Original Paper

Electronic Health Record–Based Absolute Risk Prediction Model for Esophageal Cancer in the Chinese Population: Model Development and External Validation

Yuting Han[1*], PhD; Xia Zhu[2,3*], PhD; Yizhen Hu[1], PhD; Canqing Yu[1,4], PhD; Yu Guo[5], MSc; Dong Hang[2,3], PhD; Yuanjie Pang[1], DPhil; Pei Pei[5], MSc; Hongxia Ma[2,3], PhD; Dianjianyi Sun[1,4], PhD; Ling Yang[6,7], PhD; Yiping Chen[6,7], DPhil; Huaidong Du[6,7], PhD; Min Yu[8], MSc; Junshi Chen[9], MD; Zhengming Chen[7], DPhil; Dezheng Huo[10], PhD; Guangfu Jin[2,3], PhD; Jun Lv[1,4], PhD; Zhibin Hu[2,3], PhD; Hongbing Shen[2,3], PhD; Liming Li[1,4], MPH

[1]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China

[2]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China

[3]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, China International Cooperation Center for Environment and Human Health, Nanjing Medical University, Nanjing, China

[4]Peking University Center for Public Health and Epidemic Preparedness and Response, Beijing, China

[5]Chinese Academy of Medical Sciences, Beijing, China

[6]Medical Research Council Population Health Research Unit, University of Oxford, Oxford, United Kingdom

[7]Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

[8]Zhejiang Center for Disease Control and Prevention, Hangzhou, China

[9]China National Center for Food Safety Risk Assessment, Beijing, China

[10]Department of Public Health Sciences, The University of Chicago, Chicago, IL, United States

[*]these authors contributed equally

Corresponding Author:
Jun Lv, PhD
Department of Epidemiology and Biostatistics
School of Public Health
Peking University
No 38 Xueyuan Rd, Haidian District
Beijing, 100191
China
Phone: 86 10 82801528
Fax: 86 10 82801528
Email: lvjun@bjmu.edu.cn

Abstract

Background: China has the largest burden of esophageal cancer (EC). Prediction models can be used to identify high-risk individuals for intensive lifestyle interventions and endoscopy screening. However, the current prediction models are limited by small sample size and a lack of external validation, and none of them can be embedded into the booming electronic health records (EHRs) in China.

Objective: This study aims to develop and validate absolute risk prediction models for EC in the Chinese population. In particular, we assessed whether models that contain only EHR-available predictors performed well.

Methods: A prospective cohort recruiting 510,145 participants free of cancer from both high EC-risk and low EC-risk areas in China was used to develop EC models. Another prospective cohort of 18,441 participants was used for validation. A flexible parametric model was used to develop a 10-year absolute risk model by considering the competing risks (full model). The full model was then abbreviated by keeping only EHR-available predictors. We internally and externally validated the models by using the area under the receiver operating characteristic curve (AUC) and calibration plots and compared them based on classification measures.

Results: During a median of 11.1 years of follow-up, we observed 2550 EC incident cases. The models consisted of age, sex, regional EC-risk level (high-risk areas: 2 study regions; low-risk areas: 8 regions), education, family history of cancer (simple

model), smoking, alcohol use, BMI (intermediate model), physical activity, hot tea consumption, and fresh fruit consumption (full model). The performance was only slightly compromised after the abbreviation. The simple and intermediate models showed good calibration and excellent discriminating ability with AUCs (95% CIs) of 0.822 (0.783-0.861) and 0.830 (0.792-0.867) in the external validation and 0.871 (0.858-0.884) and 0.879 (0.867-0.892) in the internal validation, respectively.

Conclusions: Three nested 10-year EC absolute risk prediction models for Chinese adults aged 30-79 years were developed and validated, which may be particularly useful for populations in low EC-risk areas. Even the simple model with only 5 predictors available from EHRs had excellent discrimination and good calibration, indicating its potential for broader use in tailored EC prevention. The simple and intermediate models have the potential to be widely used for both primary and secondary prevention of EC.

(JMIR Public Health Surveill 2023;9:e43725) doi: 10.2196/43725

KEYWORDS

esophageal cancer; prediction model; absolute risk; China; prospective cohort; screening; primary prevention; development; external validation; electronic health record

Introduction

China has the largest burden of esophageal cancer (EC), accounting for around half of the global incident cases and deaths in 2018 [1,2]. The prevalence, disability-adjusted life years, and direct medical expenditures are projected to continue to increase [3]. Upper endoscopy has been widely performed for screening and diagnosing EC, but the cost and potential harm of invasive procedures as well as the need for expertise and endoscopy skills training preclude a population-wide application, which may partially explain the poor prognosis of EC. Thus, identifying a high-risk population for endoscopy through prediction models would be more feasible and effective.

In China, 4 diagnostic models (ie, estimating the probability of prevalent EC) have been developed to act as a prescreening tool for endoscopy, with an area under the receiver operating characteristic curve (AUC) ranging from 0.681 to 0.843 [4-7]. However, these models were all developed from populations in high-risk rural areas and may not apply to low-risk rural and urban areas, where a large proportion of cases countrywide occurs [8]. Only 1 model was externally validated [5]. Besides diagnostic models, a few prognostic models (ie, predicting the absolute risk of EC in 5 or more years) have recently been developed from European cohorts [9-12]. These models can be used not only for early detection but also for primary prevention such as intensive lifestyle interventions. However, such models may not perform well for the Chinese population since the risk factor profile is different and the predominant subtype of EC is esophageal squamous cell carcinoma (ESCC) among the Chinese, while esophageal adenocarcinoma is the predominant subtype in the western population [13]. To the best of our knowledge, there is only 1 prognostic model in China, which was developed in a high-risk rural area [14]; this model was based on a case-control design, which was prone to selection bias and recall bias. Moreover, the limited EC cases (n=244) and the lack of external validation could induce overfitting and encroach generalizability.

The national Basic Public Health Service program in China requires establishing health records for all residents [15]. The efficiency and extensive use of population risk stratification for tailored prevention can be greatly improved by embedding prediction models within the electronic health record (EHR) system, that is, by directly estimating the risk of EC and identifying high-risk individuals for EC based on information from EHRs. However, some predictors in the existing models (eg, food temperature, eating speed) are not available in EHRs and need to be additionally collected even in high-risk areas of EC.

To address the above limitations, we used a large prospective cohort of 0.5 million people from both high EC-risk and low EC-risk areas of China for model development and another prospective cohort for external validation. We first constructed a 10-year absolute risk prediction model for EC with the inclusion of established and probable EC risk factors to maximize model performance. Then, we simplified the model by keeping predictors that are available in the Basic Public Health Service health records. We assessed whether the simple but potentially widely applicable model showed acceptable performance in both cohorts.

Methods

Data for Model Development

Data from the China Kadoorie Biobank (CKB), a large-scale nationwide prospective cohort of 512,725 participants aged 30-79 years, were used for model development. The baseline survey was performed between 2004 and 2008 in 10 geographically defined regions (5 urban and 5 rural). The details of the study design and survey methods have been reported previously [16]. Incident cases of EC and all-cause mortality were identified through linkage with the mortality and disease registries and national health insurance claims database, supplemented with local residential records and annual active confirmation. The International Classification of Diseases, 10th revision was used to code all EC (C15) by trained staff who were blinded to the baseline information. The adjudication of the incident cancer cases is ongoing, with medical records of 1283 EC cases having been retrieved, in which 1246 (97.1%) were confirmed as EC and 830 had pathological diagnoses. After excluding 41 cases with subtype reported as unknown, 92.7% (731/789) of the cases were classified as ESCC.

Data for External Validation

An independent prospective cohort from Changzhou of the Jiangsu province, a low EC-risk rural area in China, was used for external validation. In brief, 20,803 participants aged 30 years and older were recruited from 23 villages in 2004-2005. Incident EC cases and all-cause mortality were identified through active follow-up in 2008-2009, 2012-2013, and 2018-2019, and through linkage with the disease and mortality registries. Trained staff who were blinded to baseline information further confirmed suspected cases of nonfatal cancer by reviewing local medical records or visiting village doctors.

Ethics Approval

The study protocol for CKB was approved by the ethics review committee of the Chinese Center for Disease Control and Prevention (Beijing, China: 005/2004) and the Oxford Tropical Research Ethics Committee, University of Oxford (UK: 025–04). The Changzhou cohort was approved by the ethical review committee of the Nanjing Medical University (Nanjing, China), and written informed consent was collected from all the participants.

Predictor Variables

At baseline, all participants in the CKB and Changzhou cohort completed a questionnaire and had physical measurements taken. Candidate predictors were identified based on established risk factors for EC and factors that have been included in previous EC prediction models [17,18]. Candidate predictors included age, sex, smoking, alcohol use, education, household income, marital status, family history of EC, BMI, waist circumference, physical activity, hot food consumption, and consumption of fresh vegetables, fresh fruit, red meat, and preserved vegetables. To model the large geographic disparity in EC incidence in China, we created a variable to denote living in a high-risk or low-risk area. Of the 10 study regions of CKB, we assigned Hui county in Henan province and Pengzhou in Sichuan province to high-risk areas, according to the most recent guideline for EC in China [19,20]. The criteria for defining high-risk areas are described in Multimedia Appendix 1 [19,21-26]. The details of baseline prevalence and incidence of EC by study region are shown in Multimedia Appendix 2. Because data on the family history of EC and hot food consumption were not recorded at baseline in the CKB, we used family history of cancer and hot tea consumption as surrogates for the above-established risk factors. The details of the assessment of predictors are shown in Multimedia Appendix 1.

Statistical Methods

In the CKB cohort, participants who were previously diagnosed with cancer (n=2578) or had missing data on BMI (n=2) were excluded, leaving 510,145 participants for development. In the Changzhou cohort, participants who were previously diagnosed with cancer (n=239), out of the age range of 30-79 years (n=1902), had a recorded implausible censoring date for loss to follow-up (n=5), or had missing data on candidate predictors (n=216) were excluded, leaving 18,441 participants for external validation. Participants were considered at risk from enrollment to the first date of diagnosis of EC, death, loss to follow-up, or end of follow-up (CKB: December 31, 2017; Changzhou cohort: January 31, 2019).

Model Development

Based on the whole CKB data set, we separately fitted a model for EC and a model for all-cause mortality. For the EC model, a flexible parametric model on the cumulative hazard scale was used to estimate the baseline hazards and hazard ratios of the predictors for EC, with age as the time scale [27]. Age was modeled using restricted cubic splines with boundary knots at 30 and 90 and internal knots at 60 and 70. The established risk factors of EC (age, sex, smoking, and alcohol use) and regional risk level (high-risk/low-risk areas) entered the model directly. Two strategies were employed for the selection of the other predictors. First, other candidate predictors were included in the full model and kept if $P<.05$. Second, the predictor selection was repeated using stepwise backward elimination. Two strategies selected the same set of predictors. The variable grouping was determined using the Bayesian information criteria. All 2-way interactions were tested, but none of those significantly improved model performance. Further, we simplified the full model by keeping only predictors available in the health records. As age is the most important predictor, we also constructed an age-only model for comparison. Therefore, 4 models were constructed, with predictors included in the model: (1) age-only: age; (2) simple model: age, sex, regional risk level, education, and family history of cancer, which are available for all residents in the health records; (3) intermediate model: simple model plus smoking, alcohol use, and BMI, which are additionally available for residents aged 65 years and older, and diabetic or hypertensive patients in the health records; (4) full model: intermediate model plus physical activity, hot tea consumption, and fresh fruit consumption, which go beyond the available health records but have the potential to improve the risk prediction. We then used the same settings of the flexible parametric model to model the hazards of all-cause mortality, with sex, residence area (urban/rural), and regional risk level in the model. We used cause-specific hazard models to account for the competing risks. Briefly, the 10-year absolute risk (AR) of EC for a participant who is age a is calculated as

$$\mathrm{AR}(\mathrm{a}+10/\mathrm{a})=\int_{a}^{a+10} h_1(t)\exp\left[-\int_{a}^{t}(h_1(u)+h_2(u))du\right]\mathrm{d}t$$

Model Validation

The methods for model validation are detailed in Multimedia Appendix 1. In brief, we externally validated the age-only, simple, and intermediate models, but not the full model, because data on physical activity in metabolic equivalent of task-hours and hot tea consumption were unavailable in the Changzhou cohort. We also conducted an internal validation in the CKB by using data splitting and 500-sample bootstrapping. Cancer-free participants whose retention in the cohorts was less than 10 years were included to test calibration but were excluded from other validation measures, since it was unknown whether they could have experienced an EC if they had been followed up to 10 years.

Discrimination was quantified by calculating the AUC. Calibration was assessed by plotting the observed risk obtained using Kaplan-Meier analyses against the predicted risk by decile. Because of the large geographical variation in the incidence of EC in China, we recalibrated the models by using the method proposed by the World Health Organization Cardiovascular Disease Risk Chart Working Group with a slight modification [21]. Further, continuous Net Reclassification Improvement and Integrated Discrimination Improvement were used to evaluate the added predictive ability of additional predictors [28,29]. In the internal validation using data splitting, calibration and discrimination were also assessed in subgroups defined by regional risk level, residence area, sex, age group, and special population aged 65 years and older or with diabetes or hypertension who are of particular concern to the Basic Public Health Service. To offer a reference for primary care practices, we estimated a range of performance indices corresponding to a series of cutoffs.

Several sensitivity analyses were conducted. First, we separately developed 2 models for high EC-risk (high-risk model) and low EC-risk (low-risk model) areas by using the same strategy as the primary analyses and assessed their discrimination and calibration in the corresponding areas. Second, we restricted EC cases to (1) pathologically confirmed cases, (2) cases that were pathologically confirmed as ESCC, (3) cases that were pathologically confirmed but not as ESCC, and (4) cases that were pathologically confirmed but not as ESCC (scenario 3) or that were not pathologically confirmed. In the above 4 scenarios, we excluded EC cases that did not meet the corresponding criteria and examined the discriminating ability of the models (Multimedia Appendix 3). Third, since some asymptomatic EC cases might be undiagnosed, we excluded the EC cases documented in the first year of follow-up and used the same strategy to develop and validate the models.

Results

The mean age of the 510,145 participants in the CKB and 18,441 participants in the Changzhou cohort was 52.0 (SD 10.7) years and 51.2 (SD 12.1) years, respectively. The details of the baseline characteristics of the predictors are described in Table 1. During a median of 11.1 (IQR 10.2-12.1) years of follow-up of the CKB, we identified 2550 EC cases, with an incidence (per 100,000 person-years) of 46.2. High EC-risk areas had a significantly higher incidence than low EC-risk areas (132.2 vs 20.2, respectively). In the Changzhou cohort, 114 EC cases were identified during a median follow-up of 13.6 (IQR 13.5-14.4) years, with an incidence of 47.1.

Table 2 and Multimedia Appendix 4 list the hazard ratios and 95% CIs for predictors of EC and all-cause mortality in the CKB. Male, living in high-risk areas, less educated, having a family history of cancer, smoking, alcohol use, underweight, less physical activity, preferring burning hot tea, and rare intake of fresh fruits were associated with a greater risk of EC.

In the external validation, the simple and intermediate models exhibited similar and excellent discriminating ability with AUCs (95% CIs) of 0.822 (0.783-0.861) and 0.830 (0.792-0.867), respectively (Figure 1). In the internal validation, the AUCs (95% CIs) of the simple, intermediate, and full models were 0.871 (0.858-0.884), 0.879 (0.867-0.892), and 0.883 (0.871-0.895), respectively (Figure 1). Although there were only limited increases in the AUCs with more predictors included in the models, continuous Net Reclassification Improvement and Integrated Discrimination Improvement indicated improved accuracy of the predicted risks for both cases and those that were not cases (Multimedia Appendix 5). In the less biased internal validation method of bootstrapping, the above results were not greatly altered (Multimedia Appendix 6). The original simple and intermediate models significantly underestimated the risk of EC in the Changzhou cohort. The recalibration parameters, b and k, were 1.22 and 1.97, respectively. Age-specific observed risks of EC used to calculate b and k are shown in Multimedia Appendix 7. After recalibration, the calibration plot showed excellent agreement between the observed and predicted risks for the simple and intermediate models (Figure 2). In the internal validation, the predicted risk of the simple, intermediate, and full models agreed well with the observed risk by a tenth of the predicted risk, except for the top 2 deciles where slight underestimations seemed to have occurred (Figure 3).

The density of the predicted risks of models in cases was greater than that in those that were not cases (Multimedia Appendix 8 and Multimedia Appendix 9). The performance of the models across a series of cutoffs is presented in Multimedia Appendix 10. Compared with their counterparts, the models discriminated better in low-risk areas, rural areas, women, or middle-aged adults younger than 65 years without diabetes and hypertension in the internal validation (Multimedia Appendix 11). The predicted risks agreed well with the observed risks in all subgroups.

In the sensitivity analysis, we separately developed 2 models for high-risk and low-risk areas. The included predictors and the hazard ratios (95% CIs) are listed in Multimedia Appendix 12. When these 2 models were applied in their corresponding validation set, the model for low-risk areas performed better than the models in the primary analyses (Multimedia Appendix 5 and Multimedia Appendix 13). When we took the availability and results of pathology reports into consideration, models had excellent discriminating ability in all scenarios (Multimedia Appendix 3 and Multimedia Appendix 14). Excluding EC cases occurring in the first year of follow-up did not alter the performance of the models (Multimedia Appendix 15).

Table 1. Baseline characteristics of the participants by disease status in the China Kadoorie Biobank and Changzhou cohort.

	China Kadoorie Biobank			Changzhou cohort		
	EC[a] case (n=2550)	Not an EC case (n=507,595)	Total (N=510,145)	EC case (n=114)	Not an EC case (n=18,327)	Total (N=18,441)
Age (years), mean (SD)	60.7 (8.7)	52.0 (10.7)	52.0 (10.7)	60.9 (9.1)	51.1 (12.1)	51.2 (12.1)
Male, n (%)	1757 (68.9)	207,477 (40.9)	209,234 (41)	77 (67.5)	7611 (41.5)	7688 (41.7)
Urban, n (%)	468 (18.4)	224,300 (44.2)	224,768 (44.1)	0 (0)	0 (0)	0 (0)
High-risk area[b], n (%)	1692 (66.4)	116,715 (23)	118,407 (23.2)	0 (0)	0 (0)	0 (0)
Family history of cancer, n (%)	720 (28.2)	84,948 (16.7)	85,668 (16.8)	26 (22.8)	3354 (18.3)	3380 (18.3)
High level of physical activity[c], n (%)	565 (22.2)	126,739 (25)	127,304 (25)	—[d]	—	—
Highest education, n (%)						
Illiterate or primary school	1917 (75.2)	257,088 (50.6)	259,005 (50.8)	70 (61.4)	8372 (45.7)	8442 (45.8)
Middle or high school	598 (23.5)	220,780 (43.5)	221,378 (43.4)	43 (37.7)	9841 (53.7)	9884 (53.6)
College or university	35 (1.4)	29,727 (5.9)	29,762 (5.8)	1 (0.9)	114 (0.6)	115 (0.6)
Current smoking (cigarettes or equivalent per day), n (%)						
<30	979 (38.4)	114,815 (22.6)	115,794 (22.7)	49 (43)	4230 (23.1)	4279 (23.2)
≥30	235 (9.2)	19,155 (3.8)	19,390 (3.8)	8 (7)	713 (3.9)	721 (3.9)
Daily alcohol use (grams of pure alcohol per day), n (%)						
<30	70 (2.7)	11,503 (2.3)	11,573 (2.3)	9 (7.9)	1123 (6.1)	1132 (6.1)
30-59	161 (6.3)	14,884 (2.9)	15,045 (2.9)	9 (7.9)	1049 (5.7)	1058 (5.7)
≥60	386 (15.1)	19,085 (3.8)	19,471 (3.8)	31 (29)	1956 (10.7)	1989 (10.8)
BMI (kg/m^2), n (%)						
<18.5	175 (6.9)	21,965 (4.3)	22,140 (4.3)	4 (3.5)	984 (5.4)	988 (5.4)
18.5-23.9	1482 (58.1)	263,169 (51.8)	264,651 (51.9)	76 (66.7)	9967 (54.4)	10,043 (54.4)
≥24.0	893 (35)	222,461 (43.8)	223,354 (43.8)	34 (29.8)	7376 (40.3)	7410 (40.2)
Tea temperature preference, n (%)						
Not daily drinker/warm tea drinker	2090 (82)	426,628 (84)	428,718 (84)	—	—	—
Hot tea	311 (12.2)	59,425 (11.7)	59,736 (11.7)	—	—	—
Burning hot tea	149 (5.8)	21,542 (4.2)	21,691 (4.3)	—	—	—
Fresh fruit consumption[e], n (%)						
Daily	165 (6.5)	95,715 (18.9)	95,880 (18.8)	4 (3.5)	813 (4.4)	817 (4.4)
Weekly	658 (25.8)	207,716 (40.9)	208,374 (40.8)	98 (86)	15,921 (86.9)	16,019 (86.9)
Less than weekly	1727 (67.7)	204,164 (40.2)	205,891 (40.4)	12 (10.5)	1591 (8.7)	1603 (8.7)

[a]EC: esophageal cancer.

[b]High-risk area denotes Hui county in Henan province and Pengzhou in Sichuan province in our study.

[c]High-level physical activity was defined as age-specific and sex-specific upper quarter of total physical activity level measured by metabolic equivalent of task-hours per day.

[d]Not available.

[e]Data on the fresh fruit consumption of 2 participants in the Changzhou cohort were missing.

Table 2. Hazard ratios (95% CIs) for the predictor variables of esophageal cancer in the China Kadoorie Biobank.

	Cases (n)	Cases/person years (1/100,000)	Age-only model, HR[a] (95% CI)	Simple model, HR (95% CI)	Intermediate model, HR (95% CI)	Full model, HR (95% CI)
Spline basis of age (knots: 30, 60, 70, 90)						
First	N/A[b]	N/A	3.56 (3.29-3.86)	3.28 (3.03-3.56)	3.30 (3.04-3.59)	3.28 (3.02-3.56)
Second	N/A	N/A	1.17 (1.11-1.23)	1.15 (1.09-1.21)	1.14 (1.09-1.20)	1.14 (1.09-1.20)
Third	N/A	N/A	1.02 (1.00-1.04)	1.00 (0.98-1.02)	1.00 (0.98-1.02)	1.00 (0.98-1.02)
Sex						
Male	1757	79.12	N/A	Reference	Reference	Reference
Female	793	24.02	N/A	0.31 (0.28-0.34)	0.40 (0.36-0.44)	0.42 (0.37-0.46)
High-risk area[c]						
No	858	20.22	N/A	Reference	Reference	Reference
Yes	1692	132.22	N/A	6.31 (5.81-6.86)	6.07 (5.58-6.60)	5.61 (5.14-6.13)
Highest education						
Illiterate or primary school	1917	69.06	N/A	Reference	Reference	Reference
Middle or high school	598	24.68	N/A	0.60 (0.55-0.66)	0.65 (0.59-0.72)	0.68 (0.62-0.76)
College or university	35	10.82	N/A	0.32 (0.23-0.44)	0.37 (0.26-0.52)	0.45 (0.32-0.63)
Family history of cancer						
No	1830	39.88	N/A	Reference	Reference	Reference
Yes	720	77.14	N/A	1.71 (1.57-1.86)	1.78 (1.63-1.94)	1.74 (1.59-1.89)
Current smoking						
No	1336	32.72	N/A	N/A	Reference	Reference
Cigarettes or equivalent per day among smokers						
<30	979	79.48	N/A	N/A	1.15 (1.05-1.27)	1.12 (1.02-1.24)
≥30	235	112.95	N/A	N/A	1.25 (1.07-1.47)	1.24 (1.06-1.45)
Daily alcohol use						
No	1933	38.43	N/A	N/A	Reference	Reference
Grams of pure alcohol per day among alcohol consumers						
<30	70	56.84	N/A	N/A	0.97 (0.77-1.24)	1.03 (0.81-1.30)
30-59	161	100.03	N/A	N/A	1.39 (1.18-1.64)	1.43 (1.21-1.69)
≥60	386	185.48	N/A	N/A	2.01 (1.78-2.26)	2.06 (1.82-2.32)
BMI (kg/m^2)						
<18.5	175	77.77	N/A	N/A	Reference	Reference
18.5-23.9	1482	51.70	N/A	N/A	0.73 (0.62-0.85)	0.76 (0.65-0.88)
≥24.0	893	36.73	N/A	N/A	0.61 (0.52-0.72)	0.64 (0.54-0.76)
Physical activity						
Low	1985	48.10	N/A	N/A	N/A	Reference
High[d]	565	40.49	N/A	N/A	N/A	0.80 (0.73-0.88)
Tea temperature preference						
Not a daily drinker or warm tea drinker	2090	44.96	N/A	N/A	N/A	Reference

	Cases (n)	Cases/person years (1/100,000)	Age-only model, HR[a] (95% CI)	Simple model, HR (95% CI)	Intermediate model, HR (95% CI)	Full model, HR (95% CI)
Hot tea	311	48.27	N/A	N/A	N/A	1.03 (0.91-1.17)
Burning hot tea	149	64.89	N/A	N/A	N/A	1.49 (1.25-1.77)
Fresh fruit consumption						
Daily	165	15.74	N/A	N/A	N/A	Reference
Weekly	658	28.97	N/A	N/A	N/A	1.07 (0.90-1.27)
Less than weekly	1727	78.38	N/A	N/A	N/A	1.79 (1.51-2.12)

[a]HR: hazard ratio.

[b]N/A: not applicable.

[c]High-risk area denotes Hui county in Henan province and Pengzhou in Sichuan province in our study.

[d]High-level physical activity was defined as age-specific and sex-specific upper quarter of total physical activity level measured by metabolic equivalent of task-hours per day.

Figure 1. Receiver operating characteristic curves and corresponding areas under the receiver operating characteristic curve for the esophageal cancer prediction models. (A) Internal validation in the China Kadoorie Biobank using data splitting. (B) External validation in the Changzhou cohort. The models included age (age-only model), sex, regional risk level, education, family history of cancer (simple model), smoking, alcohol use, BMI (intermediate model), physical activity, hot tea consumption, and fresh fruit consumption (full model). AUC: area under the receiver operating characteristic curve.

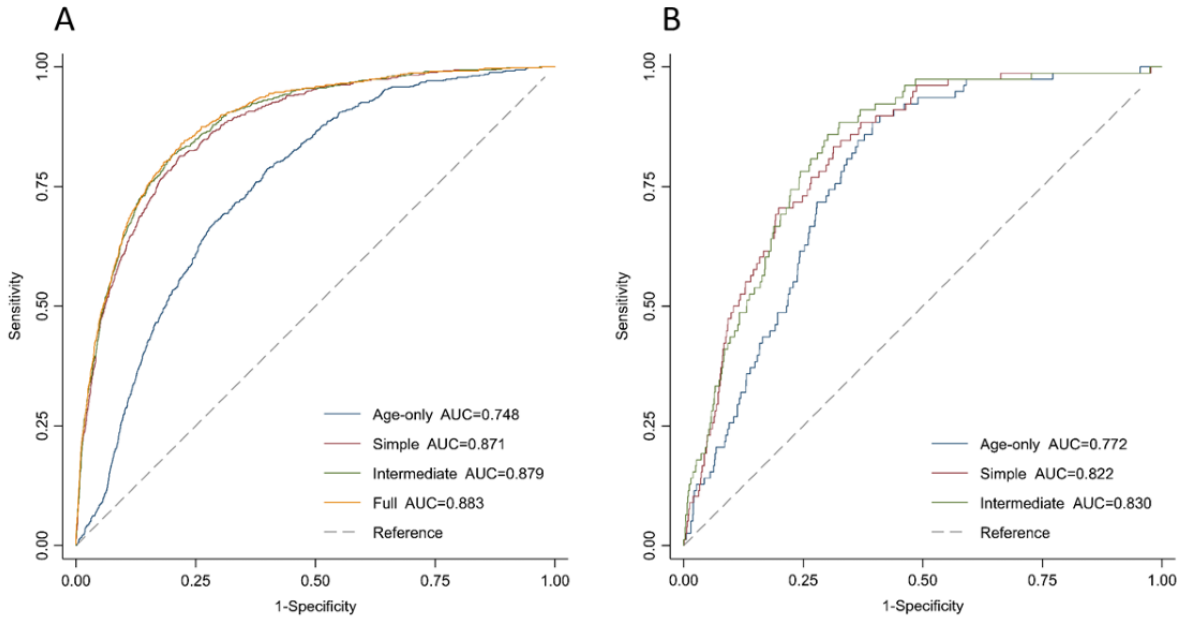


Figure 2. Calibration plot of the esophageal cancer prediction models in the Changzhou cohort. Calibration of the original (A) age-only, (C) simple, and (E) intermediate models. Calibration of the recalibrated (B) age-only, (D) simple, and (F) intermediate models. The observed 10-year risk was estimated by Kaplan-Meier analysis and plotted against model-predicted risk by decile. Models were recalibrated using the method proposed by the World Health Organization Cardiovascular Disease Risk Chart Working Group with a slight modification. For details, see Multimedia Appendix 1.

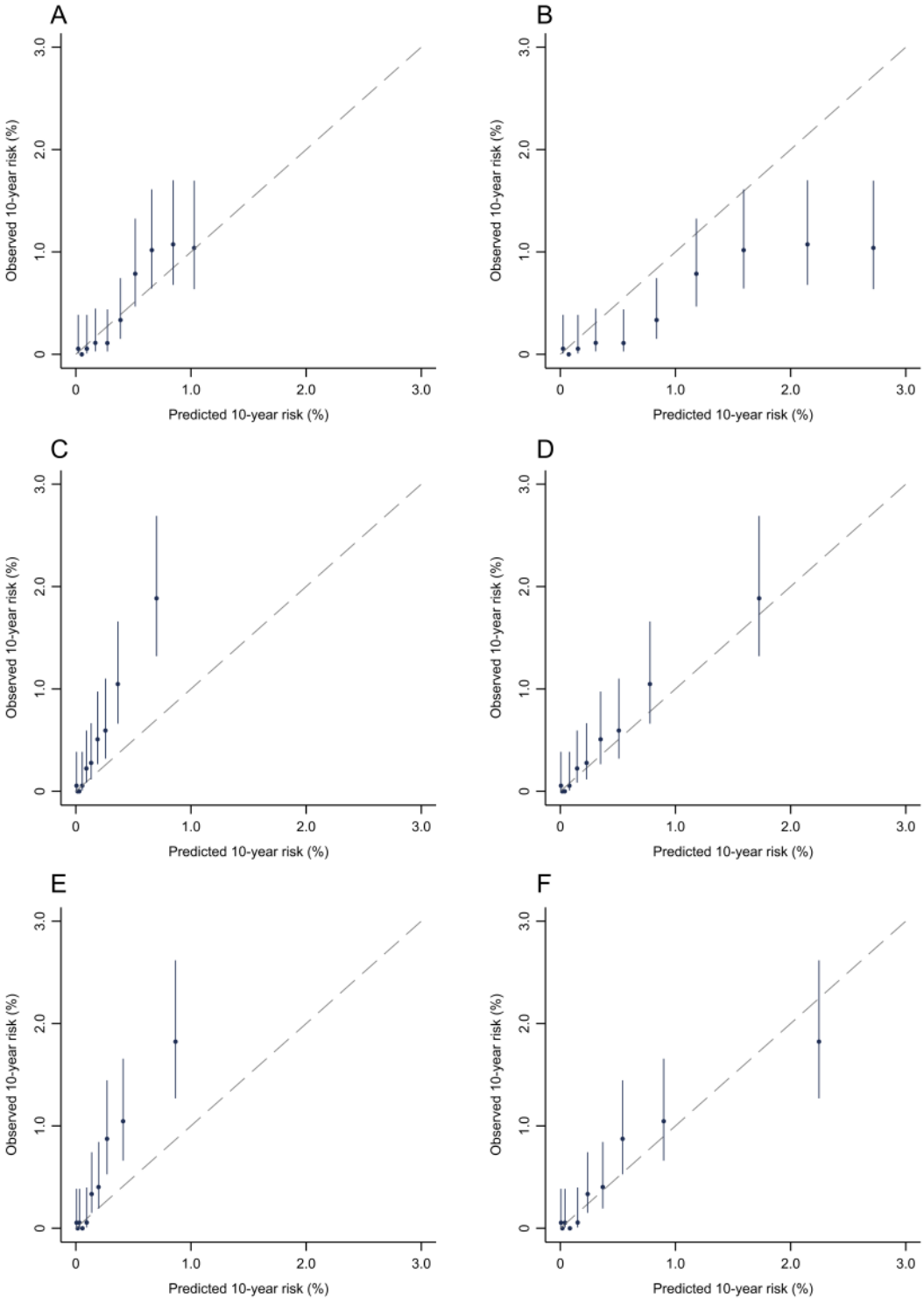
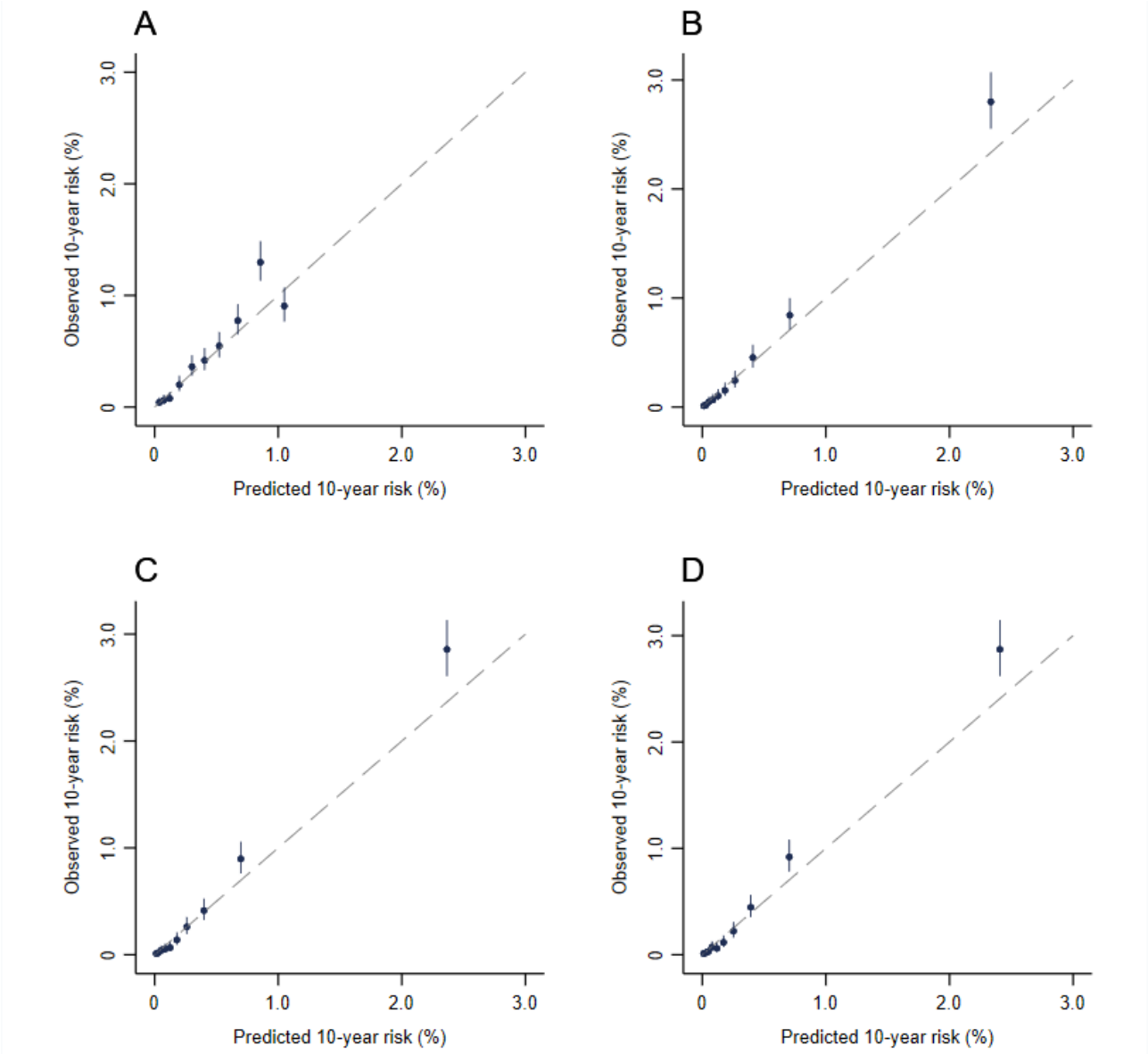


Figure 3. Calibration plot of the esophageal cancer prediction models in the China Kadoorie Biobank by using data splitting. (A) Age-only model. (B) Simple model. (C) Intermediate model. (D) Full model. Models were fitted to a random two-thirds of the China Kadoorie Biobank data and evaluated on the remaining one-third. The calibration plots in the validation set were plotted. The observed 10-year risk was estimated by Kaplan-Meier analyses and plotted against model-predicted risk by decile.



Discussion

In a large prospective cohort study, we developed 3 nested 10-year EC absolute risk prediction models for Chinese adults aged 30-79 years. The models included age, sex, regional EC-risk level, education, family history of cancer (simple model), smoking, alcohol use, BMI (intermediate model), physical activity, hot tea consumption, and fresh fruit consumption (full model). The simple and intermediate models were externally validated in an independent prospective cohort and they exhibited excellent discrimination and good calibration. The performance of these models was compromised by keeping only predictors available in the health records but only to a small and acceptable extent.

The models that we constructed included established risk factors for EC (eg, age, smoking, alcohol use) and factors associated with increased EC risk in the CKB and in previous studies [17,18]. A previous review attributed the geographical variation in the incidence of EC in China to some unique factors in high-risk areas, such as exposure to carcinogens (eg, nitrosamines, their precursors) via water, food, and other sources [30]. To capture this variation as well as to denote some unmeasured unique factors in high-risk areas, we included regional risk level in our models. Although this predictor contributed the most to the model performance with a hazard ratio of around 6 (Table 2), our stratified validation showed that the other predictors still maintained excellent performance in both high-risk and low-risk areas (Multimedia Appendix 11).

Previous models included clinical symptoms such as dysphagia and poststernal pain to identify high-risk individuals with prevalent EC for further endoscopies [4-6]. In contrast, our models, which were intended to identify individuals at risk for developing EC in the next 10 years, did not include clinical symptoms. In a prior model developed based on a hospital-based

case-control study, 25 single-nucleotide polymorphisms, in addition to age, smoking, and alcohol use, resulted in an increased AUC from 0.639 to 0.707 [7]. Some other factors such as exposure to cooking fumes, pesticides, or salty foods were also included in previous models. To develop a parsimonious model that can be potentially used widely, we did not consider genetic variants and less well-established risk factors. Nevertheless, the AUCs of our models were still higher than those of most of the previous Chinese models (range 0.681-0.843) [4-7,14].

As expected, our finding that the simple and intermediate models retained similar performance as the full model despite the fewer predictors included is reasonable since the lost information due to the removal of the predictors was more or less supplemented by other correlated predictors. A previous study showed that the discriminatory information needed for the same unit of increase of AUC exponentially increased with AUC [31]. Thus, an already high AUC of >0.8 for the simple model can only be improved by highly informative predictors. Given the similar performance and excellent discriminating ability, it is acceptable to use the simple or intermediate model in situations where the EHRs are complete and up-to-date and easily implemented in a lower-cost way than an organized screening program. Further, we noticed that the same predictors could contribute differently in subpopulations. For example, the inclusion of lifestyle factors barely improved the discriminating ability in women (Multimedia Appendix 11), which may be explained by the low prevalence and dosage intensity of smoking and alcohol use in Chinese women.

For most prediction models, underestimations or overestimations are commonly observed in an external validation, which were also observed in our study. However, across the groups defined by the deciles of predicted risks, the observed risks proportionally increased with increased predicted risk rather than an irregular misestimation. More importantly, the underestimation disappeared after recalibration. Such results implied that the predictors in our models are predictive, the coefficients estimated in the CKB are robust and generalizable, and the underestimation was mainly caused by the mismatch of EC incidence between the CKB and Changzhou cohort.

Unlike the models in previous studies, our models calculated the absolute risk of EC instead of the relative risk and could facilitate primary prevention of EC. The essence of intuition of the absolute risk can not only raise population awareness and motivate adherence to lifestyle changes but also enhance effective communication between health professionals and individuals and help health professionals identify high-risk populations for intensive lifestyle interventions. Further, several predictors in our models are modifiable, such as smoking, alcohol use, and BMI, which could be treated as targets of intervention.

Our study has several strengths. We used a large prospective cohort with the largest number of EC cases from urban, rural, high EC-risk, and low EC-risk areas in China for model development and used another prospective cohort for external validation. This method ensures that our models are robust and potentially generalizable to a wide range of areas. To the best of our knowledge, ours is the first study to develop and externally validate EC models by using 2 independent prospective Chinese cohorts. Last but not the least, we developed and validated an abbreviated version of the risk prediction model that could be easily embedded within the EHR system and enable an efficient and automatic population risk stratification. To facilitate the usage of our models, we provide an easy-to-use Stata code and example in Multimedia Appendix 16 (Stata calculator, which is a modified version of the code shared by Dr Muller) [22].

Some limitations of our study merit consideration. First, some EC cases in the CKB had only clinical diagnoses but no pathological diagnoses for various reasons. Therefore, we could not exclude cases of esophageal adenocarcinoma. However, more than 90% of the EC cases are ESCC in China [32], which was confirmed by our ongoing adjudication of incident EC cases in the CKB. More importantly, models maintained high discriminating abilities when we restricted EC cases to those with a pathological diagnosis of ESCC or those without a pathological diagnosis of ESCC. Second, we did not collect information on the family history of EC specifically and their preference for hot foods and drinks in the baseline survey. Therefore, we used the family history of any cancer and preference for very hot tea consumption as surrogates. Third, although we found limited improvement in AUC by including more predictors in the model (full model), whether other established risk factors of EC, such as disease history of the esophagus and genetic predisposition of EC could further improve the AUC warrants future research. Fourth, we only externally validated our models in low EC-risk rural areas. Further validations in other areas are warranted.

In summary, using data from 2 prospective cohorts, we developed and validated 3 nested 10-year EC absolute risk prediction models for Chinese adults, which may be particularly useful for populations in low EC-risk areas. Even the simple model with only 5 predictors available from residents' EHRs showed excellent discrimination and good calibration, indicating its potential for broader use in tailored EC prevention. Further research is needed to assess the real-world performance in aiding population-wide stratification, identify optimal risk cutoffs for initiating intensive lifestyle interventions and endoscopy screening, and establish an optimal screening protocol (including multistage screening) for individuals or regions with different risks.

Acknowledgments

We thank the participants in this study and the members of the survey teams in each of the 10 regional centers as well as the project development and management teams based at Beijing, Oxford, and the 10 regional centers. This work was supported by the National Natural Science Foundation of China (82192904, 82192901, 82192900, 82125033, 81922061). Y Han is supported by the China Postdoctoral Science Foundation (BX20220018, 2022M720007). The China Kadoorie Biobank baseline survey and

the first resurvey were supported by a grant from the Kadoorie Charitable Foundation in Hong Kong. The long-term follow-up is supported by grants from the UK Wellcome Trust (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z), grants from the National Key R&D Program of China (2016YFC0900500), National Natural Science Foundation of China (81390540, 91846303, 81941018), and Chinese Ministry of Science and Technology (2011BAI09B01). The funders had no role in the study design, data collection, data analysis and interpretation, writing of the report, or the decision to submit the paper for publication.

Data Availability

Details of how to access the China Kadoorie Biobank data and details of the data release schedule are available from [33].

Authors' Contributions

JL and LL conceived and designed the study with equal contribution to this work. LL, ZC, and JC, as members of the China Kadoorie Biobank (CKB) steering committee, designed and supervised the conduct of the CKB study, obtained funding, and together with CY, YG, YP, PP, DS, LY, YC, HD, and MY acquired the data for the CKB study. D Hang, HM, GJ, ZH, and HS designed and supervised the conduct of the Changzhou cohort study. Y Han and Y Hu analyzed the CKB data, and XZ analyzed the Changzhou cohort data. Y Han wrote the first draft of the manuscript. JL, D Huo, and LL contributed to the interpretation of the results and critical revision of the manuscript for important intellectual content and approved the final version of the manuscript. All authors reviewed and approved the final manuscript. JL and LL are the guarantors.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods.
[DOCX File , 24 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Prevalence and incidence of esophageal cancer by study region.
[DOCX File , 30 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Design of the sensitivity analysis in consideration of pathology reports.
[DOCX File , 379 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Hazard ratios (95% CIs) for predictor variables of all-cause mortality in the China Kadoorie Biobank.
[DOCX File , 24 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Comparison of the area under the receiver operating characteristic curve, continuous Net Reclassification Improvement, and Integrated Discrimination Improvement of esophageal cancer prediction models in the China Kadoorie Biobank and Changzhou cohort.
[DOCX File , 26 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Comparison of the area under the receiver operating characteristic curve, continuous Net Reclassification Improvement, and Integrated Discrimination Improvement of esophageal cancer prediction models in the China Kadoorie Biobank by using bootstrap.
[DOCX File , 25 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Age-specific observed risk of esophageal cancer in low-risk areas of the China Kadoorie Biobank and Changzhou cohort.
[DOCX File , 24 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Discriminating ability of the recalibrated prediction models in the Changzhou cohort.
[DOCX File , 395 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Discriminating ability of the prediction models in the China Kadoorie Biobank using data splitting.
[DOCX File , 394 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Performance of the esophageal cancer prediction model across different predicted risk cutoffs in the China Kadoorie Biobank and Changzhou cohort.
[DOCX File , 45 KB-Multimedia Appendix 10]

Multimedia Appendix 11

Discrimination and calibration of the intermediate model in subcohorts of the China Kadoorie Biobank by data splitting.
[DOCX File , 569 KB-Multimedia Appendix 11]

Multimedia Appendix 12

Hazard ratios (95% CIs) for predictor variables of esophageal cancer prediction models developed separately in high-risk and low-risk areas of the derivation subcohort of the China Kadoorie Biobank.
[DOCX File , 31 KB-Multimedia Appendix 12]

Multimedia Appendix 13

Model performance of 2 esophageal cancer prediction models developed separately in high-risk and low-risk areas of the derivation subcohort of the China Kadoorie Biobank and applied in the corresponding validation subcohort.
[DOCX File , 74 KB-Multimedia Appendix 13]

Multimedia Appendix 14

Discriminating ability of the esophageal cancer prediction models in the China Kadoorie Biobank by data splitting in consideration of pathology reports.
[DOCX File , 195 KB-Multimedia Appendix 14]

Multimedia Appendix 15

Model performance of the esophageal cancer prediction models in the China Kadoorie Biobank using data splitting after excluding esophageal cancer cases occurring in the first year of follow-up.
[DOCX File , 1658 KB-Multimedia Appendix 15]

Multimedia Appendix 16

Stata calculator.
[ZIP File (Zip Archive), 62 KB-Multimedia Appendix 16]

References

1. Arnold M, Ferlay J, van Berge Henegouwen MI, Soerjomataram I. Global burden of oesophageal and gastric cancer by histology and subsite in 2018. Gut 2020 Sep;69(9):1564-1571. [doi: 10.1136/gutjnl-2020-321600] [Medline: 32606208]
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 May;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
3. Li Y, Xu J, Gu Y, Sun X, Dong H, Chen C. The Disease and Economic Burdens of Esophageal Cancer in China from 2013 to 2030: Dynamic Cohort Modeling Study. JMIR Public Health Surveill 2022 Mar 02;8(3):e33191 [FREE Full text] [doi: 10.2196/33191] [Medline: 34963658]
4. Chen W, Li H, Ren J, Zheng R, Shi J, Li J, et al. Selection of high-risk individuals for esophageal cancer screening: A prediction model of esophageal squamous cell carcinoma based on a multicenter screening cohort in rural China. Int J Cancer 2021 Jan 15;148(2):329-339 [FREE Full text] [doi: 10.1002/ijc.33208] [Medline: 32663318]

5. Liu Z, Guo C, He Y, Chen Y, Ji P, Fang Z, et al. A clinical model predicting the risk of esophageal high-grade lesions in opportunistic screening: a multicenter real-world study in China. Gastrointest Endosc 2020 Jun;91(6):1253-1260.e3. [doi: 10.1016/j.gie.2019.12.038] [Medline: 31911077]
6. Liu M, Liu Z, Cai H, Guo C, Li X, Zhang C, et al. A Model To Identify Individuals at High Risk for Esophageal Squamous Cell Carcinoma and Precancerous Lesions in Regions of High Prevalence in China. Clin Gastroenterol Hepatol 2017 Oct;15(10):1538-1546.e7. [doi: 10.1016/j.cgh.2017.03.019] [Medline: 28342951]
7. Chang J, Huang Y, Wei L, Ma B, Miao X, Li Y, et al. Risk prediction of esophageal squamous-cell carcinoma with common genetic variants and lifestyle factors in Chinese population. Carcinogenesis 2013 Aug;34(8):1782-1786. [doi: 10.1093/carcin/bgt106] [Medline: 23536576]
8. Chen R, Zheng RS, Zhang SW, Zeng HM, Wang SM, Sun KX, et al. [Analysis of incidence and mortality of esophageal cancer in China, 2015]. Zhonghua Yu Fang Yi Xue Za Zhi 2019 Nov 06;53(11):1094-1097. [doi: 10.3760/cma.j.issn.0253-9624.2019.11.004] [Medline: 31683393]
9. Kunzmann AT, Thrift AP, Cardwell CR, Lagergren J, Xie S, Johnston BT, McMenamin, et al. Model for Identifying Individuals at Risk for Esophageal Adenocarcinoma. Clin Gastroenterol Hepatol 2018 Aug;16(8):1229-1236.e4. [doi: 10.1016/j.cgh.2018.03.014] [Medline: 29559360]
10. Xie S, Ness-Jensen E, Medefelt N, Lagergren J. Assessing the feasibility of targeted screening for esophageal adenocarcinoma based on individual risk assessment in a population-based cohort study in Norway (The HUNT Study). Am J Gastroenterol 2018 Jun;113(6):829-835. [doi: 10.1038/s41395-018-0069-9] [Medline: 29748563]
11. Wang Q, Lagergren J, Xie S. Prediction of individuals at high absolute risk of esophageal squamous cell carcinoma. Gastrointest Endosc 2019 Apr;89(4):726-732.e2. [doi: 10.1016/j.gie.2018.10.025] [Medline: 30616974]
12. Wang Q, Ness-Jensen E, Santoni G, Xie S, Lagergren J. Development and Validation of a Risk Prediction Model for Esophageal Squamous Cell Carcinoma Using Cohort Studies. Am J Gastroenterol 2021 Apr;116(4):683-691. [doi: 10.14309/ajg.0000000000001094] [Medline: 33982937]
13. Gao QY, Fang JY. Early esophageal cancer screening in China. Best Pract Res Clin Gastroenterol 2015 Dec;29(6):885-893. [doi: 10.1016/j.bpg.2015.09.018] [Medline: 26651250]
14. Shen Y, Xie S, Zhao L, Song G, Shao Y, Hao C, et al. Estimating Individualized Absolute Risk for Esophageal Squamous Cell Carcinoma: A Population-Based Study in High-Risk Areas of China. Front Oncol 2020;10:598603. [doi: 10.3389/fonc.2020.598603] [Medline: 33489898]
15. Yuan B, Balabanova D, Gao J, Tang S, Guo Y. Strengthening public health services to achieve universal health coverage in China. BMJ 2019 Jun 21;365:l2358 [FREE Full text] [doi: 10.1136/bmj.l2358] [Medline: 31227480]
16. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, China Kadoorie Biobank (CKB) collaborative group. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011 Dec;40(6):1652-1666 [FREE Full text] [doi: 10.1093/ije/dyr120] [Medline: 22158673]
17. World Cancer Research Fund, American Insititute for Cancer Research. Diet, nutrition, physical activity and oesophageal cancer. Continuous Update Project Expert Report. URL: https://www.wcrf.org/wp-content/uploads/2021/02/oesophageal-cancer-report.pdf [accessed 2023-02-28]
18. Yu C, Tang H, Guo Y, Bian Z, Yang L, Chen Y, China Kadoorie Biobank Collaborative Group. Hot Tea Consumption and Its Interactions With Alcohol and Tobacco Use on the Risk for Esophageal Cancer: A Population-Based Cohort Study. Ann Intern Med 2018 Apr 03;168(7):489-497 [FREE Full text] [doi: 10.7326/M17-2000] [Medline: 29404576]
19. He J, Chen WQ, Li ZS, Li N, Ren JS, Tian JH, Expert Group of China Guideline for the Screening, Early Detection and Early Treatment of Esophageal Cancer, Work Group of China Guideline for the Screening, Early Detection and Early Treatment of Esophageal Cancer. [China guideline for the screening, early detection and early treatment of esophageal cancer (2022, Beijing)]. Zhonghua Zhong Liu Za Zhi 2022 Jun 23;44(6):491-522. [doi: 10.3760/cma.j.cn112152-20220517-00348] [Medline: 35754225]
20. Bai Y, Yang F, Liu C, Li DF, Wang S, Lin R, National Clinical Research Center for Digestive Diseases (Shanghai), Chinese Society of Digestive Endoscopy, Digestive Endoscopy Professional Committee of Chinese Endoscopist Association, Cancer Endoscopy Professional Committee of China Anti-Cancer Association. Expert consensus on the clinical application of high-frequency electrosurgery in digestive endoscopy (2020, Shanghai). J Dig Dis 2022 Jan;23(1):2-12. [doi: 10.1111/1751-2980.13074] [Medline: 34953023]
21. WHO CVD Risk Chart Working Group. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Glob Health 2019 Oct;7(10):e1332-e1345 [FREE Full text] [doi: 10.1016/S2214-109X(19)30318-3] [Medline: 31488387]
22. Muller DC, Johansson M, Brennan P. Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study. J Clin Oncol 2017 Mar 10;35(8):861-869 [FREE Full text] [doi: 10.1200/JCO.2016.69.2467] [Medline: 28095156]
23. Millwood IY, Li L, Smith M, Guo Y, Yang L, Bian Z, China Kadoorie Biobank collaborative group. Alcohol consumption in 0.5 million people from 10 diverse regions of China: prevalence, patterns and socio-demographic and health-related correlates. Int J Epidemiol 2013 Jun;42(3):816-827 [FREE Full text] [doi: 10.1093/ije/dyt078] [Medline: 23918852]

24. Du H, Bennett D, Li L, Whitlock G, Guo Y, Collins R, China Kadoorie Biobank Collaborative Group. Physical activity and sedentary leisure time and their associations with BMI, waist circumference, and percentage body fat in 0.5 million adults: the China Kadoorie Biobank study. Am J Clin Nutr 2013 Mar;97(3):487-496 [FREE Full text] [doi: 10.3945/ajcn.112.046854] [Medline: 23364014]
25. Qin C, Guo Y, Pei P, Du H, Yang L, Chen Y, et al. The Relative Validity and Reproducibility of Food Frequency Questionnaires in the China Kadoorie Biobank Study. Nutrients 2022 Feb 14;14(4):794 [FREE Full text] [doi: 10.3390/nu14040794] [Medline: 35215443]
26. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 06;162(1):W1-73 [FREE Full text] [doi: 10.7326/M14-0698] [Medline: 25560730]
27. Lambert PC, Royston P. Further Development of Flexible Parametric Models for Survival Analysis. The Stata Journal 2009 Aug 01;9(2):265-290. [doi: 10.1177/1536867x0900900206]
28. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008 Jan 30;27(2):157-72; discussion 207. [doi: 10.1002/sim.2929] [Medline: 17569110]
29. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011 Jan 15;30(1):11-21 [FREE Full text] [doi: 10.1002/sim.4085] [Medline: 21204120]
30. Lin Y, Totsuka Y, Shan B, Wang C, Wei W, Qiao Y, et al. Esophageal cancer in high-risk areas of China: research progress and challenges. Ann Epidemiol 2017 Mar;27(3):215-221. [doi: 10.1016/j.annepidem.2016.11.004] [Medline: 28007352]
31. Gail MH, Pfeiffer RM. Breast Cancer Risk Model Requirements for Counseling, Prevention, and Screening. J Natl Cancer Inst 2018 Sep 01;110(9):994-1002 [FREE Full text] [doi: 10.1093/jnci/djy013] [Medline: 29490057]
32. Arnold M, Soerjomataram I, Ferlay J, Forman D. Global incidence of oesophageal cancer by histological subtype in 2012. Gut 2015 Mar;64(3):381-387. [doi: 10.1136/gutjnl-2014-308124] [Medline: 25320104]
33. China Kadoorie Biobank. URL: https://www.ckbiobank.org/ [accessed 2023-02-28]

Abbreviations

AUC: area under the receiver operating characteristic curve
CKB: China Kadoorie Biobank
EC: esophageal cancer
EHR: electronic health record
ESCC: esophageal squamous cell carcinoma

Edited by Y Khader; submitted 21.10.22; peer-reviewed by S He, P Han; comments to author 20.12.22; revised version received 09.01.23; accepted 03.02.23; published 15.03.23

Please cite as:
Han Y, Zhu X, Hu Y, Yu C, Guo Y, Hang D, Pang Y, Pei P, Ma H, Sun D, Yang L, Chen Y, Du H, Yu M, Chen J, Chen Z, Huo D, Jin G, Lv J, Hu Z, Shen H, Li L
Electronic Health Record–Based Absolute Risk Prediction Model for Esophageal Cancer in the Chinese Population: Model Development and External Validation
JMIR Public Health Surveill 2023;9:e43725
URL: https://publichealth.jmir.org/2023/1/e43725
doi: 10.2196/43725
PMID: 36781293