

RESEARCH

Open Access



# Development of a gastric cancer risk calculator for questionnaire-based surveillance of Iranian dyspeptic patients

Kimiya Gohari<sup>1,2</sup>, Samaneh Saberi<sup>1</sup>, Maryam Esmaili<sup>1</sup>, Mohammad Tashakoripour<sup>3</sup>, Mahmoud Eshagh Hosseini<sup>3</sup>, Azin Nahvijou<sup>4</sup>, Mohammad Ali Mohagheghi<sup>4</sup>, Anoshirvan Kazemnejad<sup>2\*</sup> and Marjan Mohammadi<sup>1\*</sup>

## Abstract

**Background** Gastric cancer (GC) is considered a silent killer, taking more than three quarters of a million lives annually. Therefore, prior to further costly and invasive diagnostic approaches, an initial GC risk screening is desperately in demand.

**Methods** In order to develop a simple risk scoring system, the demographic and lifestyle indices from 858 GC and 1132 non-ulcer dyspeptic (NUD) patients were analysed. We applied a multivariate logistic regression approach to identify the association between our target predictors and GC *versus* NUD. The model performance in classification was assessed by receiver operating characteristic (ROC) analysis. Our questionnaire covering 64 predictors, included known risk factors, such as demographic features, dietary habits, self-reported medical status, narcotics use, and SES indicators.

**Results** Our model segregated GC from NUD patients with the sensitivity, specificity, and accuracy rates of 85.89, 63.9, and 73.03%, respectively, which was confirmed in the development dataset (AUC equal to 86.37%,  $P < 0.0001$ ). Predictors which contributed most to our GC risk calculator, based on risk scores (RS) and shared percentages (SP), included: 1) older age group [ $> 70$  (RS:+ 241, SP:7.23), 60–70 (RS:+ 221, SP:6.60), 50–60 (RS:+ 134, SP:4.02), 2) history of gastrointestinal cancers (RS:+ 173, SP:5.19), 3) male gender (RS:+ 119, SP:3.55), 4) non-Fars ethnicity (RS:+ 89, SP:2.66), 5) illiteracy of both parents (RS:+ 78, SP:2.38), 6) rural residence (RS:+ 77, SP:2.3), and modifiable dietary behaviors (RS:+ 32 to + 53, SP:0.96 to 1.58).

**Conclusion** Our developed risk calculator provides a primary screening step, prior to the subsequent costly and invasive measures. Furthermore, public awareness regarding modifiable risk predictors may encourage and promote lifestyle adjustments and healthy behaviours.

**Keywords** Gastric cancer, Nonulcer dyspepsia, Risk prediction, Calculator, Logistic regression

Marjan Mohammadi is the main corresponding author and Anoshirvan Kazemnejad is the co-corresponding author.

\*Correspondence:

Anoshirvan Kazemnejad  
kazem\_an@modares.ac.ir

Marjan Mohammadi  
marjan.mohammadi@pasteur.ac.ir; marjan.mohammadi2010@gmail.com  
Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Introduction

According to the International Agency for Research on Cancer (IARC), gastric cancer (GC) is responsible for more than 769,000 global deaths, equating to one in every 13 deaths, for the year 2020 [1]. GC is more prevalent amongst the male population, such that ~49 in 100,000 males suffer from this disease, which is more than twice its prevalence in females (~21 in 100,000) [2]. Stomach cancer mainly involves older people, with the average age of diagnosis being 68 and more than half of people diagnosed are 65 or older [3].

GC is a multistep and multifactorial process involving genetic and environmental factors [4]. Besides age and gender as known risk factors, there is much evidence that unhealthy diets [5, 6], alcohol abuse [7], smoking [8–10] and other factors such as genetics, environmental and behavioural factors [11–14] enhance the risk of GC development.

Considering the incidence rates of GC in most countries are expected to decrease through 2030, reductions in smoking, prevalence of *Helicobacter pylori* infection and diet improvement will be the likely contributing factors [15].

It is most desired to estimate the primary disease risk using general information, consuming the least time and resources [16]. This can be made possible by combining clinical knowledge with applied data science [17]. The ultimate product should be an optimal tool, readily performed by anyone without expert knowledge, using their personal information. Ideally, the application of such tools will help increase awareness and ultimately reduce the burden of disease on the community and the health care system [18].

Classification methods are usually used to develop a risk score and identify high-risk individuals in a population [19–22]. In this study, we have used multivariate analysis to identify individual predictors differentiating GC from NUD patients. For this purpose, we carried out a logistic regression approach, using 64 predictors of known risk factors including demographic features, dietary habits, self-reported medical status, narcotics use, and SES indicators. Developing a time and cost-effective algorithm that uses the personal medical history and lifestyle habits to screen subjects for GC risk, can provide a tool for filtering dyspeptic patients prior to the more invasive screening approaches.

## Materials and methods

### Study setting

This hospital-based observational study was conducted on a group of Iranian gastric cancer (GC) patients ( $n=858$ ), who were consecutively (July 2003 to Jan 2020) referred to the National cancer Institute of Iran (NCII).

Our GC cases were diagnosed with histologically confirmed gastric adenocarcinoma. The non-ulcer dyspeptic (NUD) patients ( $n=1132$ ) were those who had referred for upper gastroscopy, but lacked GC. NUD patients were admitted at the endoscopy unit of Amiralam Hospital. Both centers shared similar SES profiles. The anatomic location (subsite) of the tumor was classified as cardia (defined as cardioesophageal junction, oesophago-gastric junction and gastroesophageal junction) or non-cardia (all other locations in the stomach) [23]. Histopathologic studies identified the subtype of the gastric tumors, as intestinal or diffuse [24].

Trained technicians interviewed each participant at the time of recruitment, using a structured questionnaire. This questionnaire elicited 64 predictors, including demographic features, dietary habits, self-reported medical status, narcotics use, and SES indicators. Primarily, each of the questionnaire predictors, with multiple levels (with the exception of age) was turned into binary groups (S-Table-1).

In order to use the properties of data while assuming the power of 90% for testing the significance of the odds ratio in the logistic regression model, a minimum of 936 observations were required. Therefore, our data, including 1990 observations, had sufficient power for the risk score development.

## Statistical analysis

### Imputation of the missing data

We used multivariate imputation by chained equations (MICE) [25], to deal with missing data in more than one variable. In this method, two general approaches for imputing multivariate data have been applied: joint modeling (JM) [computational strategies for multivariate linear mixed-effects models with missing values [26], multilevel models with multivariate mixed response types] and full conditional specification (FCS) [multivariate imputation by chained equations-dependency networks for inference, collaborative filtering, and data visualisation [27].

To validate our imputation method, we conducted the following steps. At first, using the bootstrap method [28], based on the distribution of data, we have generated ten copies of our dataset. The multivariate imputation method imputed the missing values in each copy, and five new complete datasets were generated for all the copies. In this manner, we achieved 50 complete datasets. The distribution of all variables in the original dataset and these 50 imputed versions were compared. The variables were kept in the dataset, if the deviation in the mean (S-Fig. 1) and standard deviation (S-Fig. 2) did not exceed 0.05. Next, we randomly converted 10% of the observed values for each variable into missing and imputed

them again. This process was repeated 1000 times and all the variables' biases were calculated (S-Fig. 3). The cut-off value for the bias variation was set at 2%. We aimed to maintain the imputation bias under this cut-off value.

### Model development

We used Chi-square test to measure associations between predictors and outcomes (Table 1). Statistical significance was determined using 2-sided *P*-values, with values <0.05 considered as statistically significant. We have presented a univariate analysis of all predictors and assessed the association between each of them with GC *vs.* NUD, without taking into consideration the other predictors. In this step, we emphasized on the distribution of each (Table 1). In the next step we performed multivariate logistic regression analysis on 70% of randomly selected observations, and determined the associations with each predictor, while adjusting for all others (Table 2) [29]. The probability of having GC *vs.* NUD, based on the logistic model, was calculated [30].

The probability of being GC *versus* NUD was computed using logistic regression:

$$P(GC) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}$$

Where  $\beta_0$  is the intercept term and  $\beta_1, \beta_2, \dots, \beta_k$  are the coefficients associated with the input features  $X_1, X_2, \dots, X_k$ .

This study divided patients into two risk groups based on an assigned cut-off point, derived from fixing the sensitivity rate at a minimum of 90%, while maximizing the specificity rate. Accordingly, the best threshold for the risk score was identified. We defined the shared percentage for every predictor in our risk calculator, as the contribution of each variable in predicting GC *vs.* NUD, as clinical outcomes. This measure is the proportion of the standardized regression coefficient (point estimates) for each predictor relative to their total sum (Table 2 and Fig. 1). The final risk score for each predictor was calculated by the multiplication of their pertinent point estimate by 100.

### Model validation

We used the train-test split method [31] for determining the performance criteria (AUC, sensitivity, specificity, precision, false-positive, false-negative, and accuracy rates) of our logistic model (Fig. 2), as well as to assess its internal validity.

These performance criteria were calculated as follows:

The accuracy rate, which measures the overall correctness of the classification, was calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The sensitivity rate (true positive rate or recall), which measures the proportion of actual positive instances that were correctly identified, was calculated as:

$$Sensitivity = \frac{TP}{TP + FN}$$

The specificity rate (true negative rate), which measures the proportion of actual negative instances that were correctly identified, was calculated as:

$$Specificity = \frac{TN}{TN + FP}$$

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

To do this, the data were randomly divided into the development (70%) and validation (30%) subsets. The performance criteria of our GC risk calculator were determined by examining calibration and discrimination measures. Calibration refers to how closely the predicted probability of having GC agrees with the observed GC status and is assessed by the Hosmer-Lemeshow test [32]. The discrimination rate expresses the ability of the model to differentiate between individuals with GC *versus* NUD. This was evaluated by calculating the area under the ROC curve (AUC) [33]. An AUC value of 50 and 100 was considered as having no *versus* perfect discrimination, respectively. Risk thresholds that gave a combination of more than 85% sensitivity rates and maximum specificity rates were derived from the list provided by the ROC curve analysis. All statistical analysis and data visualizations were done in R statistical software environment.

## Results

### Descriptive information

Our observational study included 858 GC [development=610 and validation=248] and 1132 NUD [development=783 and validation=349] patients, who were entered into this hospital-based study.

All of the 64 questionnaire predictors, with multiple levels, were converted into binary categories, as presented in S-Table 1. We have also presented the distribution of each of these predictors amongst GC *versus* NUD patients, without any adjustments for other predictors in Table 1. The results of the Chi-square test showed that the distribution of most (47/64) of the predictors were different between GC and NUD patients (Table 1). The association between predictors and GC *vs.* NUD affected the model, and although most variables were independently associated with GC (Table 1),

**Table 1** Distribution of predictors amongst GC versus NUD patients

	Predictors	levels	Overall, N= 1990 n (%)	GC, N= 858 n (%)	NUD, N= 1132 n (%)	P value <sup>1</sup>
<b>1. Demographic</b>						
1	Age	≤ 50 > 50 – 60 > 60 – 70 > 70	824 (41%) 477 (24%) 431 (22%) 258 (13%)	143 (17%) 214 (25%) 308 (36%) 193 (22%)	681 (60%) 263 (23%) 123 (11%) 65 (5.7%)	<b>&lt; 0.001</b>
2	Ethnicity	Fars Non-Fars	642 (32%) 1348 (68%)	177 (21%) 681 (79%)	465 (41%) 667 (59%)	<b>&lt; 0.001</b>
3	Gender	Female Male	878 (44%) 1112 (56%)	223 (26%) 635 (74%)	655 (58%) 477 (42%)	<b>&lt; 0.001</b>
<b>2. Diet</b>						
4	Canned food	Never Ever	1317 (66%) 673 (34%)	581 (68%) 277 (32%)	736 (65%) 396 (35%)	0.201
5	Carbonated soft drinks	Never to low Medium to high	1576 (79%) 414 (21%)	690 (80%) 168 (20%)	886 (78%) 246 (22%)	0.212
6	Cheese	Never to low Medium to high	574 (29%) 1416 (71%)	274 (32%) 584 (68%)	300 (27%) 832 (73%)	<b>0.008</b>
7	Chicken	Never to low Medium to high	1119 (56%) 871 (44%)	486 (57%) 372 (43%)	633 (56%) 499 (44%)	0.732
8	Coffee	Never to low Medium to high	1937 (97%) 53 (2.7%)	846 (99%) 12 (1.4%)	1091 (96%) 41 (3.6%)	<b>0.002</b>
9	Cooking method	Boiling Other than boiling	811 (41%) 1179 (59%)	359 (42%) 499 (58%)	452 (40%) 680 (60%)	0.401
10	Cooking oil	Unsaturated Saturated or both	740 (37%) 1250 (63%)	243 (28%) 615 (72%)	497 (44%) 635 (56%)	<b>&lt; 0.001</b>
11	Cooking salt	Never to low Medium to high	626 (31%) 1364 (69%)	237 (28%) 621 (72%)	389 (34%) 743 (66%)	<b>0.001</b>
12	Dinner time	Early Late	1279 (64%) 711 (36%)	495 (58%) 363 (42%)	784 (69%) 348 (31%)	<b>&lt; 0.001</b>
13	Drinking water (childhood)	City plumbing Other than city plumbing	821 (41%) 1169 (59%)	221 (26%) 637 (74%)	600 (53%) 532 (47%)	<b>&lt; 0.001</b>
14	Eggs	Never to low Medium to high	1265 (64%) 725 (36%)	481 (56%) 377 (44%)	784 (69%) 348 (31%)	<b>&lt; 0.001</b>
15	Fish	Never to low	516 (26%)	224 (26%)	292 (26%)	0.902

**Table 1** (continued)

	Predictors	levels	Overall, N = 1990 n (%)	GC, N = 858 n (%)	NUD, N = 1132 n (%)	P value <sup>1</sup>
16	Fruits	Medium to high	1474 (74%)	634 (74%)	840 (74%)	<b>&lt; 0.001</b>
		Never to low	1503 (76%)	611 (71%)	892 (79%)	
17	Milk	Medium to high	487 (24%)	247 (29%)	240 (21%)	<b>&lt; 0.001</b>
		Never to low	1332 (67%)	526 (61%)	806 (71%)	
18	Minerals	Medium to high	658 (33%)	332 (39%)	326 (29%)	<b>&lt; 0.001</b>
		Never	426 (21%)	114 (13%)	312 (28%)	
19	Pickled vegetables	Ever	1564 (79%)	744 (87%)	820 (72%)	0.245
		Never to low	414 (21%)	166 (19%)	248 (22%)	
20	Processed meats	Medium to high	1576 (79%)	692 (81%)	884 (78%)	<b>&lt; 0.001</b>
		Never to low	1159 (58%)	555 (65%)	604 (53%)	
21	Potato chips	Medium to high	831 (42%)	303 (35%)	528 (47%)	<b>&lt; 0.001</b>
		Never	1337 (67%)	629 (73%)	708 (63%)	
22	Red meat	Ever	653 (33%)	229 (27%)	424 (37%)	0.413
		Never to low	1076 (54%)	474 (55%)	602 (53%)	
23	Salted food	Medium to high	914 (46%)	384 (45%)	530 (47%)	0.054
		Never	1842 (93%)	783 (91%)	1059 (94%)	
24	Smoked fish	Ever	148 (7.4%)	75 (8.7%)	73 (6.4%)	<b>&lt; 0.001</b>
		Never	1806 (91%)	756 (88%)	1050 (93%)	
25	Smoked rice	Ever	184 (9.2%)	102 (12%)	82 (7.2%)	0.611
		Never	1781 (89%)	764 (89%)	1017 (90%)	
26	Table salt	Ever	209 (11%)	94 (11%)	115 (10%)	<b>&lt; 0.001</b>
		Never	1234 (62%)	450 (52%)	784 (69%)	
27	Tea	Ever	756 (38%)	408 (48%)	348 (31%)	<b>0.023</b>
		Never to low	53 (2.7%)	15 (1.7%)	38 (3.4%)	
28	Tea temperature	Medium to high	1937 (97%)	843 (98%)	1094 (97%)	<b>&lt; 0.001</b>
		Cold to warm	1281 (64%)	450 (52%)	831 (73%)	
29	Tuna fish	Hot	709 (36%)	408 (48%)	301 (27%)	<b>&lt; 0.001</b>
		Never	925 (46%)	442 (52%)	483 (43%)	
30	Vegetables	Ever	1065 (54%)	416 (48%)	649 (57%)	> 0.901
		Never to low	545 (27%)	236 (28%)	309 (27%)	
31	Vitamins	Medium to high	1445 (73%)	622 (72%)	823 (73%)	<b>&lt; 0.001</b>
		Never	389 (20%)	93 (11%)	296 (26%)	

**Table 1** (continued)

	Predictors	levels	Overall, N = 1990 n (%)	GC, N = 858 n (%)	NUD, N = 1132 n (%)	P value <sup>1</sup>
32	Yoghurt	Ever	1601 (80%)	765 (89%)	836 (74%)	<b>&lt; 0.001</b>
		Never to low	477 (24%)	159 (19%)	318 (28%)	
		Medium to high	1513 (76%)	699 (81%)	814 (72%)	
<b>3. Medical status (self-reported)</b>						
33	Colitis					0.825
		No	1968 (99%)	849 (99%)	1119 (99%)	
		Yes	22 (1.1%)	9 (1.0%)	13 (1.1%)	
34	Diabetes					> 0.932
		No	1850 (93%)	797 (93%)	1053 (93%)	
		Yes	140 (7.0%)	61 (7.1%)	79 (7.0%)	
35	Esophageal reflux					<b>0.025</b>
		No	1572 (79%)	698 (81%)	874 (77%)	
		Yes	418 (21%)	160 (19%)	258 (23%)	
36	Esophagitis					0.315
		No	1976 (99%)	850 (99%)	1126 (99%)	
		Yes	14 (0.7%)	8 (0.9%)	6 (0.5%)	
37	Family history of GC					<b>&lt; 0.001</b>
		No	1735 (87%)	712 (83%)	1023 (90%)	
		Yes	255 (13%)	146 (17%)	109 (9.6%)	
38	Family history of GI cancers					<b>&lt; 0.001</b>
		No	1475 (74%)	585 (68%)	890 (79%)	
		Yes	515 (26%)	273 (32%)	242 (21%)	
39	Family history of stomach operation					<b>0.005</b>
		No	1832 (92%)	773 (90%)	1059 (94%)	
		Yes	158 (7.9%)	85 (9.9%)	73 (6.4%)	
40	Fatty liver					0.226
		No	1970 (99%)	852 (99%)	1118 (99%)	
		Yes	20 (1.0%)	6 (0.7%)	14 (1.2%)	
41	Gastritis					<b>&lt; 0.001</b>
		No	1650 (83%)	758 (88%)	892 (79%)	
		Yes	340 (17%)	100 (12%)	240 (21%)	
42	Personal history of GI cancers					<b>&lt; 0.001</b>
		No	1926 (97%)	805 (94%)	1121 (99%)	
		Yes	64 (3.2%)	53 (6.2%)	11 (1.0%)	
43	Family history of stomach operation					<b>&lt; 0.001</b>
		No	1952 (98%)	830 (97%)	1122 (99%)	
		Yes	38 (1.9%)	28 (3.3%)	10 (0.9%)	
<b>4. Narcotics</b>						
44	Alcohol					0.415
		Never	1837 (92%)	787 (92%)	1050 (93%)	
		Ever	153 (7.7%)	71 (8.3%)	82 (7.2%)	
45	Opium					<b>&lt; 0.001</b>
		Never	1777 (89%)	732 (85%)	1045 (92%)	
		Ever	213 (11%)	126 (15%)	87 (7.7%)	
46	Smoking					<b>&lt; 0.001</b>
		Never	1414 (71%)	519 (60%)	895 (79%)	
		Ever	576 (29%)	339 (40%)	237 (21%)	

**Table 1** (continued)

	Predictors	levels	Overall, N = 1990 n (%)	GC, N = 858 n (%)	NUD, N = 1132 n (%)	P value <sup>1</sup>
47	Waterpipe	Never	1807 (91%)	775 (90%)	1032 (91%)	0.523
		Ever	183 (9.2%)	83 (9.7%)	100 (8.8%)	
48	Passive smoking (childhood)	Never	1235 (62%)	495 (58%)	740 (65%)	<b>&lt; 0.001</b>
		Ever	755 (38%)	363 (42%)	392 (35%)	
5. Socioeconomic status (SES)						
49	Birth place	Urban	990 (50%)	332 (39%)	658 (58%)	<b>&lt; 0.001</b>
		Rural	1000 (50%)	526 (61%)	474 (42%)	
50	Chemical exposure	Never	1704 (86%)	712 (83%)	992 (88%)	<b>0.003</b>
		Ever	286 (14%)	146 (17%)	140 (12%)	
51	Crowdedness	≤ 2 per room	1138 (57%)	458 (53%)	680 (60%)	<b>0.003</b>
		> 2 per room	852 (43%)	400 (47%)	452 (40%)	
52	Crowdedness (childhood)	≤ 2 per room	323 (16%)	115 (13%)	208 (18%)	<b>0.003</b>
		> 2 per room	1667 (84%)	743 (87%)	924 (82%)	
53	Drinking water	City plumbing	1777 (89%)	715 (83%)	1062 (94%)	<b>&lt; 0.001</b>
		Other than city plumbing	213 (11%)	143 (17%)	70 (6.2%)	
54	Education	> 8 yrs	551 (28%)	128 (15%)	423 (37%)	<b>&lt; 0.001</b>
		< 8 yrs	1439 (72%)	730 (85%)	709 (63%)	
55	Job-related physical activities	High	988 (50%)	500 (58%)	488 (43%)	<b>&lt; 0.001</b>
		Low	1002 (50%)	358 (42%)	644 (57%)	
56	Marital status	Other	115 (5.8%)	12 (1.4%)	103 (9.1%)	<b>&lt; 0.001</b>
		Married	1875 (94%)	846 (99%)	1029 (91%)	
57	Parents illiteracy (both)	No	636 (32%)	151 (18%)	485 (43%)	<b>&lt; 0.001</b>
		Yes	1354 (68%)	707 (82%)	647 (57%)	
58	Physical exercise per week	Never	701 (35%)	256 (30%)	445 (39%)	<b>&lt; 0.001</b>
		Ever	1289 (65%)	602 (70%)	687 (61%)	
59	Refrigerator use	Yes	1956 (98%)	840 (98%)	1116 (99%)	0.241
		No	34 (1.7%)	18 (2.1%)	16 (1.4%)	
60	Refrigerator use (childhood)	Yes	698 (35%)	200 (23%)	498 (44%)	<b>&lt; 0.001</b>
		No	1292 (65%)	658 (77%)	634 (56%)	
61	Residence place	Urban	1617 (81%)	604 (70%)	1013 (89%)	<b>&lt; 0.001</b>
		Rural	373 (19%)	254 (30%)	119 (11%)	
62	Residence place (childhood)	Urban	1073 (54%)	345 (40%)	728 (64%)	<b>&lt; 0.001</b>

**Table 1** (continued)

	Predictors	levels	Overall, N = 1990 n (%)	GC, N = 858 n (%)	NUD, N = 1132 n (%)	P value <sup>1</sup>
63	Residence type	Rural	917 (46%)	513 (60%)	404 (36%)	< 0.001
		Owned	1617 (81%)	758 (88%)	859 (76%)	
		Rented	373 (19%)	100 (12%)	273 (24%)	
64	Residence type (childhood)	Owned	1786 (90%)	807 (94%)	979 (86%)	< 0.001
		Rented	204 (10%)	51 (5.9%)	153 (14%)	

<sup>1</sup> Pearson's Chi-squared test

when adjusted for all other variables, few associations, remained statistically significant (Table 2).

The data obtained from the 64 predictors from our 1990 (GC + NUD) cases included varying degrees of missingness. To remedy this, we used the MICE method to impute the missing values. But first it was critical to validate our imputation technique and ascertain a consistent distribution for each predictor thereafter. Having done so, in the first approach amongst the 50 regenerated samples, the mean (S-Fig. 1) and standard deviation (S-Figure-2) differences, between our actual and imputed data did not exceed 0.05 and were thus acceptable. In the second approach, for all 1000 bootstrap-generated samples, the bias was determined as under 0.02 (S-Fig. 3).

### Model development

Logistic regression specified the strength of association between each of our 64 predictors and the clinical outcome (GC or NUD). The strengths of association for each of the predictors (if any), while adjusting for all other predictors, are presented via risk scores, shared percents, and odds ratios (Table 2). Taking into consideration all of the 64 predictors in our model, a risk calculator was created, scoring for GC or NUD (Fig. 1). The obtained risk score ranged from -1261 to +2077 (total range of 3338), moving from NUD towards GC. Of this range, the risk score of -451, equivalent to a shared percentage of 13.49, was assigned to subjects at reference level. The remaining 86.51 percentage of the risk score was contributed by our 64 predictors, with varying shares. Aiming for a minimum sensitivity rate of 90%, the risk score of -91, coinciding with the probability value of 0.29 (ranging from 0 to 1.0, Fig. 2) was identified as the cut-off point. Keeping in mind that each of the addressed 64 predictors contributed to the final risk score, those which were statistically significant are described below.

### GC-prone predictors

The predictors which acted towards the development of GC are considered as GC-prone. Amongst the demographic category, older age holds the first place, creating risk scores of +134 to +221 to +241, for subjects aged > 50 - 60, > 60 - 70 and > 70, in reference to those aged ≤ 50 years, respectively. These values were sequentially equivalent to 4.02, 6.60, and 7.23 shared percentage (SP) of the total range. Next in line, was being of male gender and non-Fars/mixed ethnicity, with risk scores of +119 (SP = 3.55%) and +89 (SP = 2.66%), respectively. Amongst the SES factors, the illiteracy of both parents and residence in a rural area contributed +77 and +78 risk scores, respectively, which contributed 2.35 and 2.3 shared percentages to the score. In regards to the medical status of the subjects, having a personal and family history of GI cancers provided a GC risk score of +173 (SP = 5.19) and +57 (SP = 1.72), respectively. Modifiable lifestyle behaviors, such as diet and use of narcotics took the subsequent positions. Amongst dietary habits, drinking hot tea [+53 (SP = 1.58)], consumption of medium-to-high amounts of cheese [+47 (SP = 1.42)], use of table salt [+46 (SP = 1.39)], late dinnertime [+34 (SP = 1.03)], and consumption of medium-to-high amounts of eggs [+32 (SP = 0.96)] were amongst the dietary GC-prone predictors (Table 2).

### Association with the subtype and subsite of GC

Some of the above-mentioned GC-prone predictors were also associated with the subsite and/or histologic subtype of the tumor. Amongst these, age was closely associated with the intestinal histologic subtype of GC ( $P = 0.003$ ). History of GI cancer was associated with the cardia anatomic location ( $P = 0.02$ ) and intestinal histologic subtype ( $P = 0.032$ ) of the tumor. The predictors of drinking hot tea ( $P = 0.004$ ) and consumption of table salt ( $P = 0.04$ ) were associated with the cardia subset of the GC tumors.

**Table 2** The results of the multivariate logistic regression model to explore the GC-prone versus NUD-prone predictors

Predictors		Shared Percentages	OR (95%CI)	P-Value	Risk Score
<b>1. Demographic</b>					
1	Age [> 50 – 60]	4.02	3.84 (2.56, 5.74)	<0.001	+134
	Age [> 60 – 70]	6.6	9.09 (5.7, 14.49)	<0.001	+221
	Age [> 70]	7.23	11.18 (6.36, 19.66)	<0.001	+241
2	Ethnicity [Non-Fars]	2.66	2.43 (1.74, 3.41)	<0.001	+89
3	Gender [Male]	3.55	3.27 (2.03, 5.26)	<0.001	+119
<b>2. Diet</b>					
4	Canned food [Ever]	0.01	1 (0.72, 1.39)	0.991	0
5	Carbonated soft drinks [Medium to high]	1	0.72 (0.49, 1.06)	0.095	-33
6	Cheese [Medium to high]	1.42	1.61 (1.15, 2.25)	<b>0.006</b>	+47
7	Chicken [Medium to high]	0.29	0.91 (0.67, 1.24)	0.548	-10
8	Coffee [Medium to high]	1.95	0.52 (0.18, 1.52)	0.231	-65
9	Cooking method [Other than boiling]	0.79	1.3 (0.95, 1.79)	0.102	+26
10	Cooking oil [Saturated or both]	0.61	1.23 (0.87, 1.72)	0.238	+20
11	Cooking salt [Medium to high]	0.61	1.23 (0.86, 1.75)	0.262	+20
12	Dinner time [Late]	1.03	1.41 (1.03, 1.93)	<b>0.032</b>	+34
13	Drinking water (childhood) [Other than city plumbing]	0.89	1.35 (0.89, 2.04)	0.161	+30
14	Eggs [Medium to high]	0.96	1.38 (1.01, 1.89)	<b>0.044</b>	+32
15	Fish [Medium to high]	0.19	1.07 (0.75, 1.51)	0.716	+6
16	Fruits [Medium to high]	0.3	0.9 (0.64, 1.28)	0.569	-10
17	Milk [Medium to high]	0.57	1.21 (0.87, 1.68)	0.259	+19
18	Minerals [Ever]	1.24	0.66 (0.43, 1.01)	0.055	-42
19	Pickled vegetables [Medium to high]	0.21	1.07 (0.74, 1.56)	0.717	+7
20	Processed meats [Medium to high]	0.22	0.93 (0.67, 1.29)	0.656	-7
21	Potato chips [Ever]	0.11	0.96 (0.68, 1.36)	0.83	-4
22	Red meat [Medium to high]	0.14	0.95 (0.7, 1.3)	0.767	-5
23	Salted food [Ever]	0.42	0.87 (0.47, 1.59)	0.646	-14
24	Smoked fish [Ever]	1.31	1.55 (0.84, 2.86)	0.161	+44
25	Smoked rice [Ever]	1.72	0.56 (0.33, 0.98)	<b>0.041</b>	-57
26	Table salt [Ever]	1.39	1.59 (1.14, 2.22)	<b>0.007</b>	+46
27	Tea [Medium to high]	0.3	1.1 (0.45, 2.7)	0.827	+10
28	Tea temperature [Hot]	1.58	1.7 (1.24, 2.33)	<b>0.001</b>	+53
29	Tuna fish [Ever]	0.94	0.73 (0.54, 0.99)	<b>0.044</b>	-31
30	Vegetables [Medium to high]	0.76	1.29 (0.9, 1.84)	0.164	+25
31	Vitamins [Ever]	1.81	0.55 (0.35, 0.85)	<b>0.008</b>	-60
32	Yogurt [Medium to high]	1.16	0.68 (0.47, 0.98)	<b>0.036</b>	-39
<b>3. Medical status (self-reported)</b>					
33	Colitis [Yes]	3.43	3.15 (0.66, 14.94)	0.148	+115
34	Diabetes [Yes]	0.24	0.92 (0.54, 1.56)	0.762	-8
35	Esophageal reflux [Yes]	0.36	0.89 (0.6, 1.3)	0.532	-12
36	Esophagitis [Yes]	0.29	1.1 (0.2, 6.14)	0.913	+10
37	Family history of GC [Yes]	0.64	1.24 (0.65, 2.36)	0.514	+21
38	Family history of GI cancers [Yes]	1.72	1.77 (1.14, 2.76)	<b>0.011</b>	+57
39	Family history of stomach operation [Yes]	0.6	1.22 (0.64, 2.35)	0.547	+20
40	Fatty liver [Yes]	2.59	0.42 (0.08, 2.3)	0.319	-87
41	Gastritis [Yes]	0.92	0.73 (0.48, 1.12)	0.151	-31
42	Personal history of GI cancers [Yes]	5.19	5.67 (1.87, 17.17)	<b>0.002</b>	+173
43	Personal history of stomach operation [Yes]	2.25	2.12 (0.64, 7.04)	0.218	+75
<b>4. Narcotics</b>					

**Table 2** (continued)

Predictors		Shared Percentages	OR (95%CI)	P-Value	Risk Score
44	Alcohol [Ever]	0.12	1.04 (0.57, 1.9)	0.895	+4
45	Opium [Ever]	0.51	0.84 (0.5, 1.42)	0.521	-17
46	Smoking [Ever]	0.99	1.39 (0.96, 2.03)	0.084	+33
47	Waterpipe [Ever]	0.66	1.25 (0.72, 2.16)	0.433	+22
48	Passive smoking (childhood) [Ever]	0.95	1.37 (1, 1.89)	0.053	+32
<b>5. Socioeconomic status (SES)</b>					
49	Birth place [Rural]	1.69	0.57 (0.36, 0.9)	<b>0.016</b>	-57
50	Chemical exposure [Ever]	0.08	0.97 (0.64, 1.49)	0.903	-3
51	Crowdedness > 2 per room]	0.23	0.93 (0.68, 1.27)	0.634	-8
52	Crowdedness (childhood) [> 2 per room]	1.14	0.68 (0.44, 1.07)	0.096	-38
53	Drinking water [Other than city plumbing]	0.98	1.39 (0.8, 2.4)	0.243	+33
54	Education [< 8yrs]	0.42	0.87 (0.57, 1.32)	0.51	-14
55	Job-related physical activities [Low]	0.56	1.2 (0.79, 1.85)	0.395	+19
56	Marital status [Married]	2.28	2.15 (0.95, 4.83)	0.065	+76
57	Parents illiteracy (both) [Yes]	2.35	2.19 (1.5, 3.2)	<b>&lt; 0.001</b>	+78
58	Physical exercise per week [Ever]	0.55	0.83 (0.6, 1.16)	0.274	-18
59	Refrigerator use [No]	0.79	0.77 (0.24, 2.45)	0.655	-26
60	Refrigerator use (childhood) [No]	1.39	0.63 (0.42, 0.95)	<b>0.029</b>	-46
61	Residence place [Rural]	2.3	2.16 (1.36, 3.42)	<b>0.001</b>	+77
62	Residence place (childhood) [Rural]	0.26	1.09 (0.67, 1.79)	0.729	+9
63	Residence type [Rented]	0.19	0.94 (0.63, 1.4)	0.756	-6
64	Residence type (childhood) [Rented]	1.85	0.54 (0.32, 0.9)	<b>0.019</b>	-62
65	Intercept (other unknown factors)	13.49	0.01 (0, 0.04)	<b>&lt; 0.001</b>	-451

The indicated levels in the parenthesis are assessed against the reference level for each predictor

### NUD-prone predictors

However, there were some predictors that acted towards the development of NUD; in other words, they were NUD-prone. These predictors belonged to the two categories of SES and diet. The predictors of the former category included: rented residential place (during childhood) [-62 (SP=1.85)], rural-birth place [-57 (SP=1.69)], and lacking refrigerator (during childhood) [-46 (SP=1.39)]. Predictors of the dietary habits included: taking vitamins [-60 (SP=1.81)], consumption of smoked rice [-57 (SP=1.72)], medium-to-high consumption of yoghurt [-39 (SP=1.16)], and consumption of tuna fish [-31 (SP=0.94)] (Table 2).

### Model validation

To validate the results of the above-described development model, a validation approach was taken, assessing 597 individuals (248 GC and 349 NUD; GC: NUD ratio, 1:1.41), on which the ROC analysis was performed (Fig. 3). Using our 64 predictors, we were able to differentiate GC from NUD, with an AUC of 86.37% and the sensitivity, specificity, and accuracy rates of 85.89, 63.9 and 73.03%, respectively. According to this model, the rates of

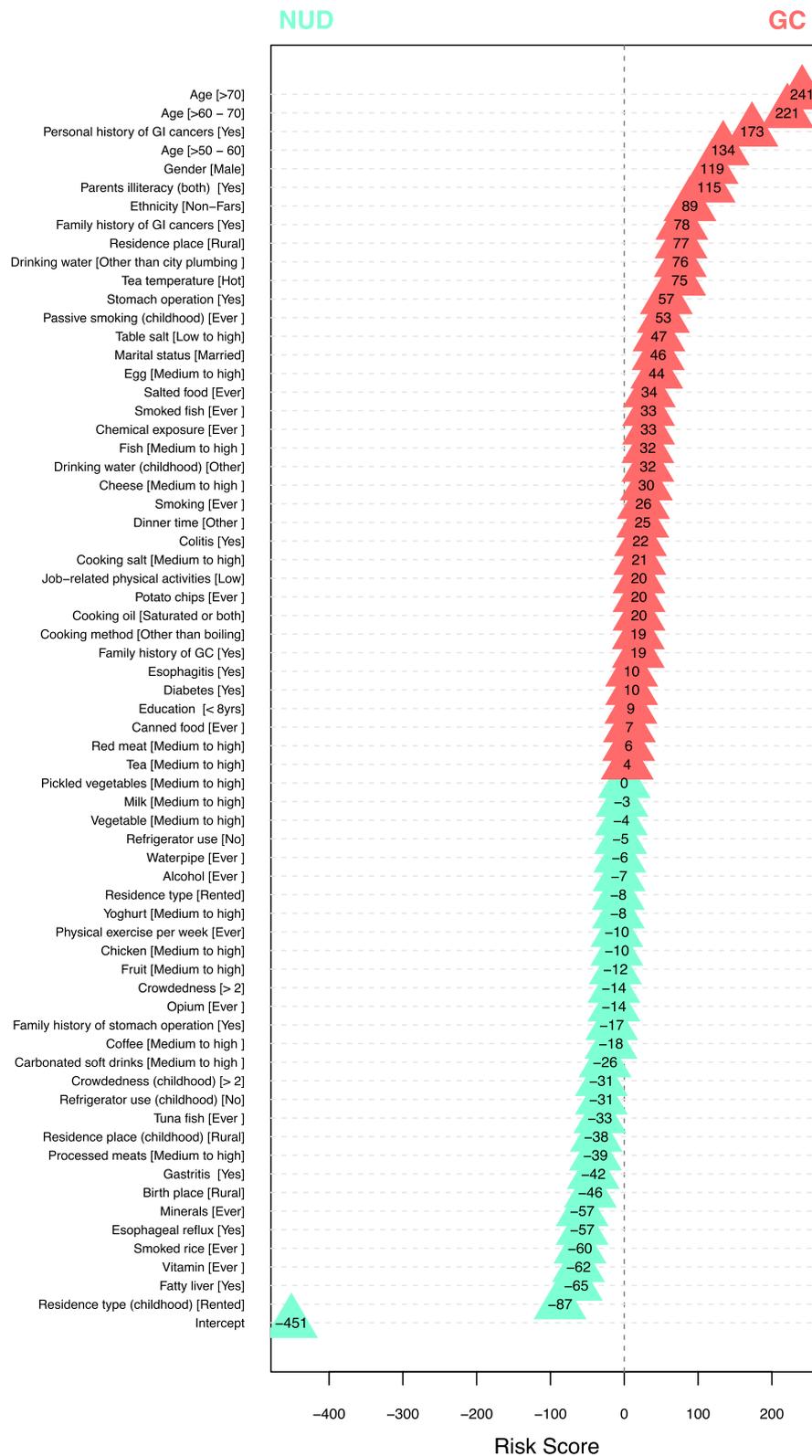
false positives and false negatives were 36.1, and 14.11%, respectively (Fig. 3).

We have also evaluated the calibration of our model by the Hosmer-Lemeshow method. Having done so, a *P* value of 0.4761 (well above 0.05) was obtained. Thus, the fitness of our model was confirmed. Our risk calculator can, thus, calculate the risk of GC *versus* NUD, based on the probability our 64 predictors, proportional to the achieved risk score (Fig. 3).

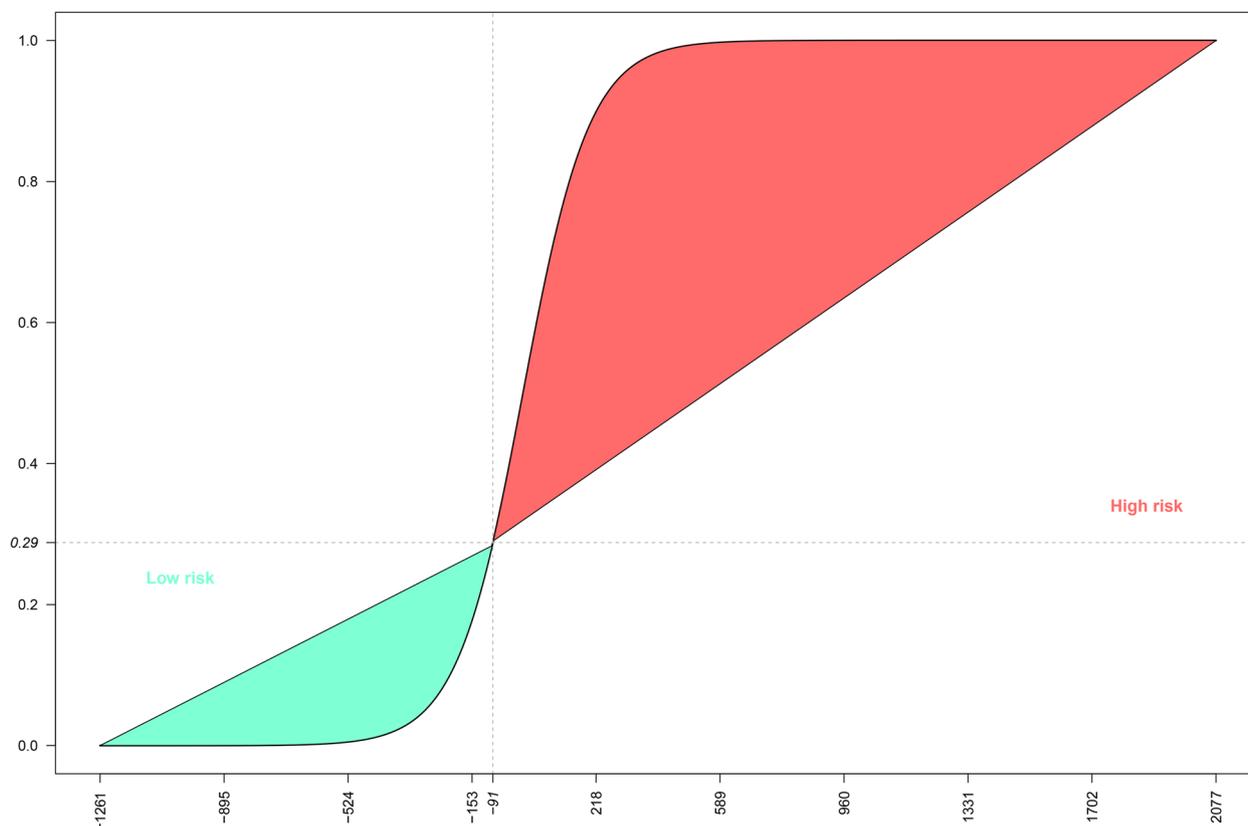
### Discussion

Gastric cancer being a silent killer, usually catches patients and their health service providers, off-guard. Being able to assign a relative risk to subjects, based on their demographic characteristics and life style behaviours, will provide an upper hand in focusing on the at-risk subjects, with subsequent stepwise clinical testing and follow-ups. The goal of this study was to develop an approach to accomplish the primary screening step based on our target predictors.

In 2023, a multicentre population-based study, carried out on over 416 thousand subjects (aged 40 – 75 years) in China, a GC risk calculator was developed, which



**Fig. 1** Risk score system for segregation of GC from NUD based on our logistic regression model



**Fig. 2** The probability of GC versus NUD based on risk scores

highlighted 11 demographic and life style variables that place individuals at risk of GC [34]. Although our study was hospital-based and has screened Iranian dyspeptic patients, the common variables between these two studies still identify age, gender (male), education (illiteracy of parents), salt intake and personal and family history of cancer as definite risk factors. In another population-based screening study on subjects (aged 40–74 years), with no history of cancer in Korea, six risk factors were identified [35], of which salt intake was a shared prominent risk factor with our hospital-based screening study.

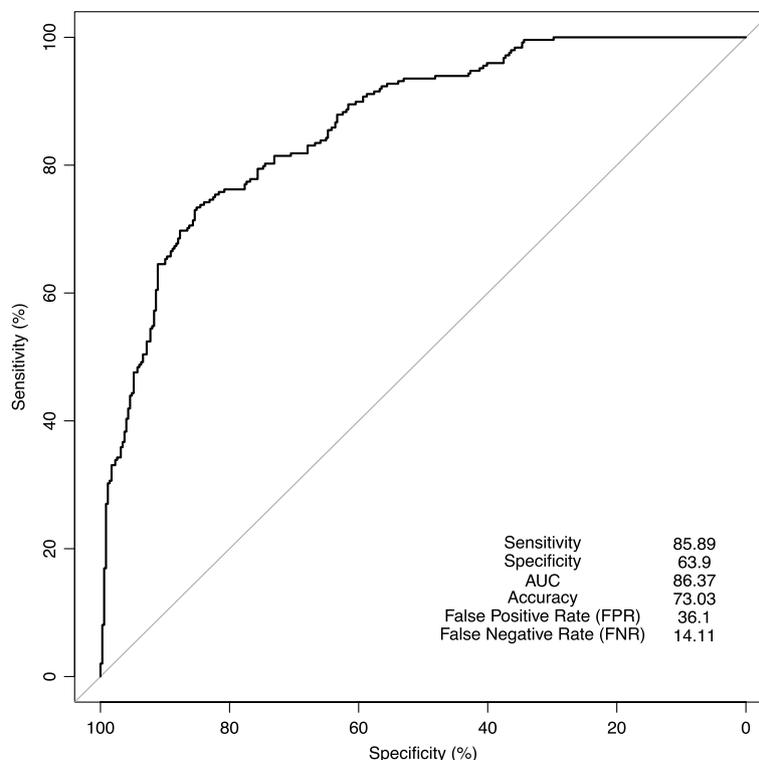
In 2019, a population-based study was conducted in China to assess the general knowledge about GC risk factors and symptoms. The analysis was performed on 1200 adults, over the age of 18 with an average age of 40, which showed that the mean score for GC knowledge was 8.85 out of 22. Of the 1200 participants, 564 (47.0%) had insufficient understanding of GC risk factors and warning symptoms. Overall, about 84% of people believed that screening helped diagnose GC. However, only 15.2% of people were screened for GC. There were various reasons for avoiding screening, including being asymptomatic, fear of diagnostic screening and its outcomes, male gender, living in rural areas, lower educational levels, etc.

[36]. Hence, lack of routine screening and the absence of specific symptoms for this fatal disease, leaves most subjects undiagnosed until the terminal stages, which accounts for GC being known as a silent killer [37, 38].

Several methodological studies on GC have been conducted over the years [13, 39–43]. A concerted strategy for the joint analysis of these investigations may allow new insights into the etiology of GC. Therefore, the ‘Stomach cancer Pooling (StoP) Project’ was set up in 2012 to join together several investigators and create a consortium of epidemiological investigations on risk factors for GC. The SToP’s final aim was to examine the role of several lifestyles and genetic determinants in the etiology of GC, through pooled analyses of individual-level data [44].

In our study we intended to investigate the effects of any potential risk factors, even if they were not statistically significant, so to create an all-inclusive risk calculator.

The GC-prone factors identified *via* our model, are also supported by previous studies, include older age [34, 45–47], male gender [34, 48], and non-Fars/mixed ethnicity [49–51], illiteracy of both parents [52–55], family history of GI cancers [56–59], drinking hot tea [60], late



**Fig. 3** ROC curve analysis of our model differentiating GC from NUD

dinnertime [61, 62], consumption of table salt [63–65], and medium to high amounts of cheese and eggs [66–69]. Having used a logistic regression model, we have developed a gastric cancer risk calculator, with the sensitivity, specificity, and accuracy rates of 85.89, 63.9, and 73.03%, which can be used by individuals or their healthcare workers, for primary screening of dyspeptic patients.

In 2007, Driver et al. [21] developed a simple scoring system that identifies men at increased risk of colorectal cancer, based on age and modifiable behaviours, such as alcohol intake, smoking status, and body mass index. They ran a logistic regression model as well as a proportional hazards model, to better simulate a screening decision, based on the information obtained. The discrimination power of the final model was about 70% (AUC = 69.5%) [21]. In comparison, our risk score had the discrimination power (AUC) of 86.37% during internal validation. Keeping in mind that this is the primary screening step, followed by simple and complex clinical testing, the limited detection rates, we have herein obtained for a primary questionnaire-based surveillance, are acceptable.

The strengths of our study include its sensible sample size and inclusion of a wide variety of target demographic and lifestyle behaviours. However, we have used a case-case setting in order to be able to add other clinical data,

on the next rounds of clinical and paraclinical screening. Having compared GC patients with non-GC (non-ulcer dyspeptic, NUD) patients, the scale bar of our risk score moves towards the direction of GC (GC-prone) or NUD (NUD-prone) and is, at best, suitable for screening dyspeptic patients, rather than the general population. Thus, our risk calculator, must be adjusted, by applying the model in a case-control (GC *versus* healthy population) setting. It must also be kept in mind that some of the highlighted risk indicators may actually be proxies for other unaddressed predictors. Furthermore, the fact that we had to turn our multinomial levels (answers), into binomial, may have oversimplified our model. Another point of concern is the external validation of this model on other sample cohorts with diverse environmental, cultural, and social characteristics.

Nevertheless, applying such an inexpensive GC risk calculator, using questionnaire-based information, can provide the first step in screening Iranian at risk patients, to be followed by more complex laboratory and clinical screenings. Furthermore, providing information about individualized GC risk status, can lead to attempts at correction of the modifiable risk behaviours. Future studies include, validation of this model in case-control settings, in different geographic locations.

**Abbreviations**

GC	Gastric cancer
NUD	Non-ulcer dyspeptic
ROC	Receiver operating characteristic
MOHME	Ministry of Health and Medical Education
StoP	Stomach cancer Pooling
SP	Shared percentage
FCS	Full conditional specification
JM	Joint modeling
MICE	Multivariate imputation by chained equations
NCII	National cancer institute of Iran
IARC	International Agency for Research on Cancer
RS	Risk scores

**Supplementary Information**

The online version contains supplementary material available at <https://doi.org/10.1186/s12876-024-03123-z>.

**Additional file 1.**

**Additional file 2.**

**Additional file 3.**

**Additional file 4.**

**Acknowledgments**

The author would like to sincerely thank Professor Emeritis Olof Nyren, (Karolinska Institutet, Stockholm, Sweden) for his expert and generous guidance in formulating the interview questionnaire. We would also like to thank every interviewer who have generously assisted in obtaining the data on our patient population over the years.

**Authors' contributions**

KG: Modelling, methodology, computational analysis, drafting the manuscript, and assistance in conceiving the concept and study design; SS: Data collection, assistance in conceiving the concept, study design and drafting the manuscript; ME: Data collection, database entry and management; MT: Data collection, MEH: Clinical data collection on endoscopy patients; AN: Data collection; AK: Methodology; MAM: Clinical data collection on gastric cancer patients; MM: Conceptualization, study design, supervision, drafting, revising and finalizing the manuscript.

**Funding**

This study was supported by a technical assistance grant (IRN-072), which was co-funded by the Islamic Development Bank, Saudi Arabia and Pasteur Institute of Iran.

**Availability of data and materials**

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

**Declarations****Ethics approval and consent to participate**

All methods were carried out following relevant guidelines and regulations. Data collections were carried out following obtaining the participants' written informed consent, according to the protocols approved by the National Committee on Ethical Issues in Medical Research, Ministry of Health and Medical Education (MOHME) of Iran; Ref No. 315.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare no competing interests.

**Author details**

<sup>1</sup>HPGC Research Group, Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran. <sup>2</sup>Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. <sup>3</sup>Gastroenterology Department, Amiralam Hospital, Tehran University of Medical Sciences, Tehran, Iran. <sup>4</sup>Cancer Research Center, Cancer Institute, Tehran University of Medical Sciences, Tehran, Iran.

Received: 22 November 2023 Accepted: 2 January 2024

Published online: 18 January 2024

**References**

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–49.
- Alvarez EM, Force LM, Xu R, Compton K, Lu D, Henrikson HJ, Kocarnik JM, Harvey JD, Pennini A, Dean FE. The global burden of adolescent and young adult cancer in 2019: a systematic analysis for the global burden of disease study 2019. *Lancet Oncol*. 2022;23(1):27–52.
- Thrift AP, Nguyen TH. Gastric Cancer epidemiology. *Gastrointest Endosc Clin N Am*. 2021;31(3):425–39.
- Bertuccio P, Chatenoud L, Levi F, Praud D, Ferlay J, Negri E, Malvezzi M, La Vecchia C. Recent patterns in gastric cancer: a global overview. *Int J Cancer*. 2009;125(3):666–73.
- Bertuccio P, Rosato V, Andreano A, Ferraroni M, Decarli A, Edefonti V, La Vecchia C. Dietary patterns and gastric cancer risk: a systematic review and meta-analysis. *Ann Oncol*. 2013;24(6):1450–8.
- Du S, Li Y, Su Z, Shi X, Johnson NL, Li P, Zhang Y, Zhang Q, Wen L, Li K, et al. Index-based dietary patterns in relation to gastric cancer risk: a systematic review and meta-analysis. *Br J Nutr*. 2020;123(9):964–74.
- de Menezes RF, Bergmann A, Thuler LC. Alcohol consumption and risk of cancer: a systematic literature review. *Asian Pac J Cancer Prev*. 2013;14(9):4965–72.
- Sasco AJ, Secretan MB, Straif K. Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer*. 2004;45(Suppl 2):S3–9.
- Ladeiras-Lopes R, Pereira AK, Nogueira A, Pinheiro-Torres T, Pinto I, Santos-Pereira R, Lunet N. Smoking and gastric cancer: systematic review and meta-analysis of cohort studies. *Cancer Causes Control*. 2008;19(7):689–701.
- Dong J, Thrift AP. Alcohol, smoking and risk of oesophago-gastric cancer. *Best Pract Res Clin Gastroenterol*. 2017;31(5):509–17.
- Cavatorta O, Scida S, Miraglia C, Barchi A, Nouvenne A, Leandro G, Meschi T, De' Angelis GL, Di Mario F. Epidemiology of gastric cancer and risk factors. *Acta Biomed*. 2018;89(8-s):82–7.
- Machlowska J, Baj J, Sitarz M, Maciejewski R, Sitarz R. Gastric Cancer: epidemiology, risk factors, classification, genomic characteristics and treatment strategies. *Int J Mol Sci*. 2020;21(11):4012.
- Guggenheim DE, Shah MA. Gastric cancer epidemiology and risk factors. *J Surg Oncol*. 2013;107(3):230–6.
- Yusefi AR, Bagheri Lankarani K, Bastani P, Radinmanesh M, Kavosi Z. Risk factors for gastric Cancer: a systematic review. *Asian Pac J Cancer Prev*. 2018;19(3):591–603.
- Lin Y, Zheng Y, Wang HL, Wu J. Global patterns and trends in gastric Cancer incidence rates (1988–2012) and predictions to 2030. *Gastroenterology*. 2021;161(1):116–127.e118.
- National Research Council Committee on Risk P. Communication. In: *Improving Risk Communication*. Washington (DC): National Academies Press (US) Copyright © 1989 by the National Academy of Sciences; 1989.
- In: *Fundamentals of Clinical Data Science*. edn. Edited by Kubben P, Dumontier M, Dekker A. Cham (CH): Springer Copyright 2019, the Editor(s) (if applicable) and The Author(s). This book is an open access publication. 2019. ISBN 978-3-319-99712-4, ISBN 978-3-319-99713-1 (eBook).
- Mishra GA, Dhivar HD, Gupta SD, Kulkarni SV, Shastri SS. A population-based screening program for early detection of common cancers

- among women in India - methodology and interim results. *Indian J Cancer*. 2015;52(1):139–45.
19. Li X, Bian D, Yu J, Li M, Zhao D. Using machine learning models to improve stroke risk level classification methods of China national stroke screening. *BMC Med Inform Decis Mak*. 2019;19(1):261.
  20. Heikes KE, Eddy DM, Arondekar B, Schlessinger L. Diabetes risk calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes. *Diabetes Care*. 2008;31(5):1040–5.
  21. Driver JA, Gaziano JM, Gelber RP, Lee IM, Buring JE, Kurth T. Development of a risk score for colorectal cancer in men. *Am J Med*. 2007;120(3):257–63.
  22. Lloyd-Jones DM, Wilson PW, Larson MG, Beiser A, Leip EP, D'Agostino RB, Levy D. Framingham risk score and prediction of lifetime risk for coronary heart disease. *Am J Cardiol*. 2004;94(1):20–4.
  23. Wittekind C, Compton CC, Greene FL, Sobin LH. TNM residual tumor classification revisited. *Cancer*. 2002;94(9):2511–6.
  24. Lauren P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. *Acta Pathol Microbiol Scand*. 1965;64:31–49.
  25. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann Transl Med*. 2016;4(2):30.
  26. Yucler RM. Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philos Trans A Math Phys Eng Sci*. 1874;2008(366):2389–403.
  27. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16(3):219–42.
  28. Wahl S, Boulesteix AL, Zierer A, Thorand B, van de Wiel MA. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. *BMC Med Res Methodol*. 2016;16(1):144.
  29. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. *Logistic regression*. Springer; 2002.
  30. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*. 2013;4(2):627–35.
  31. Refaailzadeh P, Tang L, Liu H, Liu L. *Encyclopedia of database systems*. In: Cross-validation. Springer; 2009. p. 532–8.
  32. Blaha MJ. The critical importance of risk score calibration: time for transformative approach to risk score validation? *J Am Coll Cardiol*. 2016;67(18):2131–4.
  33. Landwehr JM, Pregibon D, Shoemaker AC. Graphical methods for assessing logistic regression models. *J Am Stat Assoc*. 1984;79(385):61–71.
  34. Zhu X, Lv J, Zhu M, Yan C, Deng B, Yu C, Guo Y, Ni J, She Q, Wang T, et al. Development, validation, and evaluation of a risk assessment tool for personalized screening of gastric cancer in Chinese populations. *BMC Med*. 2023;21(1):159.
  35. Trinh TTK, Lee K, Oh JK, Suh M, Jun JK, Choi KS. Cluster of lifestyle risk factors for stomach cancer and screening behaviors among Korean adults. *Sci Rep*. 2023;13(1):17503.
  36. Liu Q, Zeng X, Wang W, Huang RL, Huang YJ, Liu S, Huang YH, Wang YX, Fang QH, He G, et al. Awareness of risk factors and warning symptoms and attitude towards gastric cancer screening among the general public in China: a cross-sectional study. *BMJ Open*. 2019;9(7):e029638.
  37. Necula L, Matei L, Dragu D, Neagu AI, Mambet C, Nedeianu S, Bleotu C, Diaconu CC, Chivu-Economescu M. Recent advances in gastric cancer early diagnosis. *World J Gastroenterol*. 2019;25(17):2029–44.
  38. Pasechnikov V, Chukov S, Fedorov E, Kikuste I, Leja M. Gastric cancer: prevention, screening and early diagnosis. *World J Gastroenterol*. 2014;20(38):13842–62.
  39. Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. *Cancer Epidemiol Biomark Prev*. 2014;23(5):700–13.
  40. Correa P. Gastric cancer: overview. *Gastroenterol Clin N Am*. 2013;42(2):211–7.
  41. Ang TL, Fock KM. Clinical epidemiology of gastric cancer. *Singap Med J*. 2014;55(12):621–8.
  42. Smyth EC, Nilsson M, Grabsch HJ, van Grieken NC, Lordick F. Gastric cancer. *Lancet*. 2020;396(10251):635–48.
  43. Venerito M, Link A, Rokkas T, Malfertheiner P. Gastric cancer - clinical and epidemiological aspects. *Helicobacter*. 2016;21(Suppl 1):39–44.
  44. Pelucchi C, Lunet N, Boccia S, Zhang ZF, Praud D, Boffetta P, Levi F, Matsuo K, Ito H, Hu J, et al. The stomach cancer pooling (StoP) project: study design and presentation. *Eur J Cancer Prev*. 2015;24(1):16–23.
  45. Choi Y, Kim N, Kim KW, Jo HH, Park J, Yoon H, Shin CM, Park YS, Lee DH. Gastric Cancer in older patients: a retrospective study and literature review. *Ann Geriatr Med Res*. 2022;26(1):33–41.
  46. Lee JG, Kim SA, Eun CS, Han DS, Kim YS, Choi BY, Song KS, Kim HJ, Park CH. Impact of age on stage-specific mortality in patients with gastric cancer: a long-term prospective cohort study. *PLoS One*. 2019;14(8):e0220660.
  47. Chung HW, Noh SH, Lim JB. Analysis of demographic characteristics in 3242 young age gastric cancer patients in Korea. *World J Gastroenterol*. 2010;16(2):256–63.
  48. Choi Y, Kim N, Kim KW, Jo HH, Park J, Yoon H, Shin CM, Park YS, Lee DH, Oh HJ, et al. Sex-based differences in histology, staging, and prognosis among 2983 gastric cancer surgery patients. *World J Gastroenterol*. 2022;28(9):933–47.
  49. Mohebbi M, Mahmoodi M, Wolfe R, Nourijelani K, Mohammad K, Zeraati H, Fotouhi A. Geographical spread of gastrointestinal tract cancer incidence in the Caspian Sea region of Iran: spatial analysis of cancer registry data. *BMC Cancer*. 2008;8:137.
  50. Malekzadeh R, Derakhshan MH, Malekzadeh Z. Gastric cancer in Iran: epidemiology and risk factors. *Arch Iran Med*. 2009;12(6):576–83.
  51. Kalan Farmanfarma K, Mahdavi N, Hassanipour S, Salehinyia H. Epidemiologic study of gastric Cancer in Iran: a systematic review. *Clin Exp Gastroenterol*. 2020;13:511–42.
  52. Tam YH, Yeung CK, Lee KH, Sihoe JD, Chan KW, Cheung ST, Mou JW. A population-based study of *Helicobacter pylori* infection in Chinese children resident in Hong Kong: prevalence and potential risk factors. *Helicobacter*. 2008;13(3):219–24.
  53. Nouraei M, Latifi-Navid S, Rezvan H, Radmard AR, Maghsudlu M, Zaer-Rezaei H, Amini S, Siavoshi F, Malekzadeh R. Childhood hygienic practice and family education status determine the prevalence of *Helicobacter pylori* infection in Iran. *Helicobacter*. 2009;14(1):40–6.
  54. Lagergren J, Andersson G, Talbäck M, Drefahl S, Bihagen E, Härkönen J, Feychting M, Ljung R. Marital status, education, and income in relation to the risk of esophageal and gastric cancer by histological type and site. *Cancer*. 2016;122(2):207–12.
  55. Stephens MR, Blackshaw GR, Lewis WG, Edwards P, Barry JD, Hopper NA, Allison MC. Influence of socio-economic deprivation on outcomes for patients diagnosed with gastric cancer. *Scand J Gastroenterol*. 2005;40(11):1351–7.
  56. Slavin TP, Weitzel JN, Neuhausen SL, Schrader KA, Oliveira C, Karam R. Genetics of gastric cancer: what do we know about the genetic risks? *Transl Gastroenterol Hepatol*. 2019;4:55.
  57. Ren JS, Freedman ND, Kamangar F, Dawsey SM, Hollenbeck AR, Schatzkin A, Abnet CC. Tea, coffee, carbonated soft drinks and upper gastrointestinal tract cancer risk in a large United States prospective cohort study. *Eur J Cancer*. 2010;46(10):1873–81.
  58. Sun WY, Yang H, Wang XK, Fan JH, Qiao YL, Taylor PR. The association between family history of upper gastrointestinal Cancer and the risk of death from upper gastrointestinal Cancer-based on Linxian dysplasia nutrition intervention trial (NIT) cohort. *Front Oncol*. 2022;12:897534.
  59. Thrift AP, Jove AG, Liu Y, Tan MC, El-Serag HB. Associations of duration, intensity, and quantity of smoking with risk of gastric intestinal metaplasia. *J Clin Gastroenterol*. 2022;56(1):e71–6.
  60. Mao XQ, Jia XF, Zhou G, Li L, Niu H, Li FL, Liu HY, Zheng R, Xu N. Green tea drinking habits and gastric cancer in Southwest China. *Asian Pac J Cancer Prev*. 2011;12(9):2179–82.
  61. Xu L, Zhang X, Lu J, Dai JX, Lin RQ, Tian FX, Liang B, Guo YN, Luo HY, Li N, et al. The effects of dinner-to-bed time and post-dinner walk on gastric Cancer across different age groups: a multicenter case-control study in Southeast China. *Medicine (Baltimore)*. 2016;95(16):e3397.
  62. Huang L, Chen L, Gui ZX, Liu S, Wei ZJ, Xu AM. Preventable lifestyle and eating habits associated with gastric adenocarcinoma: a case-control study. *J Cancer*. 2020;11(5):1231–9.
  63. Wu B, Yang D, Yang S, Zhang G. Dietary salt intake and gastric Cancer risk: a systematic review and Meta-analysis. *Front Nutr*. 2021;8:801228.
  64. Thapa S, Fischbach LA, Delongchamp R, Faramawi MF, Orloff M. Association between dietary salt intake and progression in the gastric precancerous process. *Cancers (Basel)*. 2019;11(4):467.

65. Ge S, Feng X, Shen L, Wei Z, Zhu Q, Sun J. Association between habitual dietary salt intake and risk of gastric Cancer: a systematic review of observational studies. *Gastroenterol Res Pract.* 2012;2012:808120.
66. Shen JG, Jin LD, Dong MJ, Wang LB, Zhao WH, Shen J. Low level of serum high-density lipoprotein cholesterol in gastric cancer correlates with cancer progression but not survival. *Transl Cancer Res.* 2020;9(10):6206–13.
67. Miao P, Guan L. Association of Dietary Cholesterol Intake with Risk of gastric Cancer: a systematic review and Meta-analysis of observational studies. *Front Nutr.* 2021;8:722450.
68. Shin HJ, Roh CK, Son SY, Hoon H, Han SU. Prognostic value of hypocholesterolemia in patients with gastric cancer. *Asian J Surg.* 2021;44(1):72–9.
69. Pih GY, Gong EJ, Choi JY, Kim MJ, Ahn JY, Choe J, Bae SE, Chang HS, Na HK, Lee JH, et al. Associations of serum lipid level with gastric Cancer risk, pathology, and prognosis. *Cancer Res Treat.* 2021;53(2):445–56.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.