ORIGINAL RESEARCH—CLINICAL

Using the Electronic Health Record to Develop a Gastric Cancer Risk Prediction Model

Michelle Kang Kim,[1] Carol Rouphael,[1] Sarah Wehbe,[1] Ji Yoon Yoon,[2] Juan Wisnivesky,[3,4] John McMichael,[5] Nicole Welch,[1,6] Srinivasan Dasarathy,[1,6] and Emily C. Zabor[7]

[1]Department of Gastroenterology, Hepatology, and Nutrition, Cleveland Clinic, Cleveland, Ohio; [2]Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, New York; [3]Division of General Internal Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York; [4]Division of Pulmonary and Critical Care Medicine, Icahn School of Medicine at Mount Sinai, New York, New York; [5]Department of Surgery, Digestive Disease and Surgery Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio; [6]Department of Inflammation and Immunity, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio; and [7]Department of Quantitative Health Sciences, Taussig Cancer Institute, Cleveland Clinic, Cleveland, Ohio

BACKGROUND AND AIMS: Gastric cancer (GC) is a leading cause of cancer incidence and mortality globally. Population screening is limited by the low incidence and prevalence of GC in the United States. A risk prediction algorithm to identify high-risk patients allows for targeted GC screening. We aimed to determine the feasibility and performance of a logistic regression model based on electronic health records to identify individuals at high risk for noncardia gastric cancer (NCGC). **METHODS:** We included 614 patients who had a diagnosis of NCGC between ages 40 and 80 years and who were seen at our large tertiary medical center in multiple states between 2010 and 2021. Controls without a diagnosis of NCGC were randomly selected in a 1:10 ratio of cases to controls. Multiple imputation by chained equations for missing data followed by logistic regression on imputed datasets was used to estimate the probability of NCGC. Area under the curve and the 0.632 estimator was used as the estimate for discrimination. **RESULTS:** The 0.632 estimator value was 0.731, indicating robust model performance. Probability of NCGC was higher with increasing age (odds ratio [OR] = 1.16, 95% confidence interval [CI]: 1.04–1.3), male sex (OR = 1.97; 95% CI: 1.64–2.36), Black (OR = 3.07; 95% CI: 2.46–3.83) or Asian race (OR = 4.39; 95% CI: 2.60–7.42), tobacco use (OR = 1.61; 95% CI: 1.34–1.94), anemia (OR = 1.35; 95% CI: 1.09–1.68), and pernicious anemia (OR = 6.12, 95% CI: 3.42–10.95). **CONCLUSION:** We demonstrate the feasibility and good performance of an electronic health record–based logistic regression model for estimating the probability of NCGC. Future studies will refine and validate this model, ultimately identifying a high-risk cohort who could be eligible for NCGC screening.

Keywords: Noncardia Gastric Cancer; Electronic Health Record; Logistic Regression Model; Screening; High-Risk Individuals; Cancer Disparity

Introduction

Gastric cancer (GC) is the fifth most commonly diagnosed cancer and fourth leading cause of cancer-related death in the world, with approximately 770,000 deaths in 2020.[1] It is frequently diagnosed at an advanced stage, with an overall 5-year survival rate of 31%.[2,3] While GC is aggressive and associated with poor outcomes when diagnosed late, early GC can be cured with minimally invasive endoscopic resection.[4,5] In countries where noncardia gastric cancer (NCGC) has a high incidence and prevalence, nationwide screening has been adopted resulting in earlier detection and improved overall outcomes.[6] In the United States, a lower incidence country, there are no current screening protocols for NCGC in the average population because of limited feasibility and cost-effectiveness.[7]

Despite lower incidence in the United States, NCGC remains a persistent source of cancer disparity, disproportionately affecting minority race-ethnic groups and those living in poverty.[8] Identifying high-risk populations who may benefit from screening is thus paramount. Known risk factors for NCGC include older age, sex, race, ethnicity, family history, smoking, and chronic *Helicobacter pylori* (HP) infection, variables readily available in the electronic health record (EHR).[9–13]

With adoption in the United States exceeding 81%, the EHR is a large resource of clinically relevant, real-world longitudinal data,[14,15] and has emerged as a promising data source for the identification of high-risk individuals in rare diseases and for the development of risk prediction models for common medical conditions.[16,17] EHR-based

Abbreviations used in this paper: AUC, area under the receiver operator curve; BMI, body mass index; CC, Cleveland Clinic; EGD, esophagogastroduodenoscopy; EHR, electronic health record; HP, Helicobacter pylori; ICD, International Classification of Diseases; IQR, interquartile range; MICE, multiple imputation with chain equation; NCGC, noncardia gastric cancer; PPV, positive predictive value.

Most current article

2772-5723
https://doi.org/10.1016/j.gastha.2024.07.001

models have also been developed for various cancers including cancer of the breast, esophagus, and pancreatic cancer.[18–20] Only one US-based study developed an EHR-based model to predict GC, but was limited to patients who had undergone esophagogastroduodenoscopy (EGD).[21]

In this study, we developed and assessed the performance of an EHR-based logistic regression predictive model in accurately identifying individuals at high risk for NCGC.

Methods

Data Sources, Case, and Control Selection

We conducted a retrospective case-control study using the Cleveland Clinic (CC) EHR database. The CC is one of the largest healthcare systems in the world with facilities located in multiple states including Ohio and Florida. The CC EHR includes prospectively collected data for around 12 million patients and has been available since 2010.

We identified individuals receiving care at CC between 2010 and 2021 who had a diagnosis of NCGC between the ages of 40 and 80 years through query of the EHR using International Classification of Diseases (ICD) 9 and 10 codes (C16.1–C16.9). A manual chart review was then performed to confirm accurate identification of intestinal type NCGC (adenocarcinoma); we excluded individuals with gastric neuroendocrine tumors, gastrointestinal stromal tumors, hereditary cancer syndromes, signet ring carcinoma, or linitis plastica (Figure). We also excluded cancers of the cardia as they share similar epidemiological characteristics with esophageal adenocarcinoma in the United States. Only patients with a duration of medical records of 12 months or more prior to the NCGC diagnosis were included. We randomly selected controls in a 1:10 ratio of cases to controls from approximately 4.5 million patients aged 40–80 years without a diagnosis of NCGC, who were evaluated during the same timeframe. For the NCGC cohort, age was defined at time of diagnosis; age for controls was defined at time of last encounter in the EHR.

Model Variables

Sociodemographic and clinical data available in the EHR were included in the model, including age, sex, race, ethnicity, body mass index, and tobacco history. Clinical data as obtained by ICD-9 and ICD-10 codes included clinical characteristics and comorbidities and were obtained at a time prior to the NCGC diagnosis. We included only the variables with a high degree of presence in the EHR. Features available from EGD or pathology records were not included as the objective was to identify high-risk patients from EHR features without prior endoscopy, applicable to a broader population.

Statistical Methods

Patient characteristics were summarized using the number and percentage for binary or categorical variables and median and first and third quartiles for continuous variables, according to case or control status. Missing values in covariates included in the model were assumed to be missing at random and were

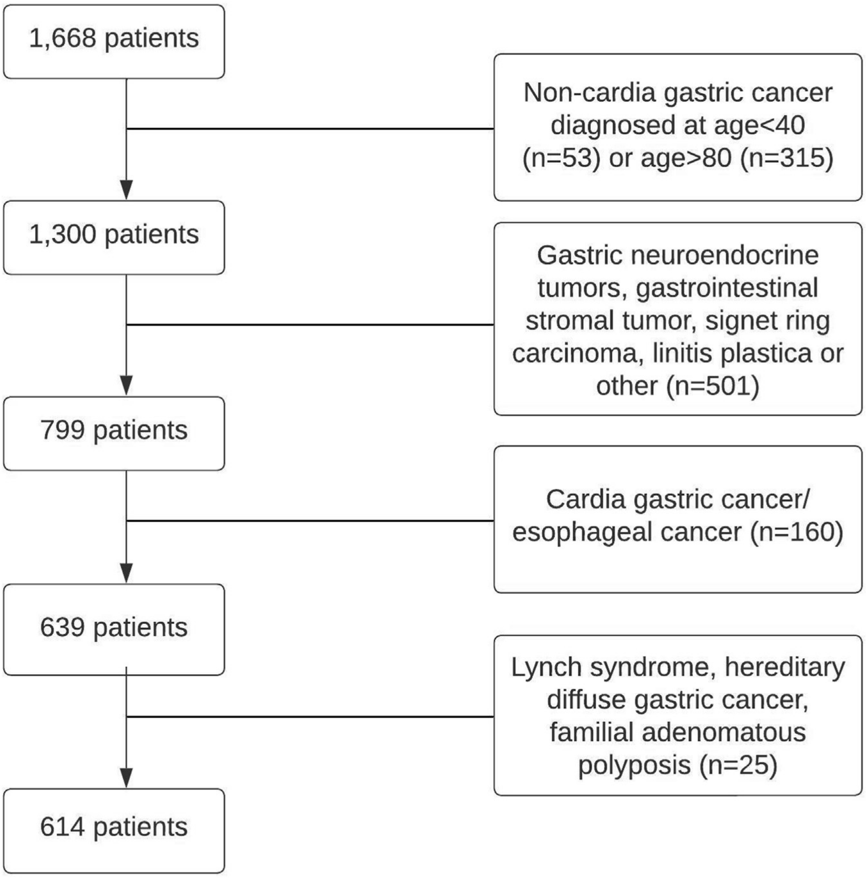
1,668 patients
Non-cardia gastric cancer diagnosed at age<40 (n=53) or age>80 (n=315)
1,300 patients
Gastric neuroendocrine tumors, gastrointestinal stromal tumor, signet ring carcinoma, linitis plastica or other (n=501)
799 patients
Cardia gastric cancer/ esophageal cancer (n=160)
639 patients
Lynch syndrome, hereditary diffuse gastric cancer, familial adenomatous polyposis (n=25)
614 patients

Figure. Flow diagram of case inclusion and exclusion.

imputed using multiple imputation by chained equations.[22] Ten datasets were imputed based on convergence to an absolute difference in predicted risk of $< 3\%$ from 2 random seeds, and only results from the first seed were retained for analysis. The imputation model conditioned on all predictors as well as case vs control status. The logistic regression model was fit to each of the imputed datasets. Final estimates of the regression coefficients and their standard errors are obtained using Rubin's approach[23] and estimates of model performance using the guidance of Marshall et al.[24]

Discrimination, which is a measure of how accurately patients are classified as having NCGC or not, was assessed using the area under the receiver operator curve (AUC). The apparent AUC was obtained for each multiple imputed dataset. Next, 500 bootstrap samples were taken of each dataset to obtain the bootstrap cross-validation AUC. The apparent and bootstrap cross-validation AUCs were combined into a 0.632 estimator of discrimination. The 0.632 estimator is considered the best estimate of model performance, with the bootstrap cross-validated estimator taken as a minimum and the apparent estimator taken as a maximum, to give a range of model performance.[25] Calibration was assessed visually using plots of the predicted risk against the observed proportion, as calculated based on subjects in a nearest neighborhood of the given predicted risk.

To identify an optimal threshold to be used to select patients for screening, we maximized the sensitivity for a fixed value of specificity based on the average predicted probability of GC across the multiple imputed datasets. Positive predictive value (PPV) was also calculated for different threshold points. All statistical analyses were conducted using R software version 4.3, using packages such as multiple imputation by chained equations, riskRegression, and cutpointr.

Results

Cohort Characteristics

Six hundred fourteen patients with NCGC and 6331 controls were identified based on prespecified inclusion criteria (Figure). Baseline characteristics of both NCGC cases and normal controls are listed in Table 1. The median age was 68 years (interquartile range [IQR] 59–74) for cases and 66 years (IQR 60–73) for controls. Male sex constituted a larger proportion in NCGC cases (58%) and was less common in the controls (42%). The majority of patients included in both cases and control groups were of White race (65% and 83%, respectively); however, there were more Black and Asian NCGC cases (28% and 3.6%, respectively) as compared to controls (14% and 1.3%, respectively). On average, approximately 6% of both NCGC cases and controls were Hispanic.

Body mass index was comparable between the 2 groups with a median of 28 kg/m^2 (IQR 24–32) for the NCGC cohort and 29 kg/m^2 (IQR 25–33) for the control cohort. The NCGC cohort had a higher proportion of ever smoking (61%) compared to the controls (48%). Dyspepsia, heartburn, and reflux occurred less frequently in the NCGC compared to the control cohort. Similarly, comorbidities including gastroesophageal reflux disease, diabetes, hypertension, hypercholesterolemia, viral hepatitis, and liver disease were all less commonly present in NCGC patients. Anemia was comparable between both groups; pernicious anemia was more common in NCGC cases.

Variable Associations

Results of a multivariable logistic regression analysis on the multiple imputed dataset are shown in Table 2. The strongest predictors of NCGC were age, male sex, Black or Asian race, ever tobacco use, anemia, and pernicious anemia. Age (per 10-year increase) was significantly associated with increased odds of NCGC (odds ratio [OR] = 1.16; 95% confidence interval [CI]: 1.04–1.30). Patients who were male (OR = 1.97; 95% CI: 1.64–2.36), Black (OR = 3.07; 95% CI: 2.46–3.83), or Asian (OR = 4.39; 95% CI: 2.60–7.42) had increased odds of NCGC. Furthermore, patients who had ever used tobacco (OR = 1.61; 95% CI: 1.34–1.94), had anemia (OR = 1.35; 95% CI: 1.09–1.68), or had pernicious anemia (OR = 6.12; 95% CI: 3.42–10.95) had increased odds of NCGC. In contrast, patients with hypertension (OR = 0.7; 95% CI: 0.56–0.87), hypercholesterolemia (OR = 0.38; 95% CI: 0.30–0.47), or liver disease (OR = 0.53; 95% CI: 0.33–0.85) had decreased odds of NCGC.

Model Performance Characteristics

The median apparent AUC across the 10 multiple imputed datasets was 0.74. The median bootstrap cross-validation AUC across the same datasets was 0.725. The median 0.632 estimator across the 10 multiple imputed datasets, which was taken as our best estimate of the AUC, was 0.731. In addition, the apparent calibration used was robust across every multiple imputed dataset (Figure A1).

Model performance characteristics including sensitivity, specificity, and PPV at different threshold points are shown in Table 3. Adjusting the threshold affected the balance between sensitivity and specificity. Setting a higher threshold lowered sensitivity while improving specificity and PPV, and vice versa. For a threshold of 0.028, sensitivity was 95.6% with a specificity of 20.1% and a PPV of 0.30%. As threshold increased to 0.172, sensitivity decreased to 39.6%, while specificity increased to 90.2% and PPV to 1.0%, approaching the desired value for screening tests.

Discussion

Using a real-world, multistate, large clinical dataset, we developed a multivariable logistic regression model to predict NCGC risk. With a relatively simple model including demographic, behavioral, and clinical features, and without the need for a prior endoscopy, we were able to predict NCGC with good accuracy (AUC 0.73). Importantly, using simple and readily available variables extracted from the EHR, at 0.172 threshold, the model reached a specificity of 90% and PPV of 1%, demonstrating the potential for providing individual-level risk prediction for screening programs.

Table 1. Clinical Characteristics of Cases and Controls

Characteristic	Case, N = 614[a]	Control, N = 6331[a]
Age (median, IQR)	68 (59, 74)	66 (60, 73)
Sex		
Female	256 (42%)	3653 (58%)
Male	358 (58%)	2678 (42%)
Race		
White	383 (65%)	5048 (83%)
Black	165 (28%)	844 (14%)
Asian	21 (3.6%)	80 (1.3%)
Other	22 (3.7%)	123 (2.0%)
Unknown	23	236
Ethnicity		
Not Hispanic	519 (93%)	5512 (95%)
Hispanic	38 (6.8%)	295 (5.1%)
Unknown	57	524
Average body mass index (BMI) (median, IQR)	28 (24, 32)	29 (25, 33)
Unknown	166	275
Smoking status		
Never	235 (39%)	3183 (52%)
Ever	360 (61%)	2925 (48%)
Unknown	19	223
Dyspepsia	20 (3.3%)	322 (5.1%)
Heartburn	9 (1.5%)	191 (3.0%)
Reflux	107 (17%)	2108 (33%)
Bloating	34 (5.5%)	559 (8.8%)
Anemia	156 (25%)	1617 (26%)
Pernicious anemia	19 (3.1%)	55 (0.9%)
Combined variable immunodeficiency	1 (0.2%)	11 (0.2%)
Gastroesophageal reflux disease	113 (18%)	2210 (35%)
Diabetes	127 (21%)	1631 (26%)
Hypertension	269 (44%)	3709 (59%)
Hypercholesterolemia	237 (39%)	4078 (64%)
Coronary artery disease	99 (16%)	1147 (18%)
Viral hepatitis	12 (2.0%)	146 (2.3%)
Liver disease	21 (3.4%)	482 (7.6%)

IQR, interquartile range.
[a]n (%).

Our results demonstrate the feasibility of an EHR-based logistic regression model for efficient NCGC prediction in general clinical practice and are a proof of concept of a low-cost and efficient methodology to identify patients at highest risk that may benefit from cancer mitigation strategies such as endoscopic screening. Although previous studies have developed risk prediction models for gastric intestinal metaplasia, no previous study has developed an EHR-based model that can be applied to a general US adult population to identify high-risk candidates appropriate for NCGC screening.[26,27] The majority of the published risk prediction models are from Asia, with more homogenous populations compared to the United States.[28,29,30] In addition, many of these models included biomarker (eg, pepsinogen) and pathology information from endoscopies (eg, intestinal metaplasia), in addition to established risk factors, limiting generalizability to the United States, where biomarkers are not readily available, and mass EGD screening is not performed. Huang et al. developed a GC risk prediction model using United States–based data. However, these models included endoscopic and pathologic variables, limiting its ability to be used in the population at large.[21]

While the United States is a lower-incidence country for GC, outcomes are especially poor, contrasting sharply high-incidence countries in the East (Japan, Korea), which have implemented screening programs. In the United States, GC encompasses a major disparity in cancer incidence and mortality. GC incidence is approximately double in non-White groups, and certain Asian subgroups demonstrate up to 14-fold risk compared to non-Hispanic White groups.[31] Indeed, GC is the foremost and second-leading cancer for disparity in mortality in Hispanic and Black groups, respectively, and these disparities have persisted.[8,32] Greater attention to GC mitigation is needed in light of these disparities and projected growth in high-risk, vulnerable populations.[33] While endoscopic screening in

Table 2. Multivariable Model

Feature	OR	LCI	UCI	*P* value
Age	1.16	1.04	1.30	.008
Male sex	1.97	1.64	2.36	<.0001
Race (Black)	3.07	2.46	3.83	<.0001
Race (Asian)	4.39	2.60	7.42	<.0001
Race (other)	2.10	1.23	3.58	.007
Ethnicity (Hispanic)	1.48	0.99	2.22	.059
Average body mass index (BMI)	0.93	0.85	1.01	.074
Ever smoker	1.61	1.34	1.94	<.0001
Dyspepsia	0.89	0.54	1.46	.653
Heartburn	0.79	0.39	1.60	.514
Reflux	1.00	0.41	2.40	.995
Bloating	0.94	0.63	1.38	.736
Anemia	1.35	1.09	1.68	.007
Pernicious anemia	6.12	3.42	10.95	<.0001
Combined variable immunodeficiency	1.13	0.14	9.25	.909
Gastroesophageal reflux disease	0.58	0.25	1.38	.220
Diabetes	1.01	0.80	1.29	.927
Hypertension	0.70	0.56	0.87	.001
Hypercholesterolemia	0.38	0.30	0.47	<.0001
Coronary artery disease	1.17	0.90	1.51	.240
Viral hepatitis	0.77	0.41	1.47	.433
Liver disease	0.53	0.33	0.85	.009

Pooled results from logistic regression on multiply imputed data.
LCI, lower confidence interval; OR, odds ratio; UCI, upper confidence interval.

Table 3. Impact of Threshold Values on Model Sensitivity, Specificity, and Positive Predictive Value

Threshold	Sensitivity (in %)	Specificity (in %)	Positive predictive value (in %)
0.028	95.6	20.1	0.30
0.037	92.2	30.0	0.33
0.060	80.6	50.0	0.40
0.121	53.4	80.0	0.67
0.172	39.6	90.2	1.00

the United States is unfeasible and of questionable benefit for the general population, targeting of high-risk groups is supported by modeling studies.[34,35]

Our findings regarding demographic and behavioral characteristics as well as medical diagnoses align with risk factors identified from existing risk prediction models from several countries.[28–30,36] Increasing age and male sex are well-established risk factors for NCGC, as are Black and Asian race.[8,37] Between 2010 and 2014 in the United States, Asians/Pacific Islanders males and Black individuals exhibited the highest rates of GC incidence when compared to other racial and gender groups.[38] Individuals with a history of tobacco use had increased odds of NCGC, consistent with prior reports of elevated risk for NCGC across various ethnic groups,[39] and tend to be higher among those with prolonged usage and higher consumption of tobacco.[11] In addition, patients with anemia or pernicious anemia had increased odds of NCGC. Retrospective studies report the presence of anemia and pernicious anemia in NCGC patients,[40] and prospective studies[41] as well as systematic reviews[42] show higher incidence rates of GC in people with pernicious anemia compared to the general population.

Our study has multiple strengths. Our approach of using the CC EHR leverages the immense longitudinal data available in a large EHR encompassing data of diverse populations from diverse geographic locations. We confirmed the accuracy of the diagnosis of NCGC with multiple measures including cross-referencing of EGD reports and manual chart review. We limited tertiary referral center bias by including only those patients with $>$ 12 months of data. In addition, our study included only NCGC, which is significant as cardia cancer is widely acknowledged to be biologically similar to esophageal cancer.[43] The comprehensive EHR data used in our analyses have the potential beyond standard clinical characteristics for more detailed risk stratification at point of care. For instance, individuals living in poverty and immigrants from high-incidence countries are additional important subgroups at risk identified by epidemiological studies that would not easily be incorporated into conventional risk stratification algorithms.[44,45,32]

There are several limitations to our study. In any EHR, missing data are a frequent and important challenge to overcome.[46] To overcome this, we included only those variables with a high degree of presence; we also used multiple imputation and sensitivity analyses to confirm the robustness of our results. These techniques were able to account for missing data in multiple variables including smoking. Clinical variables were considered present if their corresponding ICD-9/ICD-10 codes were documented in their patients' records, and the absence of the ICD-9/ICD-10 codes was interpreted as absence of the respective condition. However, we were unable to assess for 2 important factors: HP status and alcohol use, both established risk factors associated with increased risk of developing NCGC, but poorly reported in the EHR.[47–50] As our aim was to develop an EHR risk prediction model that can be applied to the general population, the solid performance of the model despite the lack of HP data ensures in fact that the model can be more widely applied. Another limitation is that we did not conduct external validation of our novel EHR-based model in NCGC screening.[51]

Conclusion

In summary, we demonstrate the feasibility of an EHR-based logistic regression model in accurately predicting the probability of NCGC. Our model included easily attainable demographic, behavioral, and medical history using ICD codes. While different thresholds of the model achieved a range of sensitivity and specificity, the most specific model was able to achieve a PPV approaching 1%. Future studies are needed to validate and refine this model.

Supplementary Materials

Material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gastha.2024.07.001.

References

1. Morgan E, Arnold M, Camargo MC, et al. The current and future incidence and mortality of gastric cancer in 185 countries, 2020-40: a population-based modelling study. EClinicalMedicine 2022;47:101404.
2. Florea A, Brown HE, Harris RB, et al. Ethnic disparities in gastric cancer presentation and screening practice in the United States: analysis of 1997-2010 surveillance, epidemiology, and end results-medicare data. Cancer Epidemiol Biomarkers Prev 2019;28(4):659–665.
3. Surveillance, Epidemiology, and End Results Program. Cancer query system: SEER incidence statistics. Bethesda: National Cancer Institute, 2023.
4. Uedo N, Iishi H, Tatsuta M, et al. Longterm outcomes after endoscopic mucosal resection for early gastric cancer. Gastric Cancer 2006;9(2):88–92.
5. Nishizawa T, Yahagi N. Long-term outcomes of using endoscopic submucosal dissection to treat early gastric cancer. Gut Liver 2018;12(2):119–124.
6. Jun JK, Choi KS, Lee HY, et al. Effectiveness of the Korean national cancer screening program in reducing gastric cancer mortality. Gastroenterology 2017;152(6):1319–1328.e7.
7. Xia JY, Aadam AA. Advances in screening and detection of gastric cancer. J Surg Oncol 2022;125(7):1104–1109.
8. CancerDisparitiesProgressReport.org. American Association for Cancer Research. 2022. http://www.CancerDisparitiesProgressReport.org. Accessed January 3, 2024.
9. Kim GH, Liang PS, Bang SJ, et al. Screening and surveillance for gastric cancer in the United States: is it needed? Gastrointest Endosc 2016;84(1):18–28.
10. Helicobacter, Cancer Collaborative G. Gastric cancer and helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts. Gut 2001;49(3):347–353.
11. Ladeiras-Lopes R, Pereira AK, Nogueira A, et al. Smoking and gastric cancer: systematic review and meta-analysis of cohort studies. Cancer Causes Control 2008;19(7):689–701.
12. Shin CM, Kim N, Yang HJ, et al. Stomach cancer risk in gastric cancer relatives: interaction between helicobacter pylori infection and family history of gastric cancer for the risk of stomach cancer. J Clin Gastroenterol 2010;44(2):e34–e39.
13. Brown LM, Devesa SS. Epidemiologic trends in esophageal and gastric cancer in the United States. Surg Oncol Clin N Am 2002;11(2):235–256.
14. Jiang J, Qi K, Bai G, et al. Pre-pandemic assessment: a decade of progress in electronic health record adoption among U.S. hospitals. Health Aff Sch 2023;1(5):qxad056.
15. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis 2023;82(3):306–311.
16. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(1):198–208.
17. Houssein EH, Mohamed RE, Ali AA. Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. Sci Rep 2023;13(1):7173.
18. Wang H, Li Y, Khan SA, et al. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. Artif Intell Med 2020;110:101977.
19. Iyer PG, Sachdeva K, Leggett CL, et al. Development of electronic health record-based machine learning models to predict Barrett's esophagus and esophageal adenocarcinoma risk. Clin Transl Gastroenterol 2023;14(10):e00637.
20. Placido D, Yuan B, Hjaltelin JX, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med 2023;29(5):1113–1122.
21. Huang RJ, Kwon NS, Tomizawa Y, et al. A Comparison of logistic regression against machine learning algorithms for gastric cancer risk prediction within real-world clinical data Streams. JCO Clin Cancer Inform 2022;6:e2200039.
22. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16(3):219–242.
23. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91(434):473–489.
24. Marshall A, Altman DG, Holder RL, et al. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol 2009;9:57.
25. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997;92(438):548–560.
26. Tan MC, Ho Q, Nguyen TH, et al. Risk score using demographic and clinical risk factors predicts gastric intestinal metaplasia risk in a U.S. Population. Dig Dis Sci 2022;67(9):4500–4508.
27. Tan MC, Sen A, Kligman E, et al. Validation of a pre-endoscopy risk score for predicting the presence of gastric intestinal metaplasia in a U.S. population. Gastrointest Endosc 2023;98(4):569–576.e1.
28. Eom BW, Joo J, Kim S, et al. Prediction model for gastric cancer incidence in Korean population. PLoS One 2015;10(7):e0132613.
29. Iida M, Ikeda F, Hata J, et al. Development and validation of a risk assessment tool for gastric cancer in a general Japanese population. Gastric Cancer 2018;21(3):383–390.
30. Taninaga J, Nishiyama Y, Fujibayashi K, et al. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: a case-control study. Sci Rep 2019;9(1):12384.

31. Shah SC, McKinley M, Gupta S, et al. Population-based analysis of differences in gastric cancer incidence among races and ethnicities in individuals age 50 years and older. Gastroenterology 2020;159(5):1705–1714.e2.
32. The burden of stomach cancer mortality by county, race, and ethnicity in the USA, 2000-2019: a systematic analysis of health disparities. Lancet Reg Health Am 2023;24:100547.
33. Vespa J, Armstrong DM, Medina L. Demographic turning points for the United States: population projections for 2020 to 2060. Washington: U.S. Census Bureau, 2020.
34. Shah SC, Canakis A, Peek RM Jr, et al. Endoscopy for gastric cancer screening is cost effective for Asian Americans in the United States. Clin Gastroenterol Hepatol 2020;18(13):3026–3039.
35. Saumoy M, Schneider Y, Shen N, et al. Cost effectiveness of gastric cancer screening according to race and ethnicity. Gastroenterology 2018;155(3):648–660.
36. In H, Solsky I, Castle PE, et al. Utilizing cultural and ethnic variables in screening models to identify individuals at high risk for gastric cancer: a pilot study. Cancer Prev Res (Phila) 2020;13(8):687–698.
37. Gu J, Chen R, Wang SM, et al. Prediction models for gastric cancer risk in the general population: a systematic review. Cancer Prev Res (Phila) 2022;15(5):309–318.
38. Ashktorab H, Kupfer SS, Brim H, et al. Racial disparity in gastrointestinal cancer risk. Gastroenterology 2017;153(4):910–923.
39. Nomura AM, Wilkens LR, Henderson BE, et al. The association of cigarette smoking with gastric cancer: the multiethnic cohort study. Cancer Causes Control 2012;23(1):51–58.
40. Tang GH, Hart R, Sholzberg M, et al. Iron deficiency anemia in gastric cancer: a Canadian retrospective review. Eur J Gastroenterol Hepatol 2018;30(12):1497–1501.
41. Hsing AW, Hansson LE, McLaughlin JK, et al. Pernicious anemia and subsequent cancer. A population-based cohort study. Cancer 1993;71(3):745–750.
42. Vannella L, Lahner E, Osborn J, et al. Systematic review: gastric cancer incidence in pernicious anaemia. Aliment Pharmacol Ther 2013;37(4):375–382.
43. Hayakawa Y, Sethi N, Sepulveda AR, et al. Oesophageal adenocarcinoma and gastric cancer: should we mind the gap? Nat Rev Cancer 2016;16(5):305–318.
44. Pabla BS, Shah SC, Corral JE, et al. Increased incidence and mortality of gastric cancer in immigrant populations from high to low regions of incidence: a systematic review and meta-analysis. Clin Gastroenterol Hepatol 2020;18(2):347–359.e5.
45. Laszkowska M, Zhang X, Kuliszewski MG, et al. Heightened risk for gastric cancer among immigrant populations in New York state from high-incidence countries. Clin Gastroenterol Hepatol 2023;21(10):2673–2675.e3.
46. Kim MK, Rouphael C, McMichael J, et al. Challenges in and opportunities for electronic health record-based data analysis and interpretation. Gut Liver 2024;18(2):201–208.
47. Wroblewski LE, Peek RM Jr, Wilson KT. Helicobacter pylori and gastric cancer: factors that modulate disease risk. Clin Microbiol Rev 2010;23(4):713–739.
48. Yang L, Kartsonaki C, Yao P, et al. The relative and attributable risks of cardia and non-cardia gastric cancer associated with helicobacter pylori infection in China: a case-cohort study. Lancet Public Health 2021;6(12):e888–e896.
49. Ishaq S, Nunn L. Helicobacter pylori and gastric cancer: a state of the art review. Gastroenterol Hepatol Bed Bench 2015;8(Suppl 1):S6–S14.
50. Yoo JE, Shin DW, Han K, et al. Association of the frequency and quantity of alcohol consumption with gastrointestinal cancer. JAMA Netw Open 2021;4(8):e2120382.
51. Thrift AP, Kanwal F, El-Serag HB. Prediction models for gastrointestinal and liver diseases: too many developed, too few validated. Clin Gastroenterol Hepatol 2016;14(12):1678–1680.

Received March 28, 2024. Accepted July 3, 2024.

Correspondence:
Address correspondence to: Michelle Kang Kim, MD, PhD, Cleveland Clinic, 9500 Euclid Avenue, A30, Cleveland, Ohio 44195. e-mail: kimm13@ccf.org.

Authors' Contributions:
Michelle Kang Kim: Conception and design of the study, interpretation of data, drafting and revision of the manuscript, and approval of final version of the manuscript. Carol Rouphael: Assembly and interpretation of data, drafting and revision of manuscript, and approval of the final version of the manuscript. Sarah Wehbe: Assembly and interpretation of data, drafting and revision of manuscript, and approval of final version of the manuscript. Ji Yoon Yoon: Drafting and revision of manuscript. Juan Wisnivesky: Drafting and revision of manuscript. John McMichael: Generation, collection, and assembly of the data. Nicole Welch: Drafting and revision of manuscript. Srinivasan Dasarathy: Conception and design of the study, drafting and revision of manuscript, and approval of the final version of the manuscript. Emily C. Zabor: Conception and design of the study, analysis and interpretation of data, and drafting and revision of the manuscript.

Conflicts of Interest:
These authors disclose the following: Juan Wisnivesky consults for PPD, Banook, and Sanofi and received grants from Sanofi, Regeneron, and Axella. Nicole Welch is partially supported by NIH K08 AA028794. Srinivasan Dasarathy is partially supported by R01 GM119174; R01 DK113196; P50 AA024333; R01 AA021890; 3U01AA026976 – 03S1; U01 AA 026976; R56HL141744; U01 DK061732. The remaining authors disclose no conflicts.

Funding:
The authors report no funding.

Ethical Statement:
This research received Institutional Review Board approval #23–151 at the Cleveland Clinic.

Data Transparency Statement:
Data and study materials can be made available upon specific request.

Reporting Guidelines:
STROBE and SAGER guidelines.