Supplementary

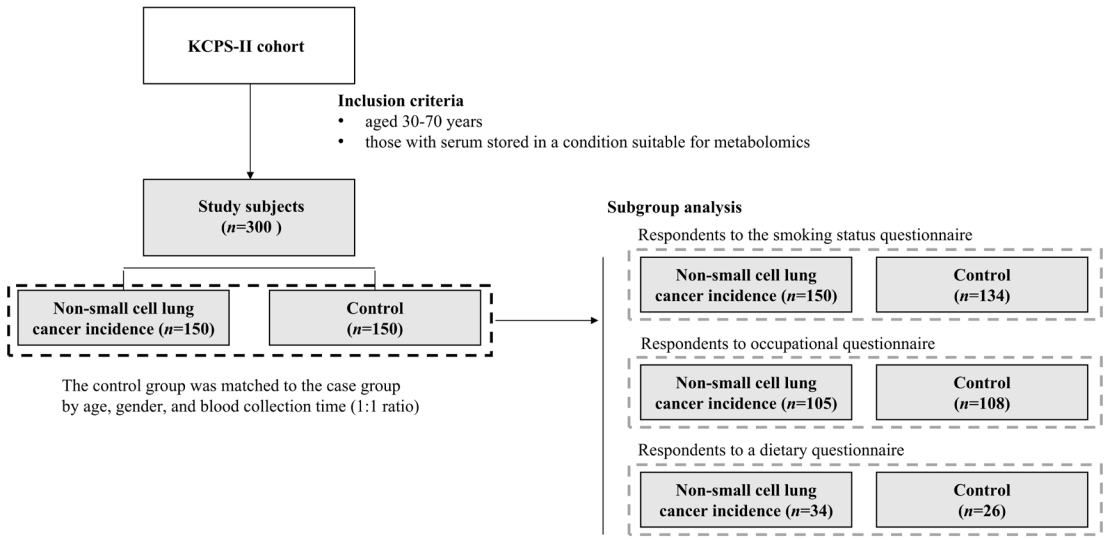

Figure S1 Participant flow chart.

https://dx.doi.org/10.21037/tlcr-2025-603

Table S1 Crosstabulation of smoking status, occupational factors, and NSCLC Incidence

Group	Smoking status (n)				Occupational factors (n)									
	Non-smoker	Ex-smoker	Current smoker	Total	Professionals	Professional-technical and administrative staff	Clerical and administrative staff	Sales and marketing workers	Service workers	Production and manual labor workers	Agricultural, forestry, and fishery workers	Homemakers (group 4)	Others (group 5)	Total
					Professional and office workers (group 1)			Sales and service workers (group 2)		Production and agricultural workers (group 3)				
Control	79	38	33	150	0	10	2	5	2	0	9	13	64	105
NSCLC incidence	52	34	48	134	1	10	8	5	2	2	9	15	56	108
Total	131	72	81	284	1	20	10	10	4	2	18	28	120	213

P from Pearson Chi-square test comparing smoking status=0.02. P from Pearson Chi-square test comparing 9 categories of occupational factors=0.51. P from Pearson Chi-square test comparing the regrouped 5 categories of occupational factors=0.66. NSCLC, non-small cell lung cancer.

https://dx.doi.org/10.21037/tlcr-2025-603

Table S2 Summary of clustering results and distribution

Cluster	Group	Type	Number	
			Train set (n=210)	Test set (n=90)
0	Cancer incidence	Adenocarcinoma/male	17	4
		Adenocarcinoma/female	34	15
		Squamous cell carcinoma/male	33	17
1	Control		86	38
	Cancer incidence	Adenocarcinom/male	1	1
		Adenocarcinoma/female	1	
2	Control		18	5
	Cancer incidence	Adenocarcinom/male	2	
3	Control		1	
	Cancer incidence	Adenocarcinom/male	17	7
Outlier	Control			2
	Cancer incidence	Adenocarcinom/male		1

Table S3 Performance evaluation

Dataset	Confusion matrix (true no cancer/ true cancer)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	Accuracy (95% CI)
Train (n=210)	TN=104, FP=1, FN=4, TP=101	0.962 (0.906–0.985)	0.990 (0.948–0.998)	0.990 (0.947–0.998)	0.976 (0.945–0.990)
Test (n=87, excl. 3 outliers)	TN=43, FP=0, FN=1, TP=43	0.977 (0.882–0.996)	1.000 (0.918–1.000)	1.000 (0.918–1.000)	0.989 (0.938–0.998)

CI, confidence interval; FN, false negative; FP, false positive; TN, true negative; TP, true positive.

https://dx.doi.org/10.21037/tlcr-2025-603

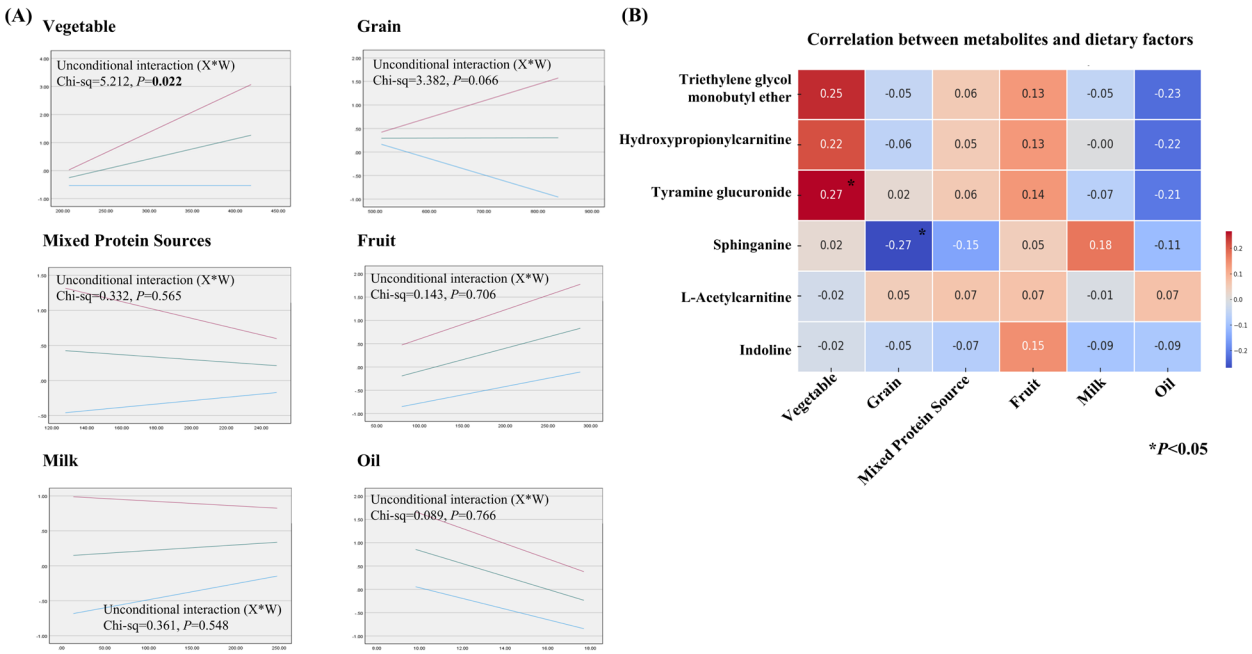

Figure S2 Subgroup analysis using dietary variables. Subgroup analysis on total lung cancer incidence (n=34) and total control (n=26). (A) Moderation effect analysis evaluating the interaction between daily dietary intake amount (g/day) and smoking status on NSCLC incidence. Independent variable: daily dietary intake amount (g/day). Moderator: smoking status. Dependent variable: lung cancer incidence (yes/no). Interaction term: daily dietary intake amount × smoking status. Lines in the plot: red line, current smoker; green line, ex-smoker; blue line, never-smoker. (B) Heatmap depicting correlation analysis between significant metabolites and dietary variables. Metabolite concentrations were log-transformed. r: Pearson's correlation coefficient. Red represents a positive correlation, and blue represents a negative correlation.

https://dx.doi.org/10.21037/tlcr-2025-603