the effect of smoking on LC development.^{3–5}

Despite the overwhelming role of smoking in LC etiology, there is also substantial LC incidence among never-smokers, with LC among never-smokers ranking as the seventh most common cancer worldwide.^{6–9} For example, in Taiwan, over 50% of participants with LC had never smoked.¹⁰ This underscores the necessity to broaden our understanding of LC etiology and pathogenesis beyond smoking-associated factors. However, there is a noticeable gap in research regarding smoking-independent DNA methylation (DNAm) changes and LC incidence risk.

In light of this, Zhao et al.¹¹ reported on methylation changes in prediagnostic blood samples that are associated with LC risk independent of smoking in a nested case-control study from the United States including 208 LC cases and 222 matched controls. Controlling for self-reported smoking and smoking-relevant alterations including a smoking methylation poly-epigenetic score (SmPEGS, consisting of 2623 smoking-associated CpG sites),¹² Zhao et al. identified 16 smoking-independent CpG sites that were significantly associated with LC risk and might offer potential for enhanced LC risk stratification.

However, given the limited sample size of the study and lack of independent validation of the results,¹¹ further external validation is warranted. We aimed to carry out such external validation in a population-based prospective cohort from Germany.

2 | METHODS

2.1 | Study design and study population

Participants for the analyses were chosen from the Epidemiologische Studie zu Chancen der Verhütung, Früherkennung und optimierten Therapie chronischer Erkrankungen in der älteren Bevölkerung (ESTHER) study, an ongoing statewide cohort study conducted in Germany. Details of the ESTHER study have been described previously.^{13–16} In brief, 9940 participants (aged 50–75 years at baseline) were enrolled by their general practitioners (GPs) during a routine health examination between July 2000 and December 2002, and then followed up every 2–3 years. Biospecimens (blood, urine, and stool) were collected at the baseline examination and then stored at –80°C. Information on sociodemographic characteristics, lifestyle factors, and health status was collected through standardized self-administered questionnaires at recruitment and each follow-up visit. Comprehensive medical data were collected from GPs' standardized health examination forms.

Prevalent and incident LC cases (ICD-10 code C34) during 17 years of follow-up (between 2000 and 2018) were determined by record linkage with the Saarland Cancer Registry (<https://krebsregister.saarland.de/>). The study was approved by the Ethics Committees of the Heidelberg Medical Faculty of Heidelberg University and the State Medical Board of Saarland. Written informed consent was obtained from each participant.

For the current analysis, we utilized four independent subsets, previously selected for genome-wide DNAm assessments across various projects^{15,17} (Figure S1). Subset I comprises 741 subjects randomly sampled from a subcohort within the ESTHER study. Subset II comprises 1030 subjects randomly selected from participants whose DNAm data were available by 2017 in the ESTHER study. Subset III consists of the first 500 men and 500 women consecutively enrolled during the first 6 months of recruitment. Subset IV employs a nested case-control design for cancer-related methylation signatures, including 266 participants with incident cancers (116 LC cases, 23 prostate cancer cases, and 129 colorectal cancer cases) and 205 randomly selected participants among those free of cancers. For subset IV, only participants with incident LC and participants free of cancers were included in the current analysis. Participants in each subset were independent and did not overlap.

Participants with missing values on the 16 CpG sites or LC incident information were excluded. Ultimately, 736, 1027, 997, and 312 participants of subsets I, II, III, and IV were included in our current study.

2.2 | Methylation assessment and data processing

Details of the DNAm assessment and data processing have been described previously.^{13,18–20} Briefly, DNAm levels of subsets I and II were measured using the Infinium Methylation EPIC BeadChip kit (EPIC, Illumina Inc.), and subset III and IV were analyzed with the Infinium Human Methylation450K BeadChip Assay (450K, Illumina Inc.). The laboratory work was carried out in accordance with the manufacturer's instructions. During data preprocessing, probes with detection *p*-value >0.01, missing values >10%, and those targeting the X and Y chromosomes were excluded. Beta values, ranging from 0 (completely unmethylated) to 1 (completely methylated), were used to evaluate the level of methylation. Leukocyte composition was estimated using Houseman's algorithms.²¹

2.3 | Statistical analysis

Standard descriptive methods were used to characterize the demographic features of the study participants at baseline.

Cox proportional hazard models were employed to estimate the associations of each CpG site with LC risk in subsets I, II, and III. In subset IV, the associations were calculated using weighted Cox regression models that accounted for the case-control sampling design. Model 1 included each CpG site only; model 2 was adjusted

for age at blood drawing, sex, batch effect, smoking status (never-, former, and current smoker), body mass index (BMI, kg/m²), and leukocyte composition (estimated by the Houseman approach²¹); and model 3 was additionally controlled for a SmPEGS score (model 3). The SmPEGS score was derived from a DNAm-based algorithm¹² which comprised 2623 CpG sites differentially expressed between current and never-smokers.²²

Hazard ratios (HRs) with corresponding 95% confidence interval (95% CI) per standard deviation (SD) increase in methylation level at single CpG sites were computed for LC risk. The ability of methylation at the single CpGs to predict LC incidence during 17 years of follow-up was evaluated using areas under the receiver-operating characteristic curves (AUCs) and their 95% CI. Due to variations in DNAm profiles across the four subsets, assessed at different times by different batches of DNAm assessment chips, HRs and corresponding 95% CIs were calculated separately and then combined by random-effects meta-analysis.

Statistical significance was defined by *p*-values <0.05 in two-sided testing. All statistical analyses were performed with R (version 4.1.2) and MedCalc (version 20.216) software.

3 | RESULTS

Table 1 presents the sociodemographic characteristics of participants in each subset at baseline. The mean ages of participants across subsets ranged from 61 to 64 years. In subsets II, III, and IV, a majority of participants who developed LC were men and had a history of smoking. The mean BMIs ranged from 27.1 to 28.1 kg/m² across all subsets. Over a 17-year follow-up period, a total of 26, 31, 35, and 107 participants developed LC in subsets I, II, III, and IV, respectively.

Table 2 shows the mean DNAm levels for 16 CpG sites among participants across the subsets. Among all 16 CpG sites, which were determined according to EPIC array,¹¹ only six CpGs could be detected by both EPIC and 450k arrays. Therefore, in subsets III and IV, we only evaluated these six CpG levels through 450K arrays. In Zhao et al.'s study, all CpG sites—except for cg02211449, which was reported to be hypomethylated in LC—exhibited hypermethylation. In our research, only cg09736286 showed consistent and significant hypermethylation in LC participants in subset I, and cg11385536 showed significant but inconsistent hypomethylation in LC participants in subset III.

Table 3 shows the association between each CpG site and LC incidence. In Zhao et al.'s study, all CpGs showed positive association with LC risk, except for cg02211449, which was negatively associated with LC risk (odds ratio [OR]=0.40). In our study, after adjusting for all variables mentioned above (model 3), five CpG sites were validated with significant association with LC risk. The findings were consistent with those reported in Zhao et al.'s study. However, it is notable that cg11385536 showed inverse associations with LC risk before (model 2: 0.95 [0.92–0.98]) and after (model 3: 1.04 [1.01–1.07]) adjusting for the SmPEGS score. Cg02211449 (model 2: 0.79

[0.68–0.92]; model 3: 0.70 [0.63–0.78]) and cg09736286 (model 2: 1.54 [1.03–2.30]; model 3: 1.64 [1.10–2.44]) showed significant associations with LC risk regardless of adjustment for the SmPEGS score. Cg19907023 (model 3: 1.64 [1.01–2.66]) and cg22032485 (model 3: 1.52 [1.04–2.21]) were significantly associated with LC risk only after adjustment for the SmPEGS score.

Table 4 shows the predictive performance of five CpG sites for LC risk, with AUCs ranging from 0.55 to 0.57.

4 | DISCUSSION

In this prospective cohort study, we validated the association between methylation at 16 smoking-independent CpG sites and LC risk. After adjusting for multiple covariates including smoking habits and the SmPEGS score, five CpG sites (cg02211449, cg11385536, cg09736286, cg19907023, and cg22032485) demonstrated a consistent and significant association with LC risk. The directions of effects of the five CpGs are the same as those reported by Zhao et al.¹¹ Nevertheless, these five CpG sites showed limited predictive ability for LC risk.

Although there are currently few studies examining the direct biological functions of the five CpG sites, we have explored the genomic regions of these identified CpG sites and the functions of their associated genes using GeneCards (<https://www.genecards.org/>)²³ (Table S1). For instance, cg11385536 is located in the CABP1 gene (COQ8/ADCK3), cg09736286 in the DIABLO gene, cg19907023 in the ASTL gene, and cg22032485 in the SSBP4 gene. The pathways associated with these genes include metabolism for CABP1, apoptosis for DIABLO, maturation arrest for ASTL, and DNA binding for SSBP4. Further research on the roles of these CpG sites in LC is warranted.

Despite smoking being the major cause of LC, LC is also one of the more common forms of cancer among never-smokers.^{8,9} The identification of smoking-independent DNAm alterations may prove to be instrumental in understanding the biological mechanisms involved in lung carcinogenesis that are not directly related to smoking. However, few studies so far have systematically compared DNAm profiles in never-smokers who did or did not get LC, likely due to the limited sample sizes of never-smokers with LC. One study revealed DNAm alterations in individuals with lung dysfunction (chronic obstructive pulmonary disease [COPD]) who have never smoked.²⁴ These DNAm patterns might be shaped by exposures to risk factors other than smoking. Exploring DNAm alterations that are independent of smoking among LC patients might likewise open new windows to enhance LC risk stratification.

Petrovic et al.²⁵ also explored smoking-independent CpG sites, but their analysis was limited to controlling for participants' smoking behaviors, without considering known smoking-related epigenetic alterations. They identified 49 smoking-independent CpG sites by adjusting for the comprehensive smoking index (CSI), which incorporates various metrics of cigarette-smoking history.²⁶ Notably, these 49 CpGs did not overlap with the 16 CpGs reported in Zhao et al.'s

TABLE 1 Population characteristics of participants at baseline from the ESTHER study.

Characteristics	Subset I (N = 736)		Subset II (N = 1027)		Subset III (N = 997)		Subset IV (N = 312)	
	Not developing LC (N = 710)	Developing LC (N = 26)	Not developing LC (N = 996)	Developing LC (N = 31)	Not developing LC (N = 962)	Developing LC (N = 35)	Not developing LC (N = 205)	Developing LC (N = 107)
Age (years; mean, SD)	61.6 (6.5)	63.3 (7.2)	62.1 (6.73)	61.1 (5.4)	62.0 (6.5)	64.0 (5.9)	62.2 (6.4)	63.7 (5.8)
Sex (N, %)								
Female	400 (56.3%)	14 (53.8%)	566 (56.8%)	12 (38.7%)	491 (51.0%)	8 (22.9%)	122 (59.5%)	32 (29.9%)
Male	310 (43.7%)	12 (46.2%)	430 (43.2%)	19 (61.3%)	471 (49.0%)	27 (77.1%)	83 (40.5%)	75 (70.1%)
Smoking status (N, %) ^a								
Never	350 (49.3%)	3 (11.5%)	519 (52.1%)	6 (19.4%)	466 (48.4%)	2 (5.71%)	95 (46.3%)	15 (14.0%)
Former	224 (31.5%)	6 (23.1%)	308 (30.9%)	11 (35.5%)	314 (32.6%)	13 (37.1%)	60 (29.3%)	41 (38.3%)
Current	114 (16.1%)	16 (61.5%)	148 (14.9%)	14 (45.2%)	160 (16.6%)	20 (57.1%)	41 (20.0%)	49 (45.8%)
Body mass index (kg/m ² , SD) ^b	27.8 (4.4)	27.1 (5.3)	27.7 (4.2)	28.1 (4.5)	27.8 (4.3)	27.6 (3.9)	27.2 (4.2)	27.1 (4.6)

Abbreviations: ESTHER, Epidemiologische Studie zu Chancen der Verhütung, Früherkennung und optimierten Therapie chronischer Erkrankungen in der älteren Bevölkerung; LC, lung cancer; N, number; SD, standard deviation.

^aData missing for 77 participants of the total cohort.

^bData missing for 7 participants of the total cohort.

TABLE 2 The mean methylation beta values of each CpG site among participants in each subset.

CpGs	Subset I (N = 736)			Subset II (N = 1027)			Subset III (N = 997)			Subset IV (N = 312)		
	Not developing LC (N = 710)	Developing LC (N = 26)	<i>p</i> ^a	Not developing LC (N = 996)	Developing LC (N = 31)	<i>p</i> ^a	Not developing LC (N = 962)	Developing LC (N = 35)	<i>p</i> ^a	Not developing LC (N = 205)	Developing LC (N = 107)	<i>p</i> ^a
Cg19907023	0.67 (0.04)	0.68 (0.03)	0.10	0.69 (0.04)	0.70 (0.04)	0.20	-	-	-	-	-	-
Cg08739693	0.61 (0.03)	0.60 (0.03)	0.50	0.61 (0.03)	0.61 (0.04)	0.60	-	-	-	-	-	-
Cg21432334	0.64 (0.03)	0.64 (0.03)	0.96	0.65 (0.02)	0.65 (0.03)	0.53	-	-	-	-	-	-
Cg12483301	0.53 (0.04)	0.53 (0.03)	0.87	0.55 (0.04)	0.55 (0.03)	0.26	-	-	-	-	-	-
Cg12796320	0.37 (0.04)	0.36 (0.02)	0.36	0.38 (0.04)	0.38 (0.03)	0.69	-	-	-	-	-	-
Cg15979932	0.64 (0.04)	0.63 (0.03)	0.61	0.63 (0.03)	0.63 (0.04)	0.81	0.72 (0.03)	0.72 (0.04)	0.77	0.66 (0.04)	0.66 (0.03)	0.22
Cg06581459	0.45 (0.05)	0.46 (0.04)	0.21	0.46 (0.05)	0.45 (0.05)	0.18	-	-	-	-	-	-
Cg09447640	0.49 (0.03)	0.49 (0.03)	0.78	0.50 (0.03)	0.50 (0.04)	0.52	-	-	-	-	-	-
Cg11343375	0.50 (0.03)	0.50 (0.03)	0.65	0.51 (0.03)	0.52 (0.03)	0.25	-	-	-	-	-	-
Cg09736286	0.70 (0.05)	0.72 (0.04)	0.02	0.71 (0.05)	0.72 (0.05)	0.16	0.80 (0.04)	0.79 (0.04)	0.29	0.76 (0.04)	0.77 (0.04)	0.05
Cg02211449	0.12 (0.03)	0.12 (0.02)	0.67	0.12 (0.03)	0.11 (0.03)	0.63	0.16 (0.03)	0.17 (0.04)	0.09	0.14 (0.03)	0.14 (0.03)	0.07
Cg11385536	0.54 (0.04)	0.54 (0.04)	0.32	0.53 (0.03)	0.53 (0.04)	0.92	0.59 (0.04)	0.58 (0.04)	0.03	0.55 (0.04)	0.56 (0.04)	0.86
Cg22032485	0.48 (0.04)	0.48 (0.04)	0.84	0.50 (0.04)	0.50 (0.04)	0.34	-	-	-	-	-	-
Cg11744966	0.80 (0.02)	0.80 (0.02)	0.10	0.81 (0.02)	0.81 (0.02)	0.22	0.91 (0.01)	0.91 (0.01)	0.11	0.87 (0.02)	0.87 (0.01)	0.49
Cg00237904	0.42 (0.04)	0.41 (0.04)	0.36	0.44 (0.04)	0.43 (0.05)	0.76	0.49 (0.04)	0.50 (0.04)	0.30	0.43 (0.04)	0.43 (0.03)	0.55
Cg18332677	0.65 (0.05)	0.66 (0.04)	0.63	0.66 (0.05)	0.67 (0.06)	0.86	-	-	-	-	-	-

Note: Bold values indicate statistical significance at $p < 0.05$.

Abbreviations: LC, lung cancer; N, number; SD, standard deviation.

^aGenerated from Student's *t*-test.

TABLE 3 Association of methylation at each CpG site (per 1 SD increase) with lung cancer risk: results of random-effects meta-analysis across four subsets and comparison with previously reported ORs.¹¹

CpGs	Model 1		Model 2		Model 3		Zhao et al. ¹¹
	HR (95% CI)	p-Value	HR (95% CI)	p-Value	HR (95% CI)	p-Value	OR (95% CI)
Cg11385536 ^a	0.94 (0.76–1.17)	0.596	0.95 (0.92–0.98)	0.001	1.04 (1.01–1.07)	0.018	2.10 (1.57, 2.80)
Cg02211449 ^a	0.96 (0.76–1.21)	0.725	0.79 (0.68–0.92)	0.002	0.70 (0.63–0.78)	<0.001	0.40 (0.28, 0.58)
Cg00237904 ^a	0.94 (0.91–0.97)	<0.001	1.09 (0.96–1.24)	0.167	1.10 (0.96–1.25)	0.172	1.94 (1.50, 2.51)
Cg09736286 ^a	1.20 (0.94–1.52)	0.144	1.54 (1.03–2.30)	0.034	1.64 (1.10–2.44)	0.015	2.99 (1.95, 4.59)
Cg15979932 ^a	1.10 (0.96–1.26)	0.179	1.00 (0.83–1.22)	0.969	1.12 (0.91–1.37)	0.277	2.01 (1.53, 2.66)
Cg11744966 ^a	1.12 (0.88–1.44)	0.355	1.17 (0.89–1.53)	0.257	1.10 (0.85–1.43)	0.461	1.87 (1.47, 2.37)
Cg19907023 ^b	1.33 (0.98–1.80)	0.070	1.54 (0.94–2.52)	0.084	1.64 (1.01–2.66)	0.046	2.10 (1.56, 2.81)
Cg08739693 ^b	0.87 (0.67–1.12)	0.273	0.91 (0.68–1.20)	0.497	0.95 (0.72–1.26)	0.732	2.17 (1.62, 2.92)
Cg21432334 ^b	0.93 (0.73–1.19)	0.578	0.85 (0.61–1.17)	0.307	0.85 (0.62–1.17)	0.318	2.28 (1.69, 3.07)
Cg12483301 ^b	1.13 (0.86–1.49)	0.379	0.94 (0.62–1.43)	0.783	1.08 (0.71–1.64)	0.733	2.81 (1.88, 4.19)
Cg12796320 ^b	0.91 (0.71–1.18)	0.483	0.91 (0.62–1.34)	0.628	1.04 (0.69–1.56)	0.849	2.20 (1.60, 3.01)
Cg06581459 ^b	0.99 (0.62–1.58)	0.962	0.83 (0.44–1.57)	0.564	0.85 (0.48–1.52)	0.592	2.26 (1.70, 3.01)
Cg09447640 ^b	1.10 (0.84–1.43)	0.499	1.10 (0.75–1.61)	0.618	1.27 (0.85–1.88)	0.240	2.37 (1.69, 3.34)
Cg11343375 ^b	1.18 (0.90–1.55)	0.239	1.15 (0.77–1.73)	0.491	1.27 (0.86–1.88)	0.236	2.42 (1.74, 3.36)
Cg22032485 ^b	1.13 (0.86–1.49)	0.381	1.23 (0.85–1.78)	0.279	1.52 (1.04–2.21)	0.030	2.47 (1.73, 3.52)
Cg18332677 ^b	1.07 (0.82–1.39)	0.632	1.10 (0.60–2.04)	0.752	1.23 (0.61–2.49)	0.562	2.77 (1.86, 4.13)

Note: Model 1 included only each CpG site. Model 2 additionally adjusted for age at blood drawing, sex, batches, smoking status (never-, former, and current smoker), body mass index (kg/m²), and leukocyte composition. Model 3 additionally adjusted for the SmPEGS score. The SmPEGS score was composed of 2623 CpG sites. In our study, six and 182 CpGs were removed from 450K data and EPIC data at data preprocessing, respectively. HR and 95% CI were generated from the Cox regression model.

Abbreviations: CI, confidence interval; HR, hazard ratio; OR, odds ratio; SD, standard deviation.

^aMeta-analysis of subsets I, II, III, and IV.

^bMeta-analysis of subsets I and II.

TABLE 4 Performance of the CpGs for lung cancer incidence prediction.

CpG	AUC (95% CI)
Cg11385536 ^a	0.55 (0.50–0.60)
Cg02211449 ^a	0.54 (0.50–0.59)
Cg09736286 ^a	0.57 (0.52–0.62)
Cg19907023 ^b	0.57 (0.49–0.65)
Cg22032485 ^b	0.55 (0.47–0.63)

Abbreviations: AUC, area under the receiver-operating characteristic curve; CI, confidence interval.

^aMeta-analysis of subsets I, II, III, and IV.

^bMeta-analysis of subsets I and II.

study, indicating the complexity in identifying smoking-independent DNAm alterations. Petrovic et al.'s approach, exclusively focusing on participants' self-reported active tobacco smoking, might not account for alternative smoking exposure pathways, such as occupational exposure,²⁷ second-hand smoke exposure, and other such factors. In our study, among the five validated CpGs, cg19907023 and cg22032485 revealed their significant associations with LC risk only after adjusting for the SmPEGS score. This result highlights the importance of adjusting for comprehensive smoking-related

effects when analyzing DNAm data. The use of the SmPEGS score to account for these effects can unmask significant associations that might otherwise be overlooked.

To our knowledge, our study is the first to provide a comprehensive external validation of the smoking-independent CpG sites identified by Zhao et al. The primary strength of this validation study is the utilization of a large, well-designed population-based cohort, with extensive follow-up records and minimal missing data. Nevertheless, this study still has limitations. It was conducted exclusively among participants recruited in Germany, which may restrict its generalizability. Also, population differences between Zhao et al.'s study, which was conducted in the United States, and our study may be one of the reasons for partial inconsistencies of the findings. Additionally, DNAm levels in subsets III and IV were analyzed with the Infinium Human Methylation450K BeadChip Assay, which prevented the validation of 10 EPIC-only CpGs. Finally, most (8 out of 11) of the "nonsignificant associations" observed in our analysis were in the same direction as the associations reported by Zhao et al. for the respective CpG sites. Lack of confirmation of these associations may be due to the limited power of our study rather than the absence of such associations.

A further limitation of our study is that subgroup analyses by LC histological subtypes, which could have provided insights into

the differential roles of CpG sites, were not performed due to missing histological data and insufficient subgroup sample sizes. Future research with larger cohorts and complete histological information is needed to explore these epigenetic associations in a subtype-specific manner, potentially improving risk stratification and personalized prevention strategies.

In summary, our findings, along with the original EWAS results by Zhao et al., may offer valuable implications for researchers seeking novel DNAm biomarkers for LC that are not confounded by smoking and provide further insight into LC carcinogenesis; this study could potentially serve as a vital reference for future investigations in this field. Further research should address the potential implications of the confirmed associations with LC risk for LC risk stratification.

AUTHOR CONTRIBUTIONS

Zitong Zhao: Formal analysis; methodology; validation; writing – original draft; writing – review and editing. **Megha Bhardwaj:** Writing – review and editing. **Ziwen Fan:** Writing – review and editing. **Xianzhe Li:** Writing – review and editing. **Petra Schrotz-King:** Writing – review and editing. **Hermann Brenner:** Conceptualization; funding acquisition; supervision; writing – review and editing.

ACKNOWLEDGMENTS

The authors thank the study participants and their general practitioners as well as laboratory and administrative staff of the ESTHER study team. The authors gratefully acknowledge the contributions of the DKFZ Genomics and Proteomics Core Facility in the processing of DNA samples and performing the laboratory work. Open Access funding enabled and organized by Projekt DEAL. [Correction added on December 5 2024, after first online publication: Projekt DEAL funding statement has been added.]

FUNDING INFORMATION

The ESTHER study was supported by the Baden-Württemberg State Ministry of Science; Research and Arts (Stuttgart, Germany); the Federal Ministry of Education and Research (Berlin, Germany); the Federal Ministry of Family Affairs, Senior Citizens, Women and Youth (Berlin, Germany); and the Saarland State Ministry for Social Affairs, Health, Women and Family Affairs (Saarbrücken, Germany). The sponsors had no role in the study design; in the collection, analysis, and interpretation of data; and preparation, review, or approval of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data that support the findings of this study may be available on reasonable request from the corresponding author (H.B.). Individual data are not publicly available, as they contain information that could compromise research participant privacy/consent.

ETHICS STATEMENT

Approval of the research protocol by an Institutional Review Board: The ESTHER study has been approved by the ethics committees of the University of Heidelberg and of the State Medical Board of Saarland, Germany.

Informed Consent: All study participants provided written informed consent.

Registry and the Registration No. of the study/trial: N/a.

Animal Studies: N/a.

CONSENT

Not applicable.

ORCID

Zitong Zhao  <https://orcid.org/0000-0003-3319-756X>

REFERENCES

- Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74:229-263. doi:[10.3322/caac.21834](https://doi.org/10.3322/caac.21834)
- Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin.* 2023;73:17-48. doi:[10.3322/caac.21763](https://doi.org/10.3322/caac.21763)
- Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenetics.* 2015;7:113. doi:[10.1186/s13148-015-0148-3](https://doi.org/10.1186/s13148-015-0148-3)
- Zhang Y, Elgizouli M, Schöttker B, Holleczer B, Nieters A, Brenner H. Smoking-associated DNA methylation markers predict lung cancer incidence. *Clin Epigenetics.* 2016;8:127. doi:[10.1186/s13148-016-0292-4](https://doi.org/10.1186/s13148-016-0292-4)
- Baglietto L, Ponzi E, Haycock P, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *Int J Cancer.* 2017;140:50-61. doi:[10.1002/ijc.30431](https://doi.org/10.1002/ijc.30431)
- Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. *Nat Rev Cancer.* 2007;7:778-790. doi:[10.1038/nrc2190](https://doi.org/10.1038/nrc2190)
- Yano T, Miura N, Takenaka T, et al. Never-smoking nonsmall cell lung cancer as a separate entity. *Cancer.* 2008;113:1012-1018. doi:[10.1002/cncr.23679](https://doi.org/10.1002/cncr.23679)
- Couraud S, Zalcman G, Milleron B, Morin F, Souquet PJ. Lung cancer in never smokers—a review. *Eur J Cancer.* 2012;48:1299-1311. doi:[10.1016/j.ejca.2012.03.007](https://doi.org/10.1016/j.ejca.2012.03.007)
- Kerpel-Fronius A, Tammemägi M, Cavic M, et al. Screening for lung cancer in individuals who never smoked: an International Association for the Study of Lung Cancer early detection and screening committee report. *J Thorac Oncol.* 2022;17:56-66. doi:[10.1016/j.jtho.2021.07.031](https://doi.org/10.1016/j.jtho.2021.07.031)
- Tseng CH, Tsuang BJ, Chiang CJ, et al. The relationship between air pollution and lung cancer in nonsmokers in Taiwan. *J Thorac Oncol.* 2019;14:784-792. doi:[10.1016/j.jtho.2018.12.033](https://doi.org/10.1016/j.jtho.2018.12.033)
- Zhao N, Ruan M, Koestler DC, et al. Epigenome-wide scan identifies differentially methylated regions for lung cancer using pre-diagnostic peripheral blood. *Epigenetics.* 2022;17:460-472. doi:[10.1080/15592294.2021.1923615](https://doi.org/10.1080/15592294.2021.1923615)
- Sugden K, Hannon EJ, Arseneault L, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. *Transl Psychiatry.* 2019;9:92. doi:[10.1038/s41398-019-0430-9](https://doi.org/10.1038/s41398-019-0430-9)

13. Zhang Y, Wilson R, Heiss J, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat Commun*. 2017;8:14617. doi:[10.1038/ncomms14617](https://doi.org/10.1038/ncomms14617)
14. Gao X, Gao X, Zhang Y, Holleccek B, Schöttker B, Brenner H. Oxidative stress and epigenetic mortality risk score: associations with all-cause mortality among elderly people. *Eur J Epidemiol*. 2019;34:451-462. doi:[10.1007/s10654-019-00493-7](https://doi.org/10.1007/s10654-019-00493-7)
15. Li X, Zhang Y, Gao X, Holleccek B, Schöttker B, Brenner H. Comparative validation of three DNA methylation algorithms of ageing and a frailty index in relation to mortality: results from the ESTHER cohort study. *EBioMedicine*. 2021;74:103686. doi:[10.1016/j.ebiom.2021.103686](https://doi.org/10.1016/j.ebiom.2021.103686)
16. Bhardwaj M, Schöttker B, Holleccek B, Brenner H. Comparison of discrimination performance of 11 lung cancer risk models for predicting lung cancer in a prospective cohort of screening-age adults from Germany followed over 17 years. *Lung Cancer (Amsterdam, Netherlands)*. 2022;174:83-90. doi:[10.1016/j.lungcan.2022.10.011](https://doi.org/10.1016/j.lungcan.2022.10.011)
17. Li X, Schöttker B, Holleccek B, Brenner H. Associations of DNA methylation algorithms of aging and cancer risk: results from a prospective cohort study. *EBioMedicine*. 2022;81:104083. doi:[10.1016/j.ebiom.2022.104083](https://doi.org/10.1016/j.ebiom.2022.104083)
18. Zhang Y, Schöttker B, Florath I, et al. Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. *Environ Health Perspect*. 2016;124:67-74. doi:[10.1289/ehp.1409020](https://doi.org/10.1289/ehp.1409020)
19. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. *Environ Res*. 2016;146:395-403. doi:[10.1016/j.envres.2016.01.026](https://doi.org/10.1016/j.envres.2016.01.026)
20. Gao X, Zhang Y, Boakye D, et al. Whole blood DNA methylation aging markers predict colorectal cancer survival: a prospective cohort study. *Clin Epigenetics*. 2020;12:184. doi:[10.1186/s13148-020-00977-4](https://doi.org/10.1186/s13148-020-00977-4)
21. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86. doi:[10.1186/1471-2105-13-86](https://doi.org/10.1186/1471-2105-13-86)
22. Joehanes R, Just AC, Marioni RE, et al. Epigenetic signatures of cigarette smoking. *Circ Cardiovasc Genet*. 2016;9:436-447. doi:[10.1161/CIRCGENETICS.116.001506](https://doi.org/10.1161/CIRCGENETICS.116.001506)
23. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics*. 2016;54:1.30.1-1.30.33. doi:[10.1002/cpbi.5](https://doi.org/10.1002/cpbi.5)
24. de Vries M, Nedeljkovic I, van der Plaat DA, et al. DNA methylation is associated with lung function in never smokers. *Respir Res*. 2019;20:268. doi:[10.1186/s12931-019-1222-8](https://doi.org/10.1186/s12931-019-1222-8)
25. Petrovic D, Bodinier B, Dagnino S, et al. Epigenetic mechanisms of lung carcinogenesis involve differentially methylated CpG sites beyond those associated with smoking. *Eur J Epidemiol*. 2022;37:629-640. doi:[10.1007/s10654-022-00877-2](https://doi.org/10.1007/s10654-022-00877-2)
26. Remen T, Pintos J, Abrahamowicz M, Siemiatycki J. Risk of lung cancer in relation to various metrics of smoking history: a case-control study in Montreal. *BMC Cancer*. 2018;18:1275. doi:[10.1186/s12885-018-5144-5](https://doi.org/10.1186/s12885-018-5144-5)
27. Alhamdow A, Lindh C, Hagberg J, et al. DNA methylation of the cancer-related genes F2RL3 and AHRR is associated with occupational exposure to polycyclic aromatic hydrocarbons. *Carcinogenesis*. 2018;39:869-878. doi:[10.1093/carcin/bgy059](https://doi.org/10.1093/carcin/bgy059)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhao Z, Bhardwaj M, Fan Z, Li X, Schrotz-King P, Brenner H. Smoking-independent DNA methylation markers for lung cancer risk: External validation in a large population-based cohort study. *Cancer Sci*. 2025;116:775-782. doi:[10.1111/cas.16414](https://doi.org/10.1111/cas.16414)