Development of Electronic Health Record-Based Machine Learning Models to Predict Barrett's Esophagus and Esophageal Adenocarcinoma Risk

Prasad G. Iyer, MD, MSc[1], Karan Sachdeva, MD[1], Cadman L. Leggett, MD[1], D. Chamil Codipilly, MD[1], Halim Abbas, PhD[2], Kevin Anderson, PhD[2], John B. Kisiel, MD, MSc[1], Shahir Asfahan, MD[3], Samir Awasthi, MD, PhD[3], Praveen Anand, PhD[3], Praveen Kumar M, MD[3], Shiv Pratap Singh, BE[3], Sharad Shukla, BE[3], Sairam Bade, BE[3], Chandan Mahto, BE[3], Navjeet Singh, BE[3], Saurav Yadav, BE[3] and Chinmay Padhye, MS[3]

INTRODUCTION: **Screening for Barrett's esophagus (BE) is suggested in those with risk factors, but remains underutilized. BE/esophageal adenocarcinoma (EAC) risk prediction tools integrating multiple risk factors have been described. However, accuracy remains modest (area under the receiver-operating curve [AUROC] ≤0.7), and clinical implementation has been challenging. We aimed to develop machine learning (ML) BE/EAC risk prediction models from an electronic health record (EHR) database.**

METHODS: **The Clinical Data Analytics Platform, a deidentified EHR database of 6 million Mayo Clinic patients, was used to predict BE and EAC risk. BE and EAC cases and controls were identified using International Classification of Diseases codes and augmented curation (natural language processing) techniques applied to clinical, endoscopy, laboratory, and pathology notes. Cases were propensity score matched to 5 independent randomly selected control groups. An ensemble transformer-based ML model architecture was used to develop predictive models.**

RESULTS: **We identified 8,476 BE cases, 1,539 EAC cases, and 252,276 controls. The BE ML transformer model had an overall sensitivity, specificity, and AUROC of 76%, 76%, and 0.84, respectively. The EAC ML transformer model had an overall sensitivity, specificity, and AUROC of 84%, 70%, and 0.84, respectively. Predictors of BE and EAC included conventional risk factors and additional novel factors, such as coronary artery disease, serum triglycerides, and electrolytes.**

DISCUSSION: **ML models developed on an EHR database can predict incident BE and EAC risk with improved accuracy compared with conventional risk factor-based risk scores. Such a model may enable effective implementation of a minimally invasive screening technology.**

KEYWORDS: artificial intelligence; prediction; algorithm; esophageal cancer

SUPPLEMENTARY MATERIAL accompanies this paper at http://links.lww.com/CTG/B14

Clinical and Translational Gastroenterology 2023;14:e00637. https://doi.org/10.14309/ctg.0000000000000637

INTRODUCTION

Esophageal adenocarcinoma (EAC) is a lethal malignancy when diagnosed after the onset of obstructive symptoms, with a 6-fold increase in incidence over the past 40 years in the West. Barrett's esophagus (BE), a metaplastic change of the distal esophageal epithelium from squamous to specialized intestinal epithelium due to chronic reflux-induced injury, is the precursor of most EACs (1) Hence, screening for BE is suggested by most gastroenterology societies (2–4) to allow identification of those at risk, followed by endoscopic surveillance and endoscopic treatment of dysplasia. This approach is cost-effective in reducing EAC incidence (5).

Despite these recommendations, BE screening rates continue to be low. Most cases of prevalent BE remains undetected, with most screening-eligible patients not being screened despite having multiple encounters with healthcare providers (6,7). Several

[1]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota, USA; [2]Center for Digital Health, Mayo Clinic, Rochester, Minnesota, USA; [3]Nference, Cambridge, Massachusetts, USA. **Correspondence:** Prasad G. Iyer, MD, MSc. E-mail: iyer.prasad@mayo.edu.

Received April 19, 2023; accepted September 1, 2023; published online September 12, 2023

© 2023 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of The American College of Gastroenterology

potential reasons exist for this phenomenon, including the invasive and expensive nature of endoscopy, challenging access to endoscopy in the community; lack of provider awareness of guidelines; and competition from other guideline-endorsed recommended cancer (colon, breast, and cervical) screening in primary care clinical encounters (8). Another challenge is that unlike colon or breast cancer screening, which are largely age-based and universal, BE screening is suggested only in those with multiple risk factors (age older than 50 years, male sex, White race, chronic reflux, obesity [central], ever smoker, and family history of BE or EAC). This difference is important because it places the responsibility of enumerating and integrating these risk factors on the provider, assessing whether the patient meets screening criteria and then taking the decision to perform screening. Although risk scores that integrate several risk factors into a single numerical value, which can be used to assess BE or EAC risk, have been proposed (9–13), they are limited by their use of variables from surveys, anthropometric measures not routinely measured (such as waist-hip ratio), and their modest accuracy (area under the curve [AUC] 0.65–0.70). We have previously demonstrated low sensitivity (45%) of the current society BE screening guidelines in detecting EAC in patients from our institution (14).

Hence, there exists the need to develop a BE/EAC risk prediction model, which is more accurate, automatically incorporates data points that are available in the electronic health record (EHR), and expands the current risk factor pool (which seems to have a ceiling for accuracy). Such a tool should ideally be developed from and be integrable into the EHR, to provide an electronic trigger to the provider, prompting consideration of screening (15). We aimed to develop an artificial intelligence-powered machine learning (ML) BE/EAC risk prediction algorithm, which determines BE/EAC risk at least a year before diagnosis, from a large deidentified EHR database.

METHODS

Data source and disease groups: overview

The Mayo Clinic (Rochester, MN)-nference (Cambridge, MA) Clinical Data Analytics Platform (CDAP) consists of approximately 6 million deidentified patient records of patients seen at all Mayo Clinic campuses (in Minnesota, Arizona, and Florida) from the 1930s to 2021. These records were computationally screened to identify patients who were diagnosed with BE and EAC. This involved a mix of natural language processing (NLP) algorithms running on patient notes and structured EHR data, as described below. Patients diagnosed with BE and EAC from CDAP formed the 2 disease-positive cohorts. A randomly selected set of patients who did not have BE/EAC and who were propensity matched with the disease-positive (case) cohorts (see criteria below) formed the control cohort. Two predictive models, 1 for each disease (BE and EAC), were developed using these cohorts. The models were built to predict the probability that a patient would develop the disease at least 1 year before diagnosis. This was achieved by including only patient data between 1 and 5 years before the diagnosis of BE or EAC (the observation period) for model development, allowing for minimization of protopathic bias.

Identification of disease-positive cohorts

Figure 1 depicts the method for identification of the 2 disease cohorts (BE and EAC). Four criteria were used to identify patients included in the BE or EAC disease cohorts. These included (i) diagnosis codes, (ii) endoscopy procedure codes, (iii) augmented curation (a NLP tool), and (iv) the presence of specific keywords in the pathology notes.

Diagnosis codes. The *International Classification of Diseases* (*ICD-9* and *ICD-10*), Systematized Nomenclature of Medicine, and Hospital Adaptation of the International Classification of Diseases codes were used to identify the diseased cohorts, and the same codes were used to exclude any cases from the control cohort. Diagnosis codes used for case identification are listed in the Supplementary Appendix (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14).

Procedure codes. The second criterion used was the presence of an endoscopy procedure code, preceding the diagnosis of BE or EAC. Only procedures that were performed within a year of the earliest and latest diagnosis dates were considered. The procedure codes used are listed in the Supplementary Data (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14).

Augmented curation. Patient notes were processed using models to check whether they were diagnosed with BE or EAC. This was a 3-step process, which involved

1. identifying synonyms for the disease,
2. getting relevant sentences from the patient notes that mentioned the disease of interest or the synonyms, and
3. using a model to check whether these sentences indicate that the patient had the disease of interest.

Various databases, such as MESH terms, DOID, MONDO Ontology, and Wikidata, were used to identify known synonyms of BE and EAC in the literature. In addition to these tools, manual reading of clinical notes and domain knowledge was also performed to come up with terms that identify BE and EAC. For example, "esophageal adenocarcinoma" and "adenocarcinoma of the esophagus" would both be the synonyms for EAC. Getting the relevant sentences involved the usage of specialized tools (16,17) to identify sentences from the patient notes that had mention of the disease or its synonyms. We then used a NLP classification model, Bidirectional Encoder Representations from Transformer (18), that was trained with more than 15,000 sentences to determine whether a sentence indicated that the patient had a disease. The precision and recall of the model have been reported to be 0.97 and 0.98, respectively. The sentences identified from the patient notes were then processed through this model to check whether the patient was diagnosed with BE or EAC.

Keywords in pathology notes. The process of identifying disease-positive patients also required that these patients have, in their pathology notes, certain terms related to the disease. For example, the word "adenocarcinoma" along with one of the following other terms—esophageal or esophagus or esophagus or oesophageal—was deemed necessary to confirm a diagnosis of EAC. This was performed to ensure that the fidelity of anatomical location and pathology was maintained. The full list of pathology note terms that were used is provided in the Supplementary Data (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14).

Identification of the control cohort and propensity matching to cases

The control cohort was created by randomly sampling patients from CDAP who did not meet any of the 4 criteria that were used

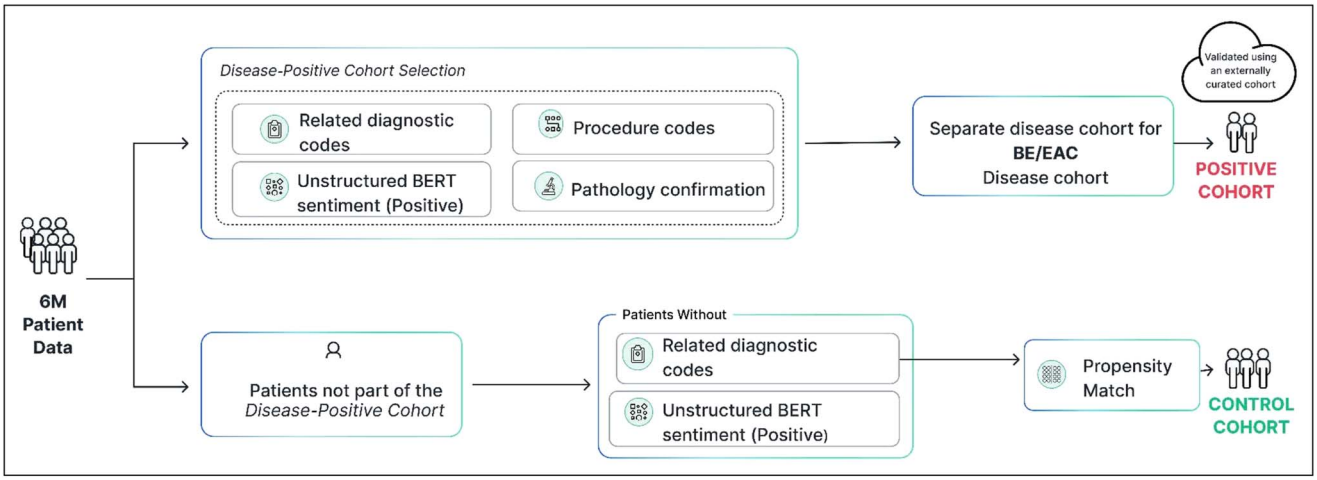

Figure 1. Process of identifying case (BE/EAC) and control cohorts from the Clinical Data Analytics Platform. BE, Barrett's esophagus; BERT, Bidirectional Encoder Representations from Transformer; EAC, esophageal adenocarcinoma.

for identification of the disease cohorts. Hence, the patients in the control cohort did not have either any structured or unstructured evidence for BE or EAC. These sampled patients were then propensity matched to cases on (i) the year of diagnosis (of the case cohort), (ii) the number of structured disease diagnoses during the observation period (see the definition above), and (iii) the proportion of hospitalization in the observation period to the disease cohort (because hospitalization leads to a larger number of medical records per encounter). Of note, the cohorts were not matched to known risk factors of BE/EAC to enable the identification of risk factors agnostic to current knowledge.

Additional inclusion and exclusion criteria

In both the case and control cohorts, patients younger than 18 years and those older than 85 years were excluded. In addition, only patients who met the data completeness criteria (defined as having 2 or more encounters in the observation period) were retained. This was done to ensure that the model had the opportunity to learn from a minimum number of encounters, which optimizes model performance.

Validation of the case identification algorithm

The case identification algorithm described earlier was tested against 2 population-based, manually identified, and annotated cohorts of patients with BE and EAC in South-Eastern Minnesota. This cohort was created using resources from the Rochester Epidemiology Project (19), which is a population-based medical record linkage system, recently expanded to 11 counties in SE Minnesota.

Model training data

The prediction model was trained on the following data elements:

Nontemporal features (variables that do not change with time):

1. Age at lead time
2. Sex
3. Race/ethnicity
4. Family history of BE or EAC
5. Smoking status, defined as current, past, or never.

Temporal features (which may appear and change over time):

6. Medications
7. Comorbidities: based on structured analysis
8. Laboratory tests: hemoglobin, aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, total bilirubin, albumin, creatinine, sodium, potassium, total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, chloride, calcium, glucose, blood urea nitrogen, lipase, amylase, gamma glutamyl transferase, prostate-specific antigen, and hemoglobin A1c. These tests were chosen based on the frequency of occurrence and clinical expertise.
9. Symptoms: identified by augmented curation on patient notes: abdominal pain, dysphagia, dyspepsia, vomiting, diarrhea, heartburn, water brash, chest pain, odynophagia, nausea, snoring, esophageal reflux, dyspnea, arthritis, backache, weight loss, cough, hoarseness, and hematemesis
10. Body mass index.

Model development and architecture

Figure 2 describes the process of data extraction from the observation time of a hypothetical patient. Data from the lead time (1 year before the anchor date: date of initial diagnosis of BE or EAC) were not used for model development, to exclude data which may be reflective of disease symptoms before diagnosis. The observation period (from which data were extracted for model development) extended from 1 year before diagnosis to 6 years before the diagnosis (i.e., a total of 5 years).

A transformer-based ML model architecture was used to develop the predictive model (20). Transformer models use attention mechanisms to capture the temporal interdependencies of words. Patient characteristics such as symptoms, diagnostic codes, medications, laboratory tests, and demographics were used in lieu of words. The temporal sequence of all these features was maintained as they occurred in the patient time line. During model development, the feature and positional vectors are passed to the transformer encoders. The transformer encoder converts the feature and positional vector into an intermediate vector representing a patient's temporal events. This intermediate vector

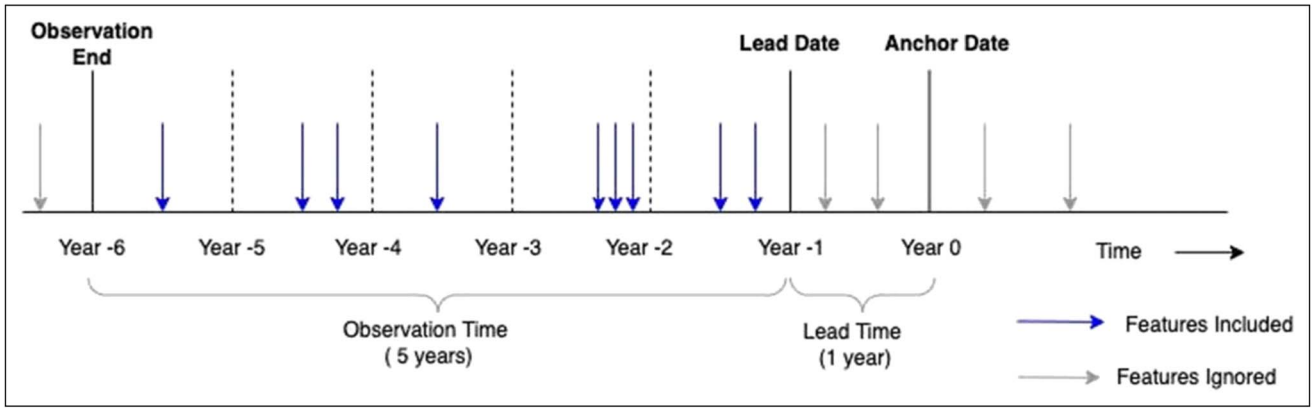

Figure 2. Time line of inclusion of data for patients with BE or EAC included in model development. Anchor date was the date of diagnosis for a patient with BE or EAC. Lead time refers to a period of 1 year before the anchor date. Events in the lead time period were not used to train the model. Observation time is the period 5 years preceding the lead date. All events in the observation period were used to train the model. BE, Barrett's esophagus; EAC, esophageal adenocarcinoma.

is then combined with nontemporal information to generate a comprehensive vector of the patient. This vector is then passed to a softmax layer to estimate the risk. Figure 3a shows the processing of a sequence of events in the patient time line.

Five randomly selected control cohorts were also created, enabling training of 5 transformer prediction models (Figure 3b). Each of the transformer models used the same disease cohort, but was trained with a different control cohort. For the BE model, the case-to-control ratio was 1:5, and for the EAC model, the case-to-control ratio was 1:10. The output of these 5 transformer models were used to train an ensemble model using logistic regression.

Figure 4 describes the data sets used for the model training process. At the outset, 10% of the data were kept aside as a holdout test data set: the Model Holdout Set (MHS). The rest of the data were used in training the transformer and ensemble models: the Development Set (DS). The DS was split into 3 sets in the ratio of 60:20:20. 60% of the DS was used to train the transformer model, the Transformer Training Set. 20% of the DS was used to choose the best epoch for the transformer, the Transformer Epoch Set (TES). The last 20% was used as an Ensemble Test Set. The TES was also used to train the ensemble model. For this, the TES was further split in the ratio of 80:20. 80% of the TES was used for training the ensemble model, the Ensemble Train Set, and 20%

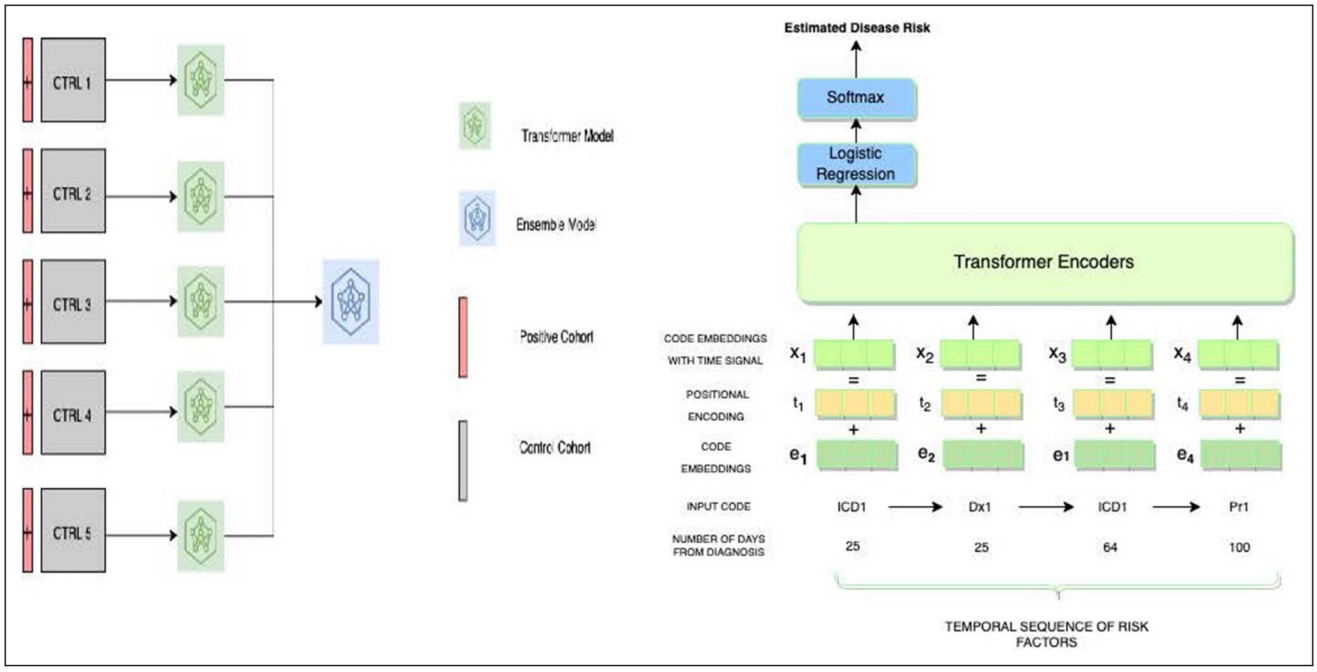

Figure 3. Description of case and control cohort utilization in model development. (**a**) Ensemble model development and architecture. Five independent control cohorts were created. Five control patients were matched to each patient with BE and 10 control patients matched to each patient with EAC. Five transformer models were developed by pairing the BE and EAC case cohort with 5 independent control cohorts. These 5 transformer models were then integrated into a single ensemble model using logistic regression. (**b**) Schematic showing the layers of the transformer model used to build the BE and EAC machine learning predictive models. BE, Barrett's esophagus; EAC, esophageal adenocarcinoma.

was used to choose to calibrate the ensemble model, the Ensemble Development Set.

Model output

The output of the ensemble model was a softmax score (ranging from 0 to 1, 0 reflecting no risk of BE/EAC and 1 reflecting 100% risk of developing BE/EAC). The threshold for dichotomization for the ensemble result (positive versus negative for incident BE or EAC) was chosen based on the Youden J method to maximize the area under the receiver-operating curve (AUROC) of the model (21). A score above the threshold indicates that the patient is at a substantial risk of being diagnosed with BE or EAC in the next year and screening should be considered. The final model performance results were reported on the MHS.

RESULTS

Validation of the case search algorithm

The logic used to generate the case cohorts (BE and EAC) from CDAP captured approximately 94% of the manually annotated BE and EAC SE Minnesota cohorts, lending validity to the case search strategy (see Supplementary Figure 1, Supplementary Digital Content 1, http://links.lww.com/CTG/B14).

Final BE, EAC, and control cohorts

Figure 5 outlines the sequential steps taken (as per the inclusion and exclusion criteria) and corresponding case counts. A total of 8,476 patients with BE and 1,539 patients with EAC were included in the final model development. A total of 252,276 controls were also identified. Baseline characteristics of the case and control cohorts identified using the electronic search strategy are presented in Table 1. Most of the BE and EAC cases were middle-aged White men with a past or current history of smoking. Controls were somewhat younger and more likely to be female than cases. Additional details of the patients included in the analysis are presented in Supplementary Figures 3a–c and 4a–c (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14).

Performance characteristics of prediction models

A threshold of ≥0.13 (on the model probability output softmax score described earlier) was chosen to define a positive BE model result. At this threshold, the sensitivity to identify BE was 76%, at a specificity of 76%, with a model AUROC of 0.84 in the MHS (Table 2). A threshold of ≥0.08 was chosen to define a positive EAC model result. In this study, the sensitivity for EAC detection was higher at 84% with a specificity of 61%, with a model AUROC of 0.84 in the MHS (Table 2).

Given that this model could be applied to the EHR to first identify those at higher risk of BE/EAC, followed potentially by a minimally invasive nonendoscopic test, the threshold to determine positivity for the BE prediction score was set somewhat lower to balance sensitivity, specificity, and overall AUROC. Conversely, for the EAC threshold, higher sensitivity was prioritized to avoid missing EAC. The confusion matrices of both models are presented in Table 3, a and b. Performance of both models at different thresholds, evaluated on the MHS, are summarized in Supplementary Tables 1 and 2 (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14).

An example of the performance characteristics of a BE risk prediction EHR-based model, followed by the administration of a nonendoscopic office-based test, such as a swallowed esophageal cell collection device, combined with methylated DNA markers (with a test sensitivity and specificity of 90%), in a population with a BE prevalence of 5% is demonstrated in Supplementary Figure 2 (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14).

Features used by the model for BE/EAC risk prediction

Of the over 7,500 variables contributing to BE prediction, Supplementary Tables 3 and 4 (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14) list some selected temporal and nontemporal determinants of BE and EAC risk, respectively. Integrated gradients were used to determine the features that the model used in its prediction (22). This method attributes a score to each feature for its contribution toward the final outcome. The attribution score for a feature is aggregated across patients.

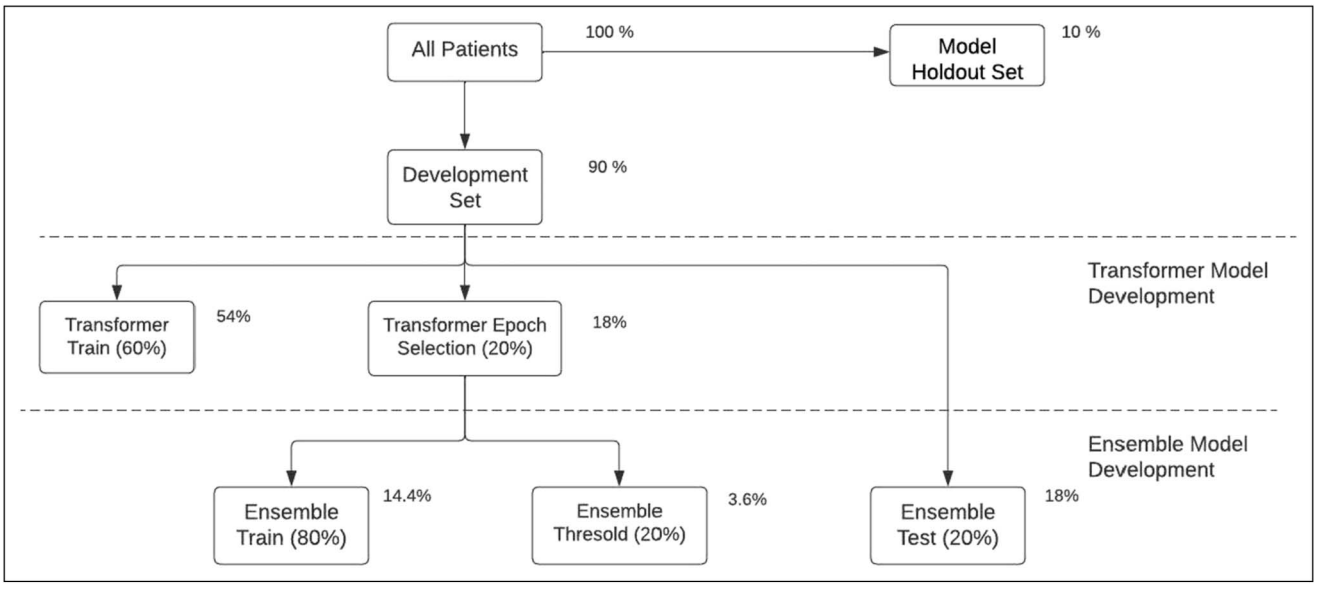

Figure 4. Distribution of CDAP data used for model development and testing. CDAP, Clinical Data Analytics Platform.

Table 1. Baseline characteristics of BE, EAC, and control cohorts identified from the Clinical Data Analytics Platform

Variables	BE (N = 8,476)	EAC (N = 1,539)	Control cohort (N = 252,276)
Age, yr, mean (SD)	63.38 ± 12.16	68.72 ± 10.04	55.24 ± 18.63
Male sex, n (%)	5,475 (64.59)	1,284 (83.43)	117,020 (46.38)
White race, n (%)	8,080 (95.32)	1,450 (94.21)	216,761 (85.92)
Ever smoker, n (%)[a]	5,250 (61.93)	1,019 (66.21)	96,569 (38.27)
BMI >30, n (%)[b]	2,067 (24.38)	389 (25.27)	64,808 (25.68)
Hospitalization, n (%)	3,710 (43.77)	739 (48.01)	111,223 (44.08)

BE, Barrett's esophagus; BMI, body mass index; EAC, esophageal adenocarcinoma.
[a]Smoking data missing in 7% of BE cases, 14% of EAC cases, and 21% of controls.
[b]BMI data were not available in 45% of BE cases, 52% of EAC cases, and 26% of controls.

Feature scores for selected (out of over 400) determinants of BE risk are presented in Supplementary Table 5 (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14). Of note, some determinants (with positive feature scores) increased risk and some (with negative feature scores) reduced risk. Some of the features that influenced BE risk include male sex, age older than 60 years, ever smoking, gastroesophageal reflux disease (GERD) diagnosis, symptoms of heartburn, dyspepsia, comorbidities such as coronary atherosclerosis, serum triglycerides, and electrolytes. Many of the features that predicted EAC were similar to those for the prediction of BE (data not shown). Notably, a history of BE was a predictor of incident EAC.

DISCUSSION

We report the development and internal validation of 2 novel ML ensemble transformer models predicting the risk of incident BE and EAC from a deidentified EHR database of over 6 million patients. These models were more accurate (AUC 0.84) than previously reported models/scoring algorithms, which use conventional risk factors and additional data obtained from the EHR or questionnaires. In addition to established risk factors, the ML model also identified potential novel predictors, such as metabolic and vascular consequences of obesity (coronary artery disease, cardiovascular disease, and hyperlipidemia), hormonal medications, and serum electrolytes.

Several risk factors of BE and EAC have been described. However, estimating BE/EAC risk in an efficient manner has remained challenging. Current guidelines suggest using GERD as an essential criterion, in addition to the use of 1–3 additional risk factors to determine the recommendation of screening. This approach also erroneously assumes that all risk factors contribute the same amount of risk. However, these criteria perform suboptimally for sensitivity and specificity for BE or EAC detection (14,23). A recent study integrated 7 established BE risk factors extracted from the VA EHR into a logistic regression model. The AUC of this model was not higher than that of the models reported before (0.68–0.7) (24).

Unsurprisingly, utilization of these risk algorithms remains limited in practice, despite recommendations for BE screening in guidelines since 2008, with most patients with incident EAC continuing to be diagnosed outside a BE screening and surveillance program, despite its presence endoscopically and/or

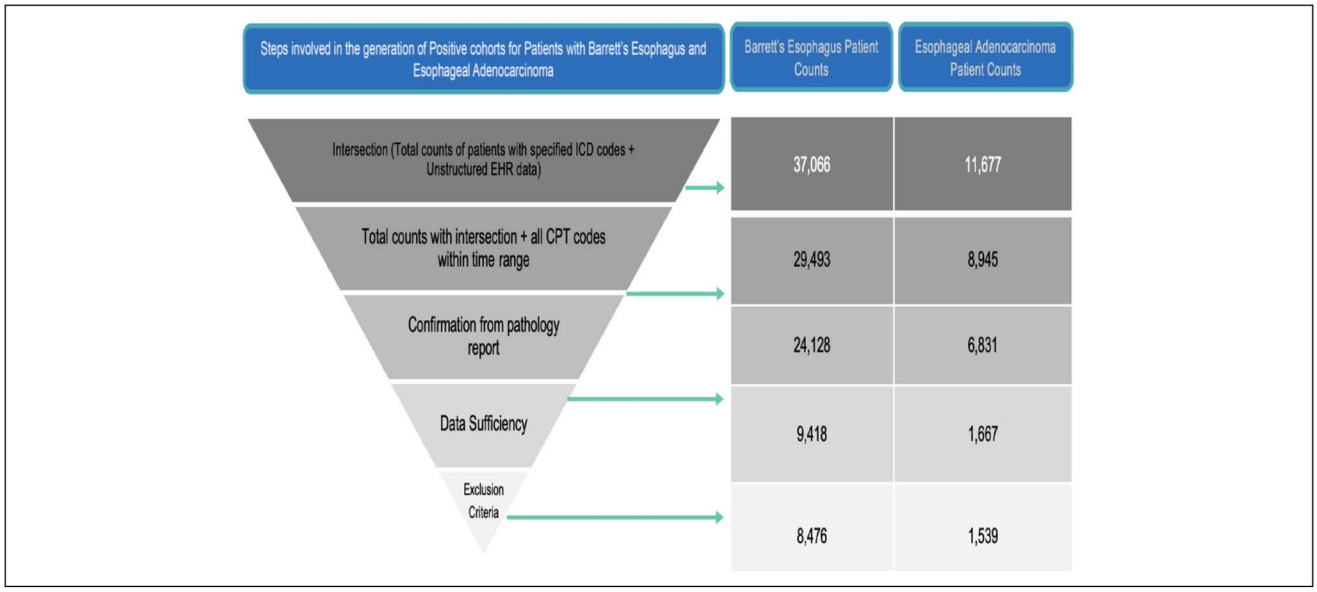

Figure 5. Sequential identification of BE and EAC cases from the CDAP, with the application of prespecified data sufficiency, inclusion criteria, and exclusion criteria. BE, Barrett's esophagus; CDAP, Clinical Data Analytics Platform; EAC, esophageal adenocarcinoma

Table 2. Performance characteristics of the BE and EAC prediction models

Model	AUROC	Sensitivity	Specificity	F1 score	PPV	NPV
BE	0.84	0.76	0.76	0.51	0.39	0.94
EAC	0.84	0.84	0.74	0.34	0.22	0.98

AUROC, area under the receiver-operating curve; BE, Barrett's esophagus; EAC, esophageal adenocarcinoma; F1 score, measure of overall accuracy; NPV, negative predictive value; PPV, positive predictive value.

histologically (25). A possible reason is that the identification and enumeration of these risk factors (either by themselves or by integrating them into combined scores, some of which use specialized data points not routinely measured and recorded in the EHR, without an actionable threshold) is time-consuming in a busy primary care practice (where this decision needs to be made most commonly). Multiple medical problems and other guideline-supported screenings compete for attention in a limited amount of time during patient visits.

An EHR-based ML risk assessment tool overcomes several of these barriers by identifying and integrating several established and novel risk factors of incident BE/EAC, automatically extracting them from the EHR and integrating them into a risk score, improving the accuracy of predicting incident BE/EAC. Although we and others have reported on the association of other metabolic consequences of obesity with BE/EAC (26–30), this ML model is the first to integrate these additional risk factors into a comprehensive risk score along with conventional risk factors. Such a score may be dichotomized by setting a threshold to recommend screening and integrated into the EHR as an electronic trigger tool, which can flag patients who are at an increased risk of BE/EAC, and also be linked to order sets for screening tools, further reducing burden on providers (15).

The threshold at which the ML model should trigger a recommendation for screening needs to be further studied and defined. The threshold will likely depend on the test that will be used as a screening tool and can likely be lower if a sensitive, minimally invasive nonendoscopic tool will be the first tool to evaluate risk. Such a tool with a resultant high negative predictive value can be useful (see Supplementary Figure 2, Supplementary Digital Content 1, http://links.lww.com/CTG/B14) in a low-prevalence population. Contrarily, with a more invasive test, such as sedated endoscopy, a higher (more specific) threshold to determine the risk of BE may be set. An alternative approach may include a definitive test, such as endoscopy for those with higher risk scores and a nonendoscopic test for those with moderately high scores. Modeling studies focused on varying thresholds and linking them to outcomes such as BE detection, EAC incidence or mortality would be needed, because empirical studies to address this issue may be challenging to conduct.

Table 3. Confusion matrices for the BE and EAC transformer prediction models (from the Model Holdout Set)

a: BE model		Actual	
		True	False
Predicted	True	586	929
	False	183	2,913
b: EAC model		**Actual**	
		True	**False**
Predicted	True	110	398
	False	21	920

BE, Barrett's esophagus; EAC, esophageal adenocarcinoma.

This study has a number of strengths. The EHR database had a substantial number of patients with BE and controls. The case identification algorithm was further validated against a manually curated and annotated population-based database of patients with BE and EAC. The large number of control patients enabled the creation of a robust ensemble model (combining 5 individual models). The transformer model (which was selected for this study) also takes into account the temporal aspects of the clinical features and eliminates the requirement of imputing missing features for a patient. Hence, the model can be used even where the entire information of a patient is not available. This is particularly relevant when implemented clinically. Although cases and controls were propensity matched to a few variables, largely to ensure comparable data points between cases and controls, they were not matched on known risk factors (such as age and sex). This explains some of the age and sex imbalance between cases and controls, and the demographics of the control cohort are reflective of the Mayo Clinic population.

The deep learning approach also allowed the identification of several additional predictors, increasing the accuracy of the model. Given that BE is a precursor for EAC, we developed 2 ML models for the prediction of both diseases independently. It is reassuring to see that despite the smaller number of EAC cases, the accuracy of the model was comparable with that of BE, and BE was identified as a strong risk factor of EAC. Model performance remained robust even after elimination of all laboratory values and medications (AUC = 0.827, detailed results in Supplementary Methods, see Supplementary Digital Content 1, http://links.lww.com/CTG/B14). Given that some of the risk models have been developed in VA populations, this model is more likely to be generalizable to other populations. However, external validation in other populations and other medical record systems (with potentially different patterns of structured and unstructured data) will be required before clinical implementation.

Deep learning models do have some limitations, particularly regarding explainability of models. While it is possible to determine whether a feature contributed to risk, it is hard to quantify the effect of the feature on the final outcome. For example, the features listed in Supplementary Table 5 (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14) have a disease and control score. A higher disease score compared with the control score would indicate that the feature is predictive of the disease. However, the actual number does not correlate with the feature weight (such as an odds or hazard ratios) in predicting the outcome. Note that these scores were calculated using methods previously described (31). This is also evident in

Supplementary Tables 4 and 5 (see Supplementary Digital Content 1, http://links.lww.com/CTG/B14) where several variables have small feature scores and have feature scores for the same variable as normal or abnormal (because they were dichotomized into normal and abnormal values). It is also hard to quantify how the cardinality and temporality of features affect a particular risk score. For example, it is hard to quantify the effect of how much a GERD diagnosis 3 years before the lead time is weighted compared with the same diagnosis 2 years before the lead time. This is an area of ongoing research in ML.

Implementation of such ML models into practice will require multiple additional technical and logistical steps. These models will have to be adapted to EHR platforms using patient identifiers and should be able to run in real time when patients are seen in the clinic or in batches with mechanisms to contact patients identified as high-risk and order diagnostic EGD. These processes will have to be tested iteratively. In addition, the medical record technology to present this to providers as an electronic trigger tool will also have to be developed and pilot tested before formal implementation. Provider input and buy-in for best practices to implement such triggers will also need to be sought to minimize burden and distraction. We have already begun to identify the relevant stakeholders, information technology support, and resources to pilot test the algorithm at a specific location in our own health system. If the algorithm has to be implemented in another EHR system, the model will likely have to be modified or adapted to the specific EHR and its performance validated in an independent test set before implementation. Ultimately, the impact of such models and BE screening tests on outcomes such as EAC incidence and mortality will have to be assessed; it can take decades to establish cancer control from a new screening process or intervention.

In conclusion, we have successfully developed and internally validated 2 novel and more accurate ML models to predict the risk of incident BE and EAC from a large EHR database. This model has the potential to be integrated into a medical EHR to predict BE/EAC risk in real time, facilitating the implementation of nonendoscopic BE detection technology as an intermediate step to definitive diagnosis with sedated endoscopy (in those testing positive with the nonendoscopic test). External validation (on EHRs which are deidentified or have patient identifiers) and clinical testing are critical next steps before implementation.

CONFLICTS OF INTEREST

Guarantor of the article: Prasad G. Iyer, MD, MSc, FACG.

Specific author contributions: P.G.I.: concept, obtaining funding, writing the initial draft, and revisions. K.S.: data collection. C.L., D.C.C., H.A., K.A., and J.B.K.: editing the manuscript. S. Asfahan, S. Awasthi, P.A., P.K.M., S.P.S., S.S., S.B., C.M., and S.Y.: model development. N.S. and C.P.: model development and editing the manuscript.

Financial support: Supported in part by a NIH grant (NCI R01CA241164), the Mayo Foundation, and the Freeman Foundation.

Potential competing interests: P.G.I.: research funding: Exact Sciences, Pentax Medical, CDx Medical, and Castle Biosciences; consultant: Exact Sciences, Pentax Medical, CDx Medical, Castle Biosciences, Ambu, and Symple Surgical. C.L.: consultant Verily Life Sciences. J.B.K.: research funding and intellectual property, Exact Sciences. The remaining authors have no disclosures.

IRB approval: Approved by the Mayo Clinic IRB.

Study Highlights

WHAT IS KNOWN

- ✓ Screening for Barrett's esophagus (BE), which is the sole known precursor for esophageal adenocarcinoma (EAC), is recommended in those with multiple risk factors.
- ✓ Screening rates in those with risk factors remain low.
- ✓ Tools to assess BE/EAC risk are modestly accurate (area under the curve [AUC] 0.7 or less) and challenging to implement in clinical practice.
- ✓ With the advent of nonendoscopic minimally invasive BE detection tools, such risk assessment tools are more relevant.

WHAT IS NEW HERE

- ✓ A machine learning-powered tool to assess the risk of incident BE/EAC was developed using variables available in the deidentified electronic health record of 6 million patients.
- ✓ This tool is more accurate than those available currently (AUC = 0.84).
- ✓ Such a tool could be integrated into the electronic health record as a clinical prompt for providers to consider BE screening at clinically appropriate thresholds.

REFERENCES

1. Curtius K, Rubenstein JH, Chak A, et al. Computational modelling suggests that Barrett's oesophagus may be the precursor of all oesophageal adenocarcinomas. Gut 2020;70(8):1435–40.
2. Shaheen NJ, Falk GW, Iyer PG, et al. Diagnosis and management of Barrett's esophagus: An updated ACG guideline. Am J Gastroenterol 2022;117(4):559–87.
3. Muthusamy VR, Wani S, Gyawali CP, et al. AGA clinical practice update on new technology and innovation for surveillance and screening in Barrett's esophagus: Expert review. Clin Gastroenterol Hepatol 2022; 20(12):2696–706.e1.
4. Asge Standards Of Practice C, Qumseya B, Sultan S, et al. ASGE guideline on screening and surveillance of Barrett's esophagus. Gastrointest Endosc 2019;90(3):335–59.e2.
5. Sami SS, Moriarty JP, Rosedahl JK, et al. Comparative cost effectiveness of reflux-based and reflux-independent strategies for Barrett's esophagus screening. Am J Gastroenterol 2021;116(8):1620–31.
6. Rubenstein JH, Evans RR, Burns JA, et al. Patients with adenocarcinoma of the esophagus or esophagogastric junction frequently have potential screening opportunities. Gastroenterology 2022;162(4):1349–51.e5.
7. Eluri S, Reddy S, Ketchem CC, et al. Low prevalence of endoscopic screening for Barrett's esophagus in a screening-eligible primary care population. Am J Gastroenterol 2022;117(11):1764–71.
8. Kolb JM, Chen M, Tavakkoli A, et al. Understanding compliance, practice patterns, and barriers among gastroenterologists and primary care providers is crucial for developing strategies to improve screening for Barrett's esophagus. Gastroenterology 2022;162(6):1568–73.e4.
9. Rubenstein JH, Morgenstern H, Appelman H, et al. Prediction of Barrett's esophagus among men. Am J Gastroenterol 2013;108(3):353–62.
10. Xie SH, Ness-Jensen E, Medefelt N, et al. 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;113(6):829–35.
11. Thrift AP, Garcia JM, El-Serag HB. A multibiomarker risk score helps predict risk for Barrett's esophagus. Clin Gastroenterol Hepatol 2014; 12(8):1267–71.
12. Kunzmann AT, Thrift AP, Cardwell CR, et al. Model for identifying individuals at risk for esophageal adenocarcinoma. Clin Gastroenterol Hepatol 2018;16(8):1229–36.e4.
13. Rubenstein JH, McConnell D, Waljee AK, et al. Validation and comparison of tools for selecting individuals to screen for Barrett's esophagus and early neoplasia. Gastroenterology 2020;158(8):2082–92.

14. Sawas T, Zamani SA, Killcoyne S, et al. Limitations of heartburn and other societies' criteria in Barrett's screening for detecting de novo esophageal adenocarcinoma. Clin Gastroenterol Hepatol 2022;20(8):1709–18.
15. Murphy DR, Meyer AN, Sittig DF, et al. Application of electronic trigger tools to identify targets for improving diagnostic safety. BMJ Qual Saf 2019;28(2):151–9.
16. Wagner T, Awasthi S, Wittenberg G, et al. Real-time biomedical knowledge synthesis of the exponentially growing world wide web using unsupervised neural networks. BioRxiv 2020:2020.04.03.020602.
17. Wagner T, Shweta F, Murugadoss K, et al. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. Elife 2020;9:e58227.
18. Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018:arXiv: 1810.04805.
19. Dhaliwal L, Codipilly DC, Gandhi P, et al. Neoplasia detection rate in Barrett's esophagus and its impact on missed dysplasia: Results from a large population-based database. Clin Gastroenterol Hepatol 2021;19(5): 922–9.e1.
20. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems 30; 2017.
21. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.
22. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In International Conference on Machine Learning, PMLR, 2017.
23. Nguyen TH, Thrift AP, Rugge M, et al. Prevalence of Barrett's esophagus and performance of societal screening guidelines in an unreferred primary care population of U.S. veterans. Gastrointest Endosc 2021;93(2): 409–19.e1.
24. 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.
25. Tan MC, Mansour N, White DL, et al. Systematic review with meta-analysis: Prevalence of prior and concurrent Barrett's oesophagus in oesophageal adenocarcinoma patients. Aliment Pharmacol Ther 2020; 52(1):20–36.
26. Leggett CL, Nelsen EM, Tian J, et al. Metabolic syndrome as a risk factor for Barrett esophagus: A population-based case-control study. Mayo Clin Proc 2013;88(2):157–65.
27. Iyer PG, Borah BJ, Heien HC, et al. Association of Barrett's esophagus with type II diabetes mellitus: Results from a large population-based case-control study. Clin Gastroenterol Hepatol 2013;11(9):1108–14.e5.
28. Sawas T, Dilmaghani S, Dhaliwal L, et al. Risk factor profiles can distinguish esophageal adenocarcinoma from Barrett's esophagus. Am J Gastroenterol 2021;116(1):198–201.
29. Saleh S, Trujillo S, Ghoneim S, et al. Effect of hormonal replacement therapy on gastroesophageal reflux disease and its complications in postmenopausal women. Clin Gastroenterol Hepatol 2023;21(2): 549–51.e3.
30. Petrick JL, Falk RT, Hyland PL, et al. Association between circulating levels of sex steroid hormones and esophageal adenocarcinoma in the FINBAR Study. PLoS One 2018;13(1):e0190325.
31. Petch J, Di S, Nelson W. Opening the black box: The promise and limitations of explainable machine learning in cardiology. Can J Cardiol 2022;38(2):204–13.

Open Access This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.