Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Incremental value of a multidomain epidemiological questionnaire for discriminating thyroid cancer case status: a case–control study

Dehao Wang
Shandong First Medical University
Hailin Sun
Liaocheng People's Hospital
Zhen Wu
Liaocheng People's Hospital
Li Liu
Liaocheng Center for Disease Control and Prevention
Jinchao Wang
Shandong First Medical University
Zhaoqing Cui
cuizhaoqing203@163.com
Liaocheng People's Hospital

Research Article

Keywords: thyroid cancer, case–control study, epidemiological questionnaire, socioeconomic factors, lifestyle, machine learning, incremental discrimination

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11195324/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Incremental value of a multidomain epidemiological questionnaire for discriminating thyroid cancer case status: a case–control study

Dehao Wang[1], Hailin Sun[2], Zhen Wu[2], Li Liu[3], Jinchao Wang[1], Zhaoqing Cui[2]*

1 School of Medicine, Shandong First Medical University, Jinan, Shandong, China

2 Department of Thyroid, Breast and Hernia Surgery, Liaocheng Hospital Affiliated to Shandong First Medical University (Liaocheng People's Hospital), Liaocheng, Shandong, China

3 Liaocheng Center for Disease Control and Prevention, Liaocheng, Shandong, China

*Corresponding author: Zhaoqing Cui; email: cuizhaoqing203@163.com

Article type: Research article

Abstract

Background: Multidomain epidemiological questionnaires capture socioeconomic, lifestyle and environmental information, but the contribution of each domain to thyroid cancer case discrimination remains unclear. We evaluated the incremental value of a seven-domain questionnaire beyond age, sex and a limited set of literature-supported variables, and compared contributions across domains.

Methods: This single-centre case–control study included 660 participants recruited from October 2023 to September 2024, comprising 330 cases and 330 controls matched by age and sex. We used 10 repeats of five-fold nested cross-validation, with preprocessing, variable selection and tuning confined to the outer training folds. We compared elastic net logistic regression, L2-penalised logistic regression, random forest, XGBoost and support vector machines. The area under the receiver operating characteristic curve (AUROC) was

calculated from each participant's mean across 10 outer-test predictions. We estimated 95% confidence intervals using 2000 bootstrap resamples. Domain addition and leave-one-domain-out ablation analyses assessed the contribution of each domain.

Results: The support vector machine had the numerically highest AUROC, at 0.933 (95% confidence interval, 0.915–0.950). Its full model exceeded the restricted model by 0.289 AUROC units (95% confidence interval, 0.247–0.333). Domain addition and ablation analyses both ranked salt use and diet, digital device use, and demographic and socioeconomic characteristics as the three leading contributors. These domains also ranked in the top three in the elastic net sensitivity analysis.

Conclusions: The full multidomain questionnaire provided additional discriminatory information beyond the restricted variable set in this case–control sample. These findings may inform questionnaire refinement for future epidemiological surveys, but require validation in independent samples.

Keywords: thyroid cancer; case–control study; epidemiological questionnaire; socioeconomic factors; lifestyle; machine learning; incremental discrimination

Background

Thyroid cancer incidence has risen over recent decades and remains higher in women than in men [1]. In 2020, global age-standardised incidence rates were 10.1 per 100000 women and 3.1 per 100000 men. Among women in East Asia, the rate reached 17.8 per 100000 [1, 2]. Mortality has not risen in parallel. This divergence has been attributed to wider use of screening and imaging, broader routine examinations and increased detection of microcarcinomas [1, 3, 4]. Similar trends have been observed in China.

Across 35 cancer registries in mainland China, age-standardised thyroid cancer incidence in 2008–2012 was 16.8 per 100000 women and 5.3 per 100000 men. Estimates from 21 urban registries attributed 83.1% of female cases and 77.3% of male cases during this period to

overdiagnosis [5]. Updated national estimates for 2013–2017 placed these proportions at 88.7% and 83.5%, respectively [3]. Previous studies have examined radiation exposure, genetic factors and family history, diet and lifestyle, and female reproductive factors in relation to thyroid cancer [6–12]. Most focused on specific exposure types, included few variables and reported association measures such as odds ratios. Exposure definitions, recall periods and approaches to confounding also differed across studies. In other cancers, studies have combined demographic characteristics, personal and family histories, lifestyle, diet and environmental exposures for case identification and risk stratification. The combined models showed some ability to distinguish cases from controls [13–15]. For thyroid cancer, the additional discrimination provided by a full multidomain questionnaire beyond age, sex and limited literature-supported variables still lacks direct evaluation. Machine learning is increasingly used to analyse complex epidemiological data. Comparative and methodological studies nevertheless indicate that performance depends on sample size, data structure, validation design and the choice of learning algorithm [16–18].

Using questionnaire data from cases and controls, we evaluated the incremental discrimination of case status provided by a full questionnaire covering seven prespecified domains. We compared this questionnaire with age, sex and a limited set of literature-supported variables, and assessed performance across learning algorithms. Domain analyses identified the main contributors to discrimination, to inform questionnaire refinement for future epidemiological surveys.

Methods

Study design and participants

This single-centre, retrospective, matched case–control study used existing epidemiological questionnaire data for model development and internal evaluation. Cases were recruited by

convenience sampling from October 2023 to September 2024. Eligible patients had a first pathological diagnosis of thyroid cancer at the Department of Thyroid and Breast Surgery, Liaocheng People's Hospital, Shandong, China. Thyroid cancer was coded C73 under the International Classification of Diseases, tenth revision (ICD-10). Controls were drawn from people attending health examinations in the Liaocheng area during the same period. Cases and controls were recruited in a 1:1 ratio with matching on age and sex. The final sample comprised 330 cases and 330 controls. The study received ethics approval (approval number 2023279), and all participants provided written informed consent.

All participants had lived in the Liaocheng area for more than 6 months during the year preceding the survey. Cases were excluded for previous or concurrent malignancies other than thyroid cancer, severe endocrine disorders or other major illnesses. Controls were excluded for thyroid-related diseases, including thyroid adenoma, nodular goitre and thyroiditis, or other malignancies. People unable to participate in the survey were excluded from both groups.

Case status, questionnaire recall periods and candidate variables

The questionnaire was compiled for this study with reference to published literature and relevant epidemiological questionnaires [10], and was tailored to the dietary habits of residents in Liaocheng. An English-language version of the questionnaire used in this study is provided in Additional file 3.

The primary outcome was thyroid cancer case status, based on the first pathological diagnosis and coded as 1 for cases and 0 for controls. Questionnaire items covered different timeframes, including current status and the preceding 30 days; analyses retained the timeframe recorded for each item. Candidate information was organised into seven domains. D1 covered demographic and socioeconomic characteristics, and D2 captured a family history of any disease listed in the questionnaire. D3 covered lifestyle and personal exposures, and D4

covered salt use and diet. The remaining domains comprised residential, occupational and environmental factors (D5), digital device use (D6) and psychosocial context (D7). Within D6, mobile phone use habit item 2 was a second optional response slot. A blank indicated that no second option had been selected and was coded as a separate category. Age and sex were retained separately as basic design variables; the full candidate set comprised 39 original variables. Medical radiation items were used in sensitivity analyses, whereas age at menarche and any previous contraceptive use were examined in the female subgroup. Before analysis, four cured-food responses recorded as '1–3 days per occasion' and one as '4–6 days per occasion' were standardised to the questionnaire's weekly frequency options. These became '1–3 days per week' and '4–6 days per week', respectively. Original values and transformation records were retained, and no participants were excluded because of these corrections.

Model development, domain analyses and internal evaluation

We specified three models with progressively broader variable sets. The design model included only age and sex. The restricted model additionally included education and a family history of any listed disease. It also included smoking history, hair-dye use, specified occupational histories, industrial facilities near the residence or workplace, and occupational exposure. The full model included age, sex and all candidate variables from D1 to D7. The primary comparison was the difference in the area under the receiver operating characteristic curve (AUROC) between the full and restricted models. We assessed domain contributions in two ways. First, each domain was added separately to age and sex to quantify its individual contribution to discrimination. Second, each domain was removed from the full model, which was then refitted to assess that domain's contribution when all remaining questionnaire information was included.

We compared five methods: elastic net logistic regression, L2-penalised logistic regression, random forest, XGBoost and support vector machine (SVM). Internal evaluation used 10

repeats of five-fold nested cross-validation [19, 20], yielding 50 outer test folds. Each outer split allocated approximately 80% of participants to training and the remainder to testing. Five-fold cross-validation within the outer training set guided variable selection and hyperparameter tuning. Missing numeric values were imputed using medians from the corresponding training set, with missingness indicators added. Missing categorical values were treated as a separate category. Variables were standardised before fitting elastic net, L2-penalised logistic regression and SVM models; random forest and XGBoost retained the original scales. Preprocessing parameters were derived from the corresponding outer training data and then applied to the test data. Because cases and controls were equally represented, we applied neither additional class-balancing resampling nor class weights during training. Except for the design model, all models used elastic net variable selection within the corresponding outer training set [21], with age and sex always retained. Candidate values were 0.03, 0.10, 0.30, 1.00 and 3.00 for the regularisation parameter C, and 0.10, 0.50 and 0.90 for the L1 ratio. Inner cross-validation AUROC determined the parameter combination, and original variables with non-zero coefficients were retained. Elastic net also served as a candidate model that directly generated prediction scores. The selector was refitted on the entire outer training set, and the selected variable set was then fixed. Each of the other four methods underwent a second stage of inner tuning. Variable selection was not repeated within the inner splits at this second stage. Consequently, these inner AUROC values did not independently validate the complete selection and tuning process. Final evaluation used outer test participants who had not contributed to model development. For L2-penalised logistic regression, we tuned the regularisation parameter C. Random forests used 400 classification trees, with tuning of tree depth, minimum leaf size and the feature-sampling proportion [22]. XGBoost used a binary logistic objective and histogram-based tree construction, with selection from eight prespecified parameter configurations [23]. For SVMs, we compared

linear and radial basis function kernels, using C values of 0.1, 1.0 and 10.0 and tuning the relevant kernel parameters [24].

Both outer and inner cross-validation splits were defined at the participant level. For the full sample, stratification preferentially combined case status, sex and age group (<40, 40–49, 50–59 and ≥60 years). Where combinations were too sparse for five folds, stratification used case status and sex, or case status alone, in that order. For the female subgroup, stratification preferentially combined case status and age group, falling back to case status alone when necessary. Each participant received 10 outer-test predictions across the 10 repeats; these were averaged before calculating participant-level AUROC. All five candidate models used the same outer splits and evaluation participants. The model with the numerically highest participant-level AUROC was selected for the main analysis.

Model interpretability

We used SHapley Additive exPlanations (SHAP) for model interpretation [25, 26]. Permutation-based SHAP was used to explain the SVM decision function. Within each outer fold, SHAP values were calculated only for participants in that test fold. Background data comprised 10 cases and 10 controls sampled from the corresponding training set, with a budget of one permutation round per test participant. Positive SHAP values indicated a contribution towards case classification, whereas negative values indicated a contribution towards control classification. SHAP values were first aggregated to original variables and then assigned to questionnaire domains. Variables not selected in an outer training set were assigned zero contribution in that fold. Global variable importance was quantified by mean absolute SHAP values. Domain attribution was calculated by summing signed SHAP values within each domain before taking the absolute value. For individual attribution analyses, we averaged the 10 outer-test explanations available for each participant. Stability was assessed from variable rankings across the 50 outer folds and the proportion of folds in which each

variable ranked among the top 10. Local explanations were generated for the case and control whose prediction scores were closest to their respective group medians. Numerical SHAP quality and variable-selection stability are shown in Additional file 1 (Fig. S11).

Sensitivity and subgroup analyses

We repeated the progressive-model and incremental domain analyses using elastic net, and compared domain and variable attribution rankings using tree-based SHAP for XGBoost. Medical radiation variables were added to the full elastic net model to assess changes in discrimination. In the female subgroup, we fitted an age-only model and a female restricted model, then added age at menarche and any previous contraceptive use. In a further analysis, the candidate algorithm was selected within each outer training set solely on the basis of inner cross-validation AUROC. The selected model was then evaluated in the corresponding outer test fold.

Statistical analysis and software

Model discrimination was assessed using AUROC [27]. We estimated 95% confidence intervals (CIs) for AUROC and between-model differences using 2000 non-parametric bootstrap resamples at the participant level, stratified by case status [28]. In each resample, cases and controls were sampled separately with replacement, and all compared models used the same resampling indices. Pairwise comparisons among the five candidate algorithms were adjusted using the Benjamini–Hochberg method [29]. Domain addition and leave-one-domain-out ablation analyses were adjusted using the Holm method [30]. All tests were two-sided. Statistical analyses used Python 3.12.14, principally with scikit-learn 1.8.0, XGBoost 3.4.0 and SHAP 0.52.0; figures were produced in R 4.5.2. Bootstrap intervals were based on saved prediction scores, without model refitting during resampling. Reporting followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement

[31] and the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD+AI) statement [32].

Results

Study population and missing data

The study included 660 participants, with 330 cases and 330 controls; cases accounted for 50.0% of the sample (Fig. 1). Mean age and the proportion of women were similar in both groups: mean age was 49.3 years (standard deviation, 12.2), and each group included 236 women (71.5%). Junior secondary education was reported by 38.2% of cases and 24.5% of controls. Median total monthly household income was 5000 Chinese yuan (CNY) (interquartile range, 3000–9000) among cases and CNY 3000 (2150–5000) among controls. Questionnaire-listed occupations, nearby enterprises or facilities, and occupational exposures were reported by 13.7%, 13.3% and 5.5% of cases, respectively. The corresponding proportions among controls were 1.2%, 6.7% and 2.1% (Table 1). Nearby enterprises or facilities referred to those near the home or workplace. Outcome data and participant identifiers were complete for the study sample. Among the 39 full-model variables, the second optional response slot for mobile phone use habits was blank for 405 participants (61.4%). The marital relationship item was missing for 18 participants (2.7%); missingness was below 1% for all remaining variables. Supplementary figures are provided in Additional file 1. Distributions and missing or blank responses for all candidate variables are provided in Additional file 2 (Table S1). Medical radiation and female reproductive variables were used only in prespecified sensitivity or subgroup analyses (Table 7; Additional file 1, Figs. S8 and S9).

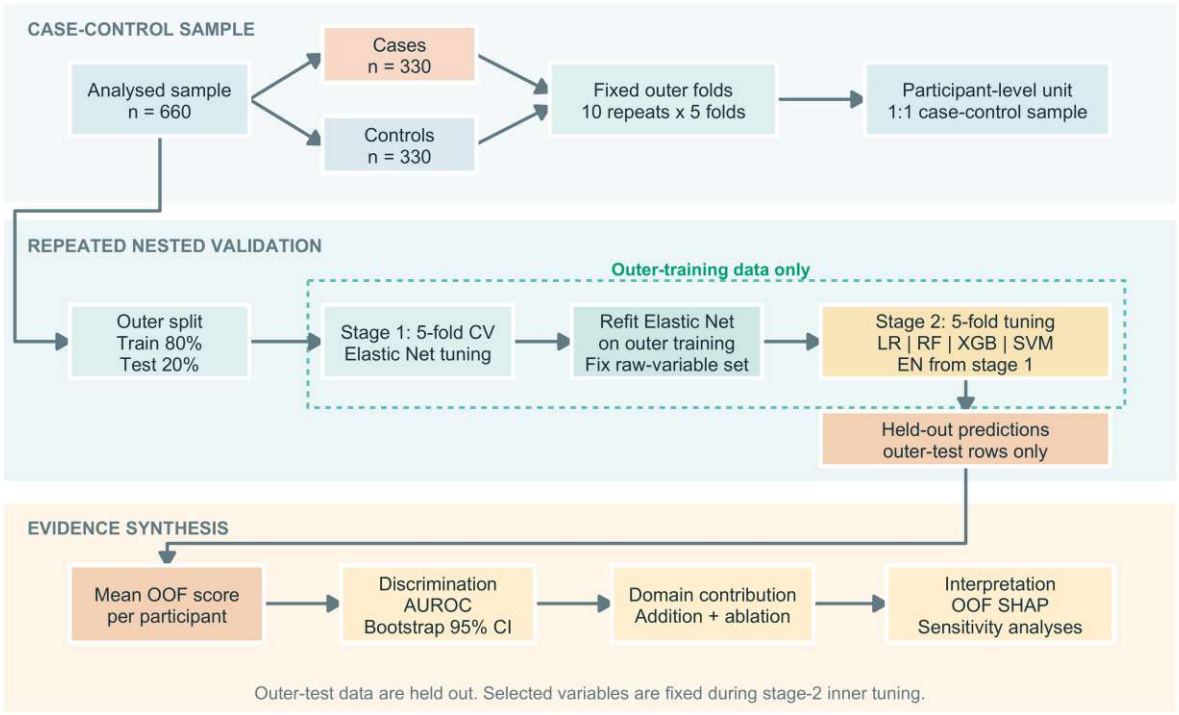


Fig. 1. Study design, sample composition and repeated nested cross-validation workflow

Cases and controls were matched by age and sex at a 1:1 ratio; model development and evaluation were performed at the participant level.

Table 1. Baseline characteristics and missing data for selected variables

Characteristic	Overall (N=660)	Cases (n=330)	Controls (n=330)	Missing, n (%)
Age, years	49.3 (12.2)	49.3 (12.2)	49.3 (12.2)	0 (0.0)
Women	472 (71.5)	236 (71.5)	236 (71.5)	0 (0.0)
Educational attainment	-	-	-	0 (0.0)
Primary school or below	180 (27.3)	80 (24.2)	100 (30.3)	-
Junior secondary school	207 (31.4)	126 (38.2)	81 (24.5)	-
Senior secondary, vocational or technical school	125 (18.9)	68 (20.6)	57 (17.3)	-
Junior college	51 (7.7)	16 (4.8)	35 (10.6)	-
University or above	97 (14.7)	40 (12.1)	57 (17.3)	-
Marital status	-	-	-	0 (0.0)
Never married	21 (3.2)	10 (3.0)	11 (3.3)	-
First marriage	606 (91.8)	319 (96.7)	287 (87.0)	-
Remarried	21 (3.2)	1 (0.3)	20 (6.1)	-

Characteristic	Overall (N=660)	Cases (n=330)	Controls (n=330)	Missing, n (%)
Divorced	3 (0.5)	0 (0.0)	3 (0.9)	-
Widowed	9 (1.4)	0 (0.0)	9 (2.7)	-
Total monthly household income, yuan/month	5000 (3000–7000)	5000 (3000–9000)	3000 (2150–5000)	5 (0.8)
Any family history of diseases listed in the questionnaire	88 (13.3)	50 (15.2)	38 (11.5)	0 (0.0)
Former or current smoking	73 (11.1)	39 (11.8)	34 (10.3)	0 (0.0)
Ever used hair dye	315 (47.7)	155 (47.0)	160 (48.5)	0 (0.0)
Ever worked in at least one occupation listed in the questionnaire	49 (7.4)	45 (13.7)	4 (1.2)	2 (0.3)
Industrial facilities near the long-term residence or workplace	66 (10.0)	44 (13.3)	22 (6.7)	0 (0.0)
Occupational exposure	25 (3.8)	18 (5.5)	7 (2.1)	0 (0.0)

Data are presented as mean (standard deviation), median (interquartile range), or n (%). Percentages for categorical variables use non-missing observations within each group as the denominator. Missing-data percentages use all 660 participants as the denominator throughout. This table is descriptive and was not used for variable selection; no between-group significance tests were performed. Distributions and missing or blank responses for all 39 full-model candidate variables are provided in Additional file 2 (Table S1).

Discrimination of candidate models

After averaging the 10 outer-test predictions for each participant, SVM had the numerically highest AUROC, at 0.933 (95% CI, 0.915–0.950). It was therefore selected as the primary model for analysis (Fig. 2; Table 2). AUROCs were 0.929 (0.910–0.946) for random forest and 0.926 (0.907–0.943) for XGBoost. Elastic net and L2-penalised logistic regression each achieved an AUROC of 0.904.

SVM exceeded both elastic net and L2-penalised logistic regression by 0.029 in AUROC, with Benjamini–Hochberg-adjusted q values of 0.005 for both comparisons. Its AUROC difference was 0.004 versus random forest (95% CI, −0.010 to 0.017; q=0.650) and 0.007

versus XGBoost (95% CI, −0.006 to 0.021; q=0.467). Both confidence intervals included 0. Performance across outer folds, prediction-score distributions and all pairwise model comparisons are shown in Additional file 1, Figs. S2, S3 and S12.

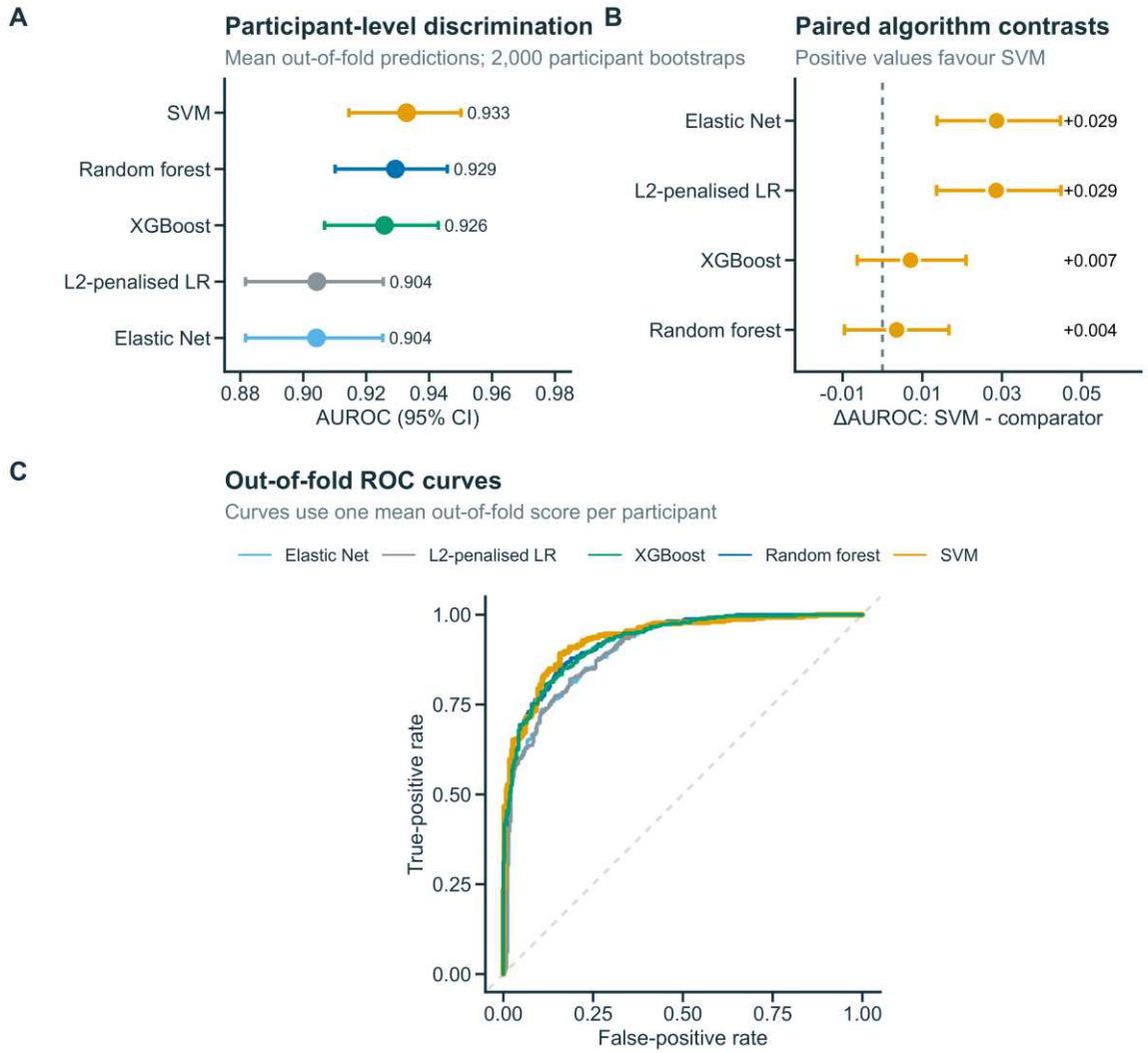


Fig. 2. Internal performance of five candidate models and AUROC differences relative to SVM
AUROC, area under the receiver operating characteristic curve; SVM, support vector machine.

Table 2. Participant-level discrimination of the five candidate models

Model	AUROC	95% CI	ΔAUROC: SVM − model	95% CI for ΔAUROC; q
Support vector machine	0.933	0.915–0.950	Reference	Not applicable
Random forest	0.929	0.910–0.946	0.004	-0.010–0.017; q=0.650
XGBoost	0.926	0.907–0.943	0.007	-0.006–0.021; q=0.467
Elastic net	0.904	0.882–0.925	0.029	0.014–0.045; q=0.005

Model	AUROC	95% CI	ΔAUROC: SVM − model	95% CI for ΔAUROC; q
L2-penalised logistic regression	0.904	0.882–0.925	0.029	0.014–0.045; q=0.005

The q values were adjusted using the Benjamini–Hochberg method for multiple comparisons. Each AUROC difference was calculated as the support vector machine AUROC minus that of the corresponding model. AUROC, area under the receiver operating characteristic curve; CI, confidence interval.

Incremental discrimination from multidomain questionnaire information

For SVM, AUROCs were 0.556 (95% CI, 0.511–0.600), 0.644 (0.603–0.688) and 0.933 (0.915–0.950) for the design, restricted and full models, respectively (Fig. 3; Table 3). The restricted model improved on the design model by 0.088 (95% CI, 0.025–0.151; P = 0.004). The full model improved on the restricted model by 0.289 (0.247–0.333; P = 0.001) and on the design model by 0.377 (0.333–0.425; P = 0.001).

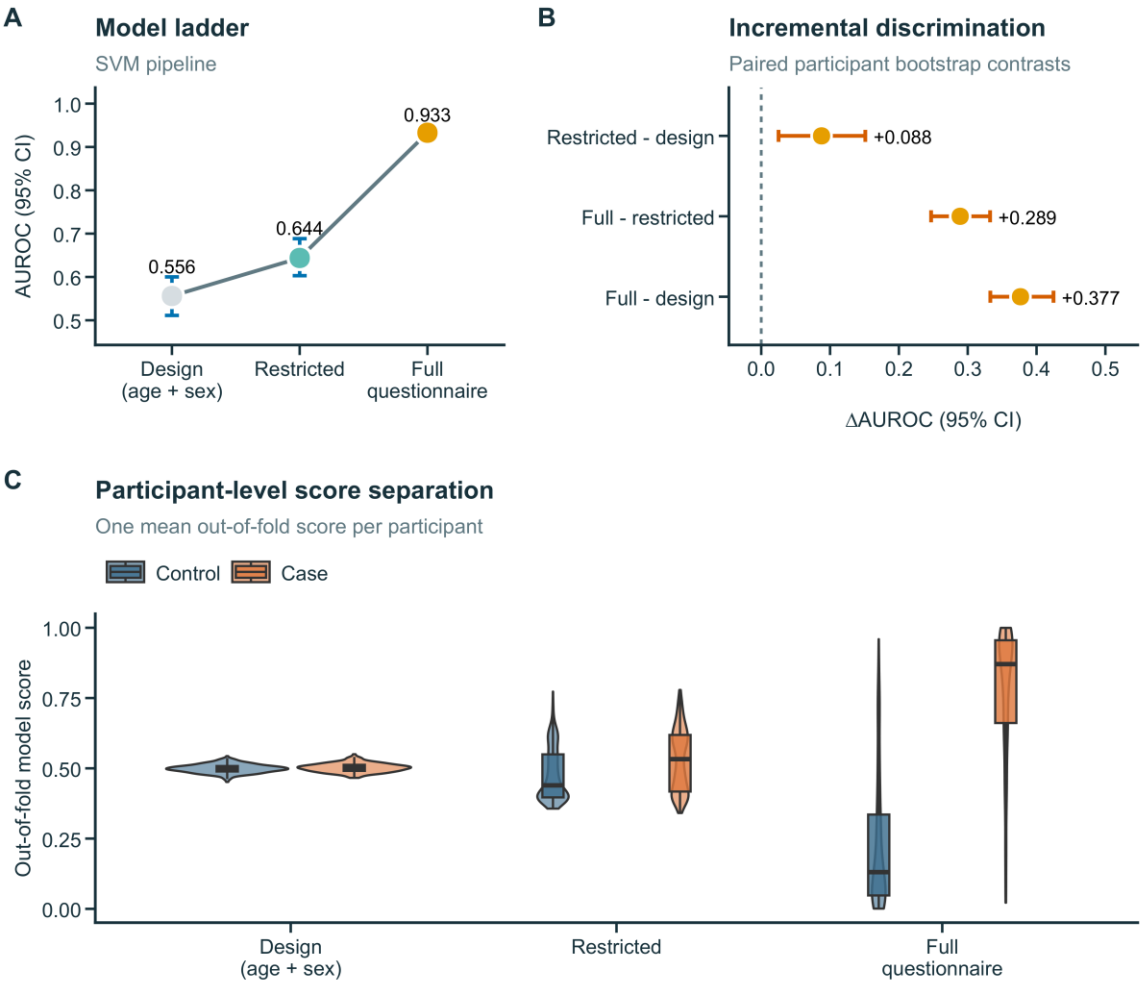


Fig. 3. Discrimination of the design, restricted and full support vector machine models

Table 3. Incremental discrimination across the three support vector machine models

Model	Variables	AUROC	95% CI	ΔAUROC vs previous model; P
Design model	Age and sex	0.556	0.511–0.600	Not applicable
Restricted model	9 basic variables	0.644	0.603–0.688	+0.088; P = 0.004
Full model	Age, sex and D1–D7, comprising 39 original variables	0.933	0.915–0.950	+0.289; P = 0.001

Confidence intervals were estimated from 2000 participant-level bootstrap samples stratified by case status. The primary incremental comparison was between the full and restricted models. AUROC, area under the receiver operating characteristic curve; CI, confidence interval.

Domain addition and leave-one-domain-out ablation

Adding digital device use (D6) to age and sex increased AUROC by 0.253 (95% CI, 0.199–0.310). Adding salt use and diet (D4) increased AUROC by 0.250 (0.195–0.305), whereas adding demographic and socioeconomic characteristics (D1) increased it by 0.182 (0.126–0.240). All three comparisons had a Holm-adjusted P value of 0.007 (Fig. 4A; Table 4). Removing D4, D6 and D1 from the full model reduced AUROC by 0.051, 0.049 and 0.025, respectively. Each reduction had a Holm-adjusted P value of 0.007 (Fig. 4B). Removing psychosocial context (D7) reduced AUROC by 0.007 (unadjusted P = 0.024; Holm-adjusted P = 0.096). Both analyses ranked D4, D6 and D1 as the top 3 domains for incremental contribution; complete comparisons are provided in Table 4.

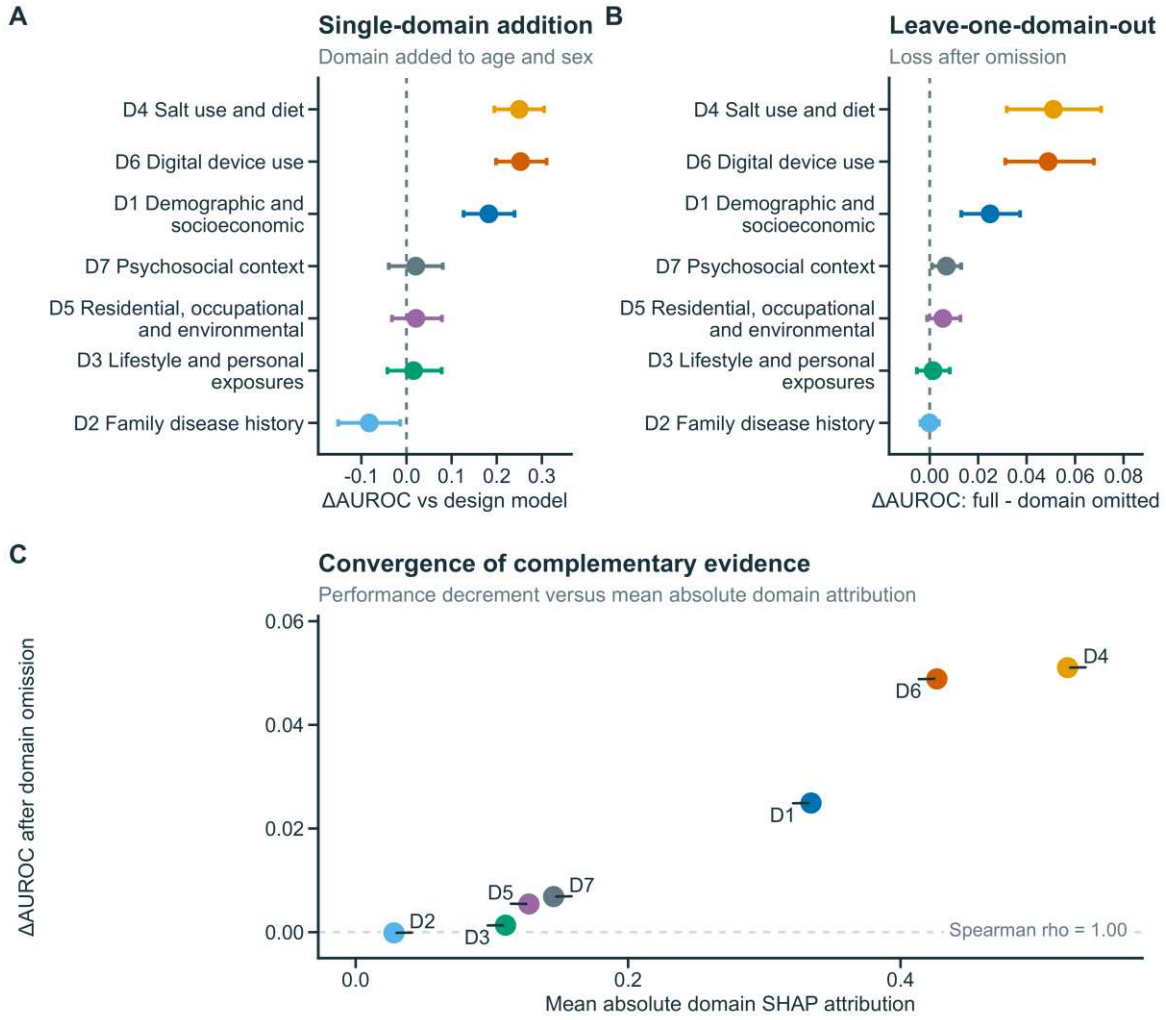


Fig. 4. Single-domain addition, leave-one-domain-out ablation and agreement between SHAP and ablation rankings

SHAP, SHapley Additive exPlanations.

Table 4. Domain addition and leave-one-domain-out ablation analyses

Domain	Addition ΔAUROC (95% CI)	Holm-adjusted P	Ablation ΔAUROC (95% CI)	Holm-adjusted P
D1 Demographic and socioeconomic characteristics	0.182 (0.126–0.240)	0.007	0.025 (0.013–0.037)	0.007
D2 Any family history of diseases listed in the questionnaire	-0.083 (-0.151–-0.014)	0.076	0.000 (-0.004–0.004)	1.000
D3 Lifestyle and personal exposures	0.016 (-0.042–0.078)	1.000	0.001 (-0.005–0.008)	1.000
D4 Salt use and diet	0.250 (0.195–0.305)	0.007	0.051 (0.032–0.071)	0.007

Domain	Addition ΔAUROC (95% CI)	Holm-adjusted P	Ablation ΔAUROC (95% CI)	Holm-adjusted P
D5 Residential, occupational and environmental factors	0.021 (-0.032–0.079)	1.000	0.005 (-0.001–0.013)	0.342
D6 Digital device use	0.253 (0.199–0.310)	0.007	0.049 (0.031–0.068)	0.007
D7 Psychosocial context	0.021 (-0.039–0.080)	1.000	0.007 (0.001–0.013)	0.096

The domain-addition analysis used the age-and-sex design model as the reference. A positive ablation value indicates a decrease in AUROC after removal of the corresponding domain. AUROC, area under the receiver operating characteristic curve; CI, confidence interval.

SHAP attribution and its stability

For SVM, mean absolute domain-level SHAP attributions were highest for D4 (0.523), followed by D6 (0.427), D1 (0.334) and D7 (0.145). D5 (0.127), D3 (0.110) and D2 (0.028) completed the descending order. This ranking matched leave-one-domain-out ablation exactly (Spearman's ρ = 1.00; Figs. 4C and 5A; Table 5).

Table 5. Domain-level SHAP attributions and ablation rankings for the support vector machine

Domain	Mean |SHAP|	SHAP rank	Ablation ΔAUROC	Ablation rank
D4 Salt use and diet	0.523	1	0.051	1
D6 Digital device use	0.427	2	0.049	2
D1 Demographic and socioeconomic characteristics	0.334	3	0.025	3
D7 Psychosocial context	0.145	4	0.007	4
D5 Residential, occupational and environmental factors	0.127	5	0.005	5
D3 Lifestyle and personal exposures	0.110	6	0.001	6
D2 Any family history of diseases listed in the questionnaire	0.028	7	0.000	7

SHAP values are expressed on the support vector machine decision-function scale. Spearman's ρ = 1.00. SHAP, SHapley Additive exPlanations.

The highest mean absolute SHAP values were observed for mobile internet duration (0.284), total monthly household income (0.281) and taste preference (0.233). Computer use (0.215)

and mobile phone use habits, item 1 (0.186), completed the top 5 (Table 6). Across 50 outer folds, these 5 variables appeared in the top 10 with frequencies of 100%, 100%, 100%, 100% and 98%, respectively (Fig. 5B,C). Signed SHAP contributions for a given variable differed in direction and magnitude across individuals. The overall attribution heatmap and representative local explanations are presented in Additional file 1, Figs. S13 and S14.

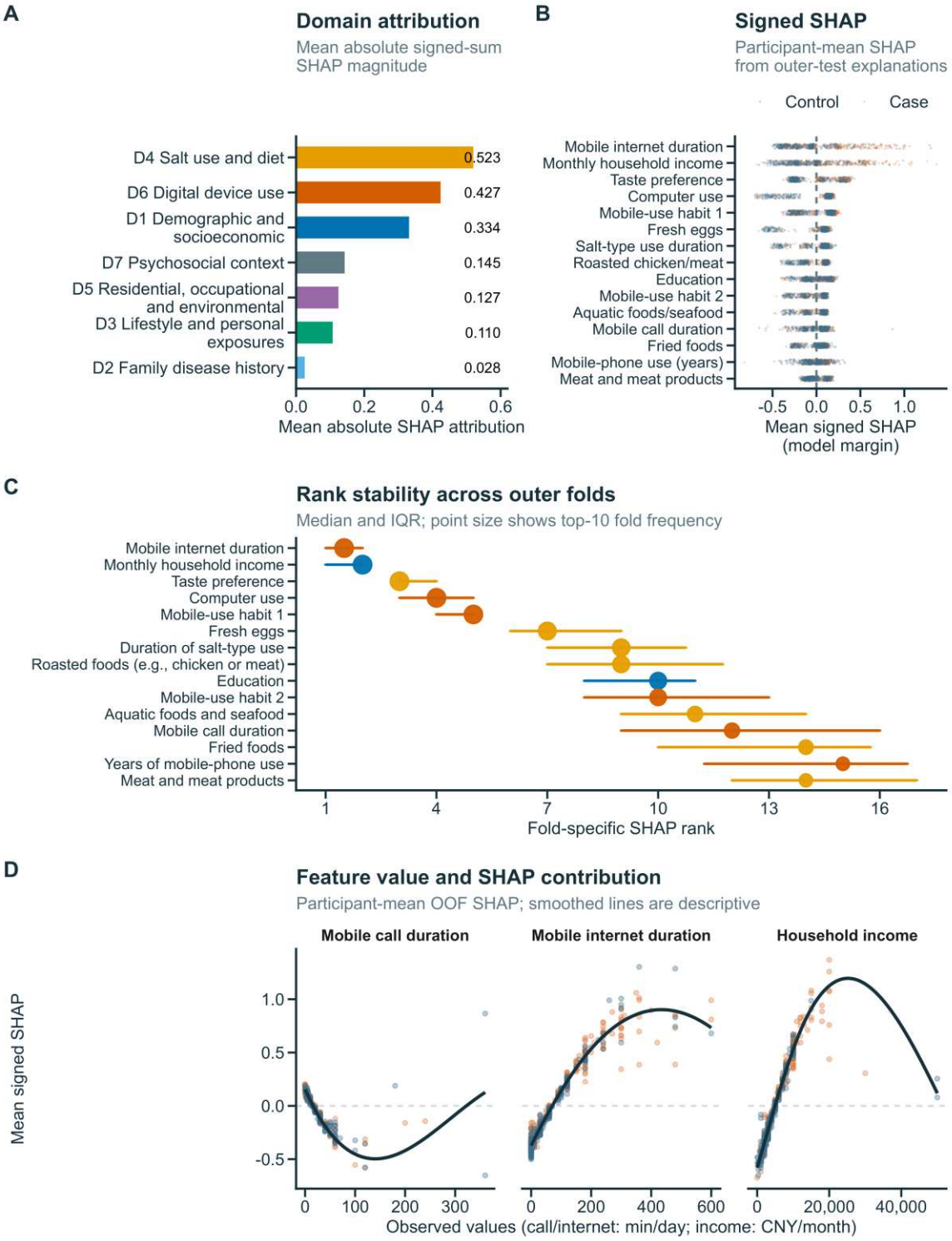


Fig. 5. Outer-test SHAP explanations and their stability for the full support vector machine model

SHAP, SHapley Additive exPlanations.

Table 6. Leading variables by SHAP importance and stability across outer folds

Variable	Domain	Mean |SHAP|	Outer folds ranking in top 10, %
Daily mobile internet use duration	D6	0.284	100%
Total monthly household income	D1	0.281	100%
Taste preference	D4	0.233	100%
Daily computer use	D6	0.215	100%
Mobile phone use habit, item 1	D6	0.186	98%
Frequency of fresh egg consumption in the past 30 days	D4	0.153	82%
Years of using the current salt type	D4	0.141	74%
Frequency of roasted food consumption (roast chicken, roast meat, etc.) in the past 30 days	D4	0.134	66%
Educational attainment	D1	0.128	64%
Mobile phone use habit, item 2	D6	0.124	58%
Frequency of aquatic food/seafood consumption in the past 30 days	D4	0.113	38%
Daily mobile phone call duration	D6	0.109	36%
Frequency of fried food consumption in the past 30 days	D4	0.102	28%
Years of mobile phone use	D6	0.099	16%
Frequency of meat and meat product consumption in the past 30 days	D4	0.097	14%

Variable importance is the mean absolute SHAP value across all outer-test explanations, with 10 explanations per participant. Stability is the proportion of the 50 outer folds in which a variable ranked among the top 10. SHAP, SHapley Additive exPlanations.

Stability of internal evaluation

Across the 50 outer-test folds, SVM achieved a median AUROC of 0.922, ranging from 0.865 to 0.961. Mean outer-test AUROC across the 10 repeats ranged from 0.915 to 0.930 (Fig. 6). Each outer-test fold contained 132 participants, comprising 66 cases and 66 controls

(Additional file 1, Fig. S1). Differences between algorithms at the fold level described internal stability only. The 50 folds were not treated as independent samples for statistical inference.

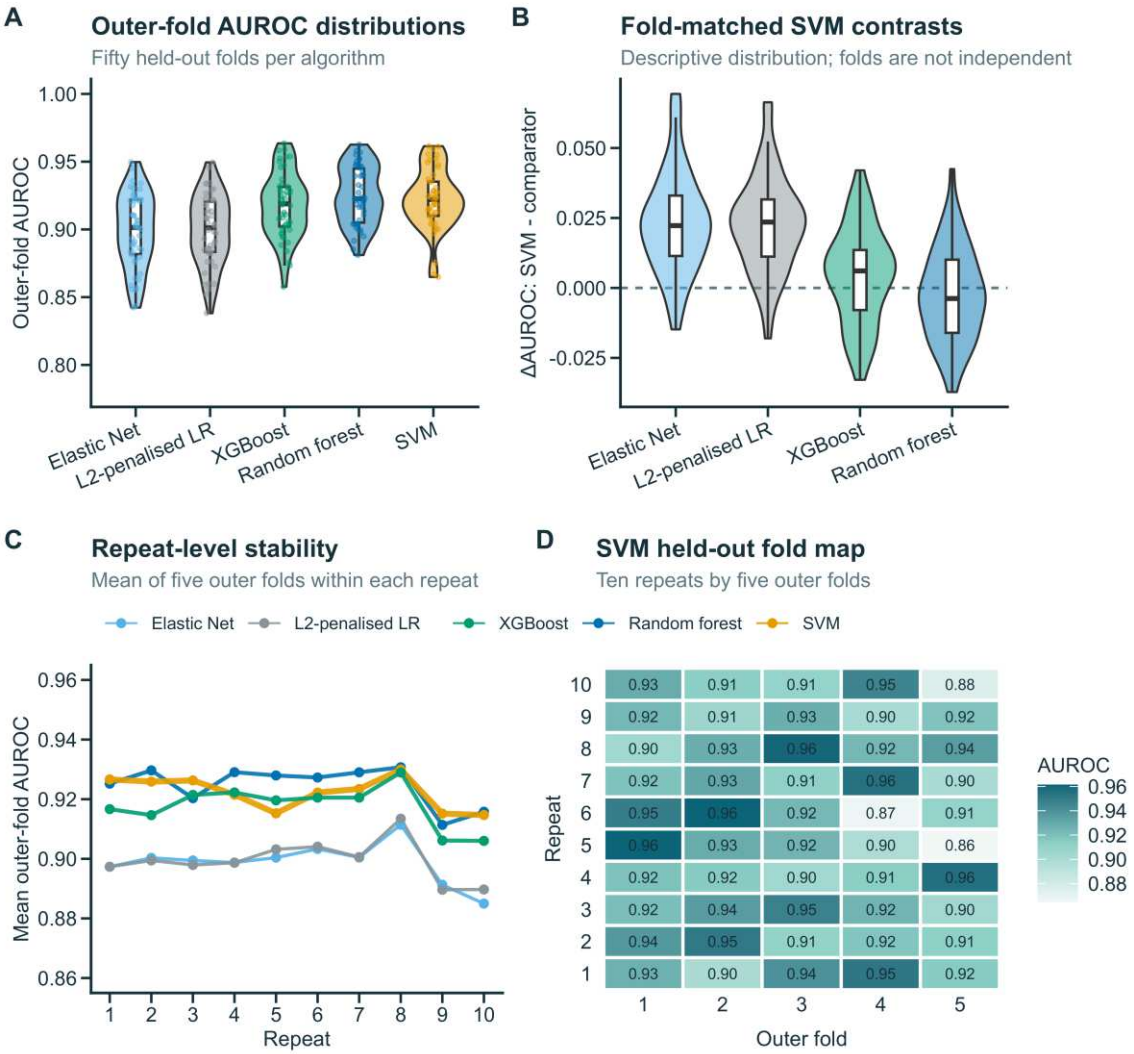


Fig. 6. Internal evaluation stability across 50 outer test folds and 10 repeats

Sensitivity and subgroup analyses

For elastic net, the design, restricted and full models achieved AUROCs of 0.498, 0.632 and 0.904, respectively. The full model exceeded the restricted model by 0.273 (95% CI, 0.228–0.314; Additional file 1, Fig. S4). Elastic net domain addition and ablation again placed D4, D6 and D1 in the top 3 (Additional file 1, Fig. S5). The domain-level TreeSHAP ranking from XGBoost correlated with the SVM domain-level SHAP ranking (Spearman's $\rho = 0.929$;

Additional file 1, Fig. S6). D6 and D4 had the highest attributions in XGBoost (Additional file 1, Figs. S6 and S7).

Adding medical radiation variables to the full elastic net model yielded an AUROC of 0.903 (95% CI, 0.879–0.925), versus 0.904 without them (Additional file 1, Fig. S8; Table 7). The female subgroup comprised 472 participants, of whom 236 were cases. The age-only model and female restricted model achieved AUROCs of 0.486 and 0.650, respectively. Adding age at menarche and any previous contraceptive use to the female restricted model yielded an AUROC of 0.692 (Additional file 1, Fig. S9).

When algorithms were selected using only inner cross-validation within each outer training set, SVM was chosen 25 times (50%) across 50 outer folds. Random forest was selected 24 times (48%) and XGBoost 1 time (2%). This analysis yielded a participant-level AUROC of 0.938 (95% CI, 0.921–0.955; Additional file 1, Fig. S10).

Table 7. Sensitivity and subgroup analyses

Model or analysis	Population	n	AUROC	95% CI	Description
Elastic net design model	All participants	660	0.498	0.453–0.543	Age and sex
Elastic net restricted model	All participants	660	0.632	0.589–0.676	9 basic variables
Elastic net full model	All participants	660	0.904	0.882–0.925	D1–D7
Full model with medical radiation variables added	All participants	660	0.903	0.879–0.925	Sensitivity analysis of medical radiation variables
Age-only model in women	Women	472	0.486	0.434–0.538	Age only
Restricted model in women	Women	472	0.650	0.601–0.699	Sex excluded
Restricted model in women with F1 added	Women	472	0.692	0.644–0.739	Age at menarche and ever use of contraception added

Model or analysis	Population	n	AUROC	95% CI	Description
Algorithm selection within each outer training set	All participants	660	0.938	0.921–0.955	Support vector machine, 25 folds; random forest, 24 folds; XGBoost, 1 fold

F1 includes age at menarche and whether contraception had ever been used. Algorithm selection within each outer training set relied only on inner cross-validation performance. AUROC, area under the receiver operating characteristic curve; CI, confidence interval.

Discussion

Expanding the restricted variable set to the full questionnaire increased SVM AUROC from 0.644 to 0.933, a gain of 0.289 (95% CI, 0.247–0.333; Table 3). Repeating the analysis with elastic net produced a gain in the same direction (Table 7). SVM discriminated better than both penalised regression models, although differences from random forest and XGBoost were not statistically significant (Table 2). Lee et al. previously combined lifestyle, occupational, reproductive and family-background factors in a case–control model of thyroid cancer among Korean women [33]. Here, we compared different sets of questionnaire information in the same participants and examined the sources of incremental discrimination. Adding salt use and diet to age and sex increased AUROC by 0.250. Removing this domain from the full model and refitting reduced AUROC by 0.051 (Table 4). These comparisons address different questions about the information a domain provides. Addition measures the gain beyond basic information; ablation measures the contribution retained when all other questionnaire domains are available [34]. Digital device use and demographic and socioeconomic characteristics also contributed in both analyses, with support after correction for multiple comparisons.

A preference for salty food was more common among cases than controls (46.1% versus 19.7%; Additional file 2, Table S1). Cases also more often reported iodised salt use and daily fresh egg consumption, but were more likely to eat kelp or laver rarely or never. This pattern cannot readily be reduced to a single food or salt choice. In a Shanghai case–control study,

Xia et al. constructed the Chinese Healthy Diet Index (CHDI) from 121 food-frequency items and household condiment consumption. Higher scores showed an L-shaped association with lower odds of thyroid cancer [9]. An Iranian case–control study identified four dietary patterns using a food-frequency questionnaire. The Western pattern was associated with higher odds of differentiated thyroid cancer, whereas the other three showed no clear associations [12]. The Shanghai Women's Health Study found that associations between dietary nitrite and thyroid cancer differed by source. Total intake showed a positive association when the highest and lowest quartiles were compared, without a monotonic gradient. Animal-source and processed-meat nitrite intake both showed positive trends [35]. Nguyen et al. observed 138 incident thyroid cancers in a prospective Korean cohort of 13973 participants. Dairy consumption frequency and meal duration were associated with incidence, whereas a preference for salty food showed no clear association [36]. These studies examined overall dietary quality, dietary patterns, or specific foods and habits in relation to thyroid cancer. Our analysis considered salt choices alongside multiple food frequencies to assess their combined information for case discrimination.

Median total monthly household income was higher among cases than controls (CNY 5000 versus 3000). Junior secondary education was also more common among cases (38.2% versus 24.5%; Table 1). Previous studies have reported differing associations between socioeconomic factors and thyroid cancer. In the China Kadoorie Biobank, higher household income and education were associated with a higher incidence of diagnosed thyroid cancer [37]. In UK Biobank, lower annual household income and higher education were both associated with greater odds of a thyroid cancer diagnosis [38]. Income and education capture different dimensions of socioeconomic circumstances, and differences in measures and study populations may also shape the observed associations.

Removing demographic and socioeconomic characteristics from the full model reduced AUROC by 0.025 (Table 4). Interpretation of this contribution requires consideration of opportunities for detection and diagnosis. A US Surveillance, Epidemiology, and End Results (SEER) study found that higher neighbourhood socioeconomic status corresponded to higher rates of diagnosed papillary thyroid cancer among people with more comprehensive insurance. This pattern also held for papillary cancers measuring 2 cm or less, but was absent in other insurance groups. The investigators suggested that differences in access to imaging and biopsy could contribute [39]. Using national Korean data, Kang et al. found higher thyroid cancer screening and diagnosis rates in higher-income groups, but higher mortality in lower-income groups [40]. Chinese cancer registry data also linked regional variation in thyroid cancer incidence to healthcare accessibility [5]. Another study estimated higher age-standardised incidence attributable to overdiagnosis in urban than rural areas [3]. Associations between household income and case status in our study may likewise partly reflect differences in opportunities for examination and diagnosis.

Cases reported a longer median daily duration of mobile internet use than controls (60 versus 30 minutes), but shorter call duration (10 versus 15 minutes). They were also more likely to report almost never using a computer (Additional file 2, Table S1). These differences therefore cannot be summarised as greater device use among cases. In a Connecticut case–control study, Luo et al. examined years of mobile phone use, daily duration and call frequency, among other measures. The confidence interval for the association between overall mobile phone use and thyroid cancer included the null [41]. A 2025 systematic review by Karipidis et al. pooled that study with two cohorts, giving a relative risk of 1.05 (95% CI, 0.88–1.26). The review rated the certainty of this evidence as low [42]. Those studies assessed exposure through mobile phone use history; our questionnaire recorded internet use, calls and computer use separately. The contribution of D6 may reflect the combined information in

these different digital activity patterns. Such activity records cannot be equated with the amount of radiofrequency exposure.

Future epidemiological surveys could prioritise salt use and diet, demographic and socioeconomic characteristics, and digital device use. Independent samples are needed to test their incremental discrimination and determine whether fewer items retain the main information. Recording thyroid examination history and routes of detection could help assess how diagnostic opportunities influence case discrimination. Clinical studies could also examine the additional value of questionnaire information alongside clinical history and ultrasound features among patients presenting with thyroid nodules.

This study has several limitations. Cases came from a single hospital, whereas controls were recruited from health examination attendees without thyroid-related disease. Differences between these sources may have influenced the observed model discrimination. Data on pathological subtype, tumour size and detection route were unavailable, limiting assessment of how much diagnostic processes explained the observed socioeconomic differences. Recent diet and digital device use were self-reported, and recall periods varied across questionnaire items. We could not establish whether all responses described the period before diagnosis. Recall bias and behavioural changes after diagnosis may therefore have affected case–control differences. Cases and controls were matched by age and sex and included at a 1:1 ratio. The sample consequently does not reflect thyroid cancer prevalence in the source population. Repeated nested cross-validation still drew on participants from the same source. The reported AUROCs primarily describe discrimination between cases and controls in this study; model scores cannot directly estimate absolute thyroid cancer risk in the general population.

Conclusions

In this sample, the full multidomain questionnaire added information for thyroid cancer case discrimination beyond the restricted variable set. Salt use and diet, digital device use, and

demographic and socioeconomic characteristics made substantial contributions. These internally evaluated findings may inform questionnaire refinement, but their incremental value requires validation in independent samples.

List of abbreviations

AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CNY, Chinese yuan; SVM, support vector machine; SHAP, SHapley Additive exPlanations; ICD-10, International Classification of Diseases, 10th Revision; CHDI, Chinese Healthy Diet Index; SEER, Surveillance, Epidemiology, and End Results; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence.

Declarations

Ethics approval and consent to participate

The Ethics Committee of Liaocheng People's Hospital approved this study (approval number, 2023279). The study followed the Declaration of Helsinki, local laws and regulations, and institutional requirements. Written informed consent was obtained from all participants in the study.

Consent for publication

Not applicable. The manuscript and supplementary materials contain no identifiable personal information.

Availability of data and materials

The raw data are not publicly available because they contain potentially identifying participant information. De-identified data and analysis code may be obtained from the

corresponding author on reasonable request, subject to approval. Approval is required from the Ethics Committee of Liaocheng People's Hospital and the relevant institutional data management authorities.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Shandong Provincial Medical and Health Science and Technology Programme (grant number 202404011050) and the Policy Guidance Project of the Liaocheng Key Research and Development Programme (grant number 2024YD16).

Authors' contributions

DW: conceptualisation, formal analysis, methodology, software, validation, visualisation, writing the original draft, and reviewing and editing the manuscript. HS: data collection and curation, investigation, resources, and reviewing and editing the manuscript. ZW: investigation, project administration, resources, and reviewing and editing the manuscript. LL: data curation, investigation, validation, and reviewing and editing the manuscript. JW: formal analysis, methodology, validation, visualisation, and reviewing and editing the manuscript. ZC: conceptualisation, funding acquisition, project administration, resources, supervision, and reviewing and editing the manuscript. All authors contributed to the study and reviewed and approved the final version for submission.

Acknowledgements

We thank all study participants for their time and contributions. We also thank staff at Liaocheng People's Hospital and the participating health examination centres for assisting with questionnaire administration and data collection.

Additional files

Additional file 1: Supplementary figures S1–S14

File name: Additional file 1 Supplementary figures.pdf. Format: PDF. Supplementary figures S1–S14, as cited in the text.

Additional file 2: Supplementary Table S1

File name: Additional file 2 Table S1.docx. Format: DOCX. Distributions and missing or blank responses for the 39 candidate variables, overall and by case status.

Additional file 3: English version of the study questionnaire

File name: Additional file 3 Questionnaire English.docx. Format: DOCX. English-language version of the thyroid cancer risk factor questionnaire used in this study.

References

1. Pizzato M, Li M, Vignat J, Laversanne M, Singh D, La Vecchia C, et al. The epidemiological landscape of thyroid cancer worldwide: GLOBOCAN estimates for incidence and mortality rates in 2020. Lancet Diabetes Endocrinol. 2022;10:264–72. https://doi.org/10.1016/S2213-8587(22)00035-3.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49. https://doi.org/10.3322/caac.21660.
3. Zhu Q, Liu J, Hu J, Zhang Y. The epidemiological landscape of thyroid cancer and estimates of overdiagnosis in China: a population-based study. Thyroid. 2025;35:307–20. https://doi.org/10.1089/thy.2024.0583.

4. Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. N Engl J Med. 2016;375:614–7. https://doi.org/10.1056/NEJMp1604412.

5. Li M, Zheng R, Dal Maso L, Zhang S, Wei W, Vaccarella S. Mapping overdiagnosis of thyroid cancer in China. Lancet Diabetes Endocrinol. 2021;9:330–2. https://doi.org/10.1016/S2213-8587(21)00083-8.

6. Saenko V, Mitsutake N. Radiation-related thyroid cancer. Endocr Rev. 2024;45:1–29. https://doi.org/10.1210/endrev/bnad022.

7. Malchoff CD. Inherited risk factors for nonmedullary thyroid carcinoma. J Clin Endocrinol Metab. 2021;106:e4287–9. https://doi.org/10.1210/clinem/dgab185.

8. Myung S-K, Lee CW, Lee J, Kim J, Kim HS. Risk factors for thyroid cancer: a hospital-based case-control study in Korean adults. Cancer Res Treat. 2017;49:70–8. https://doi.org/10.4143/crt.2015.310.

9. Xia M, Zang J, Wang Z, Wang J, Wu Y, Liu M, et al. Thyroid cancer and its associations with dietary quality in a 1:1 matched case-control study. Br J Nutr. 2023;129:283–91. https://doi.org/10.1017/S0007114522000836.

10. Wang M, Gong W-W, He Q-F, Hu R-Y, Yu M. Menstrual, reproductive and hormonal factors and thyroid cancer: a hospital-based case-control study in China. BMC Womens Health. 2021;21:13. https://doi.org/10.1186/s12905-020-01160-w.

11. Cho YA, Kim J. Dietary factors affecting thyroid cancer risk: a meta-analysis. Nutr Cancer. 2015;67:811–7. https://doi.org/10.1080/01635581.2015.1040517.

12. Sangsefidi ZS, Ghafouri-Taleghani F, Zakavi SR, Norouzy A, Kashanifar R, Pourbaferani R, et al. Major dietary patterns and differentiated thyroid cancer. Clin Nutr ESPEN. 2019;33:195–201. https://doi.org/10.1016/j.clnesp.2019.05.015.

13. Wang X, Ma K, Cui J, Chen X, Jin L, Li W. An individual risk prediction model for lung cancer based on a study in a Chinese population. Tumori. 2015;101:16–23. https://doi.org/10.5301/tj.5000205.
14. Tse LA, Wang F, Wong MC-S, Au JS-K, Yu IT-S. Risk assessment and prediction for lung cancer among Hong Kong Chinese men. BMC Cancer. 2022;22:585. https://doi.org/10.1186/s12885-022-09678-y.
15. Deng J-W, Zhou Y-L, Dai W-X, Chen H-M, Zhou C-B, Zhu C-Q, et al. Noninvasive predictive models based on lifestyle analysis and risk factors for early-onset colorectal cancer. J Gastroenterol Hepatol. 2023;38:1768–77. https://doi.org/10.1111/jgh.16243.
16. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.
17. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14:137. https://doi.org/10.1186/1471-2288-14-137.
18. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res. 2019;28:2455–74. https://doi.org/10.1177/0962280218784726.
19. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91.
20. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:2079–107.

21. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67:301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
22. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
24. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1007/BF00994018.
25. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765–74.
26. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
27. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.
28. Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7:1–26. https://doi.org/10.1214/aos/1176344552.

29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
30. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6:65–70.
31. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147:573–7. https://doi.org/10.7326/0003-4819-147-8-200710160-00010.
32. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.
33. Lee SM, Kwak KH. Risk factors and a predictive model for thyroid cancer in Korean women. Cancer Nurs. 2010;33:310–9. https://doi.org/10.1097/NCC.0b013e3181cd2844.
34. Williamson BD, Gilbert PB, Simon NR, Carone M. A general framework for inference on algorithm-agnostic variable importance. J Am Stat Assoc. 2023;118:1645–58. https://doi.org/10.1080/01621459.2021.2003200.
35. Aschebrook-Kilfoy B, Shu X-O, Gao Y-T, Ji B-T, Yang G, Li HL, et al. Thyroid cancer risk and dietary nitrate and nitrite intake in the Shanghai Women's Health Study. Int J Cancer. 2013;132:897–904. https://doi.org/10.1002/ijc.27659.
36. Nguyen LTD, Gunathilake M, Lee J, Kim J. Association between dietary habits and incident thyroid cancer: a prospective cohort study. Front Nutr. 2023;10:1104925. https://doi.org/10.3389/fnut.2023.1104925.

588 37. Lin J, Yang J, Liang M, Chen Y, Huang W, Wu L, et al. Association between
589 socioeconomic position and thyroid cancer incidence: a population-based cohort study in
590 China. Cancer Epidemiol Biomarkers Prev. 2026;35:752–61.
591 https://doi.org/10.1158/1055-9965.EPI-25-1875.
592 38. Li Y, Zhan Y, Mao W, Wang B, Dong P, Na R. Interaction between genetic risks and
593 socioeconomic factors on thyroid cancer: evidence from 0.5 million UK Biobank
594 participants. Cancers (Basel). 2023;15:5028. https://doi.org/10.3390/cancers15205028.
595 39. Altekruse S, Das A, Cho H, Petkov V, Yu M. Do US thyroid cancer incidence rates
596 increase with socioeconomic status among people with health insurance? An
597 observational study using SEER population-based data. BMJ Open. 2015;5:e009843.
598 https://doi.org/10.1136/bmjopen-2015-009843.
599 40. Kang H-Y, Kim I, Kim Y-Y, Bahk J, Khang Y-H. Income differences in screening,
600 incidence, postoperative complications, and mortality of thyroid cancer in South Korea:
601 a national population-based time trend study. BMC Cancer. 2020;20:1096.
602 https://doi.org/10.1186/s12885-020-07597-4.
603 41. Luo J, Deziel NC, Huang H, Chen Y, Ni X, Ma S, et al. Cell phone use and risk of
604 thyroid cancer: a population-based case-control study in Connecticut. Ann Epidemiol.
605 2019;29:39–45. https://doi.org/10.1016/j.annepidem.2018.10.004.
606 42. Karipidis K, Baaken D, Loney T, Blettner M, Mate R, Brzozek C, et al. The effect of
607 exposure to radiofrequency fields on cancer risk in the general and working population:
608 a systematic review of human observational studies – part II: less researched outcomes.
609 Environ Int. 2025;196:109274. https://doi.org/10.1016/j.envint.2025.109274.

Supplementary Files

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

- Additionalfile1Supplementaryfigures.pdf
- Additionalfile2TableS1.docx
- Additionalfile3QuestionnaireEnglish.docx