Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2016 May ; 25(5): 727–735. doi:10.1158/1055-9965.EPI-15-0832.

Predicting Barrett's esophagus in Families: An Esophagus Translational Research Network (BETRNet) Model Fitting Clinical Data to a Familial Paradigm

Xiangqing Sun[1], **Robert C. Elston**[1,2], **Jill S. Barnholtz-Sloan**[1,2], **Gary W. Falk**[3], **William M. Grady**[4], **Ashley Faulx**[5,6], **Sumeet K. Mittal**[7], **Marcia Canto**[8], **Nicholas J. Shaheen**[9], **Jean S. Wang**[10], **Prasad G. Iyer**[11], **Julian A. Abrams**[12], **Ye D. Tian**[1], **Joseph E. Willis**[13], **Kishore Guda**[14], **Sanford D. Markowitz**[15], **Apoorva Chandar**[5], **James M. Warfe**[1], **Wendy Brock**[5], and **Amitabh Chak**[2,5]

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University

[2]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine

[3]University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

[4]Clinical Research Division, Fred Hutchinson Cancer Research Center; Gastroenterology Division, University of Washington School of Medicine, Seattle, WA

[5]Division of Gastroenterology and Hepatology, University Hospitals Case Medical Center, Case Western Reserve University School of Medicine

[6]Division of Gastroenterology and Hepatology, Louis Stokes Veterans Administration Medical Center, Case Western Reserve University School of Medicine

[7]Department of Surgery, Creighton University School of Medicine

[8]Division of Gastroenterology, Johns Hopkins Medical Institutions, Baltimore, MD

[9]Center for Esophageal Diseases & Swallowing, University of North Carolina at Chapel Hill School of Medicine

[10]Division of Gastroenterology, Washington University School of Medicine

[11]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN

[12]Department of Medicine, Columbia University Medical Center

[13]Department of Pathology, University Hospitals Case Medical Center, Case Western Reserve University School of Medicine

[14]Division of General Medical Sciences (Oncology), Case Comprehensive Cancer Center

[15]Department of Medicine and Case Comprehensive Cancer Center, Case Medical Center, Case Western Reserve University

Corresponding author: Amitabh Chak, MD. Wearn 242, University Hospitals Case Medical Center. 11100 Euclid Avenue, Cleveland, OH 44106. Phone: 216-844-3217; Fax: 216-844-7480; ; Email: Amitabh.Chak@uhhospitals.org.

Conflict of interest: Dr. Prasad G. Iyer disclosed commercial research grants from Intromedic and Exact Sciences. No potential conflicts of interest were disclosed for other authors.

Abstract

Background—Barrett's esophagus (BE) is often asymptomatic and only a small portion of BE patients are currently diagnosed and under surveillance. Therefore, it is important to develop risk prediction models to identify high-risk individuals with BE. Familial aggregation of BE and esophageal adenocarcinoma (EAC), and the increased risk of EAC for individuals with a family history, raise the necessity of including genetic factors in the prediction model. Methods to determine risk prediction models using both risk covariates and ascertained family data are not well-developed.

Methods—We developed a Barrett's Esophagus Translational Research Network (BETRNet) risk prediction model from 787 singly ascertained BE pedigrees and 92 multiplex BE pedigrees, fitting a multivariate logistic model that incorporates family history and clinical risk factors. The eight risk factors age, sex, education level, parental status, smoking, heartburn frequency, regurgitation frequency, and use of acid suppressant, were included in the model. The prediction accuracy was evaluated on the training dataset and an independent validation dataset of 643 multiplex BE pedigrees.

Results—Our results indicate family information helps to predict BE risk, and predicting in families improves both prediction calibration and discrimination accuracy.

Conclusions—Our model can predict BE risk for anyone with family members known to have, or not have, had BE. It can predict risk for unrelated individuals without knowing any relatives' information.

Keywords

Barrett's esophagus; Esophageal adenocarcinoma; Clinical predictors; Genetic factors; family

Introduction

During the past thirty years, the incidence of esophageal adenocarcinoma (EAC) has increased dramatically up to 7-fold in the United States (1,2). EAC is a lethal cancer with 5-year survival rates lower than 20% (3). Barrett's esophagus (BE) is the only known precursor for EAC, with BE patients having an 11- to 30-fold increased risk of developing EAC (4,5). Moreover, BE and EAC aggregate in some families (6). The population prevalence of BE is estimated to be between 1-3% (7,8), and the prevalence of BE in the family members of patients who had BE is estimated to be higher than in the general population (8%) (9). However, BE is often asymptomatic and, among older people, it may actually be associated with a decreased GERD symptom burden (10,11). It is estimated that only 5% of patients who have BE are currently diagnosed and under surveillance. The majority of EAC patients do not have an antecedent diagnosis of BE and are diagnosed in the late stages of the disease (12). To detect EAC early, it is necessary to first develop a risk prediction model that identifies high-risk individuals with BE.

To date, models that include clinical/environmental risk factors for predicting BE or EAC risk have been developed using unrelated case-control or cohort data (13-17). Although the rapidly increasing incidence of EAC in recent decades indicates the importance of environmental factors in assessing the disease risk, familial aggregation of Barrett's esophagus and esophageal adenocarcinoma, and the increased risk of EAC for individuals with a family history (6,9), raise the necessity of including familial factors in the prediction model. The heritability of polygenic liability to BE and EAC has been estimated to be 35% and 25%, respectively (18). In this study, we developed a Barrett's Esophagus Translational Research Network (BETRNet) model that incorporates family members' information in addition to clinical risk factors in predicting an individual's absolute risk of BE.

Material and methods

Data

BE was rigorously defined as the presence of intestinal metaplasia in biopsies obtained from endoscopically-visible columnar mucosa in the tubular esophagus. Intestinal metaplasia on biopsies of the gastroesophageal junction or an irregular Z-line was not accepted as BE. We considered BE, EAC, and gastroesophageal junctional adenocarcinomas (JAC) to be part of the same trait, theorizing that at least a proportion of these cancers arose from BE. For assessing risk of BE in our prediction models we considered any individual with BE, EAC, or gastroesophageal junctional adenocarcinoma to be affected. Individuals were considered unaffected only if they had upper endoscopy that excluded Barrett's esophagus. To ascertain the exposure variables, we used a standardized questionnaire based on the validated MAYO GERD questionnaire (19). This modified questionnaire has previously been used in studies of familial aggregation (20).

Through the Barrett's Esophagus Translational Research Network (BETRNet), we collected two independent sets of Barrett's esophagus pedigree data, which are respectively used as training and validation data. The training set comprises 787 singly ascertained and 92 multiplex pedigrees, each of which includes two or more persons, at least one of which is affected with BE or EAC. Single ascertainment assumes that a pedigree enters the sample analyzed because of only a single proband. Our methodology for collecting this initial set of pedigrees has been previously reported (6,20,21). These pedigrees were used to estimate the parameters of the prediction model (coefficients of covariates, genotypic susceptibilities and the allele frequency at a trait locus). The independent validation dataset was collected separately but used the same trait definitions and comprised 643 multiplex familial BE pedigrees. A summary of the characteristics of the members of the training and validation datasets is shown in Table 1.

The prediction model

Using a pedigree likelihood based on a multivariate logistic model (MLM) that includes segregation at a major trait locus (22), we assessed the predictive utility of genetic factors, demographic factors (age, sex, body mass index (BMI), education level, parental status), environmental factors (smoking, alcohol consumption, use of acid suppressant medications, gastroesophageal reflux symptoms such as GERD and heartburn), and affection status among family members. The prediction model was obtained from a generalized and modified version of the SEGREG program in the S.A.G.E. package (23). In SEGREG, although there is no restriction on the size of the family, the family is assumed to be noninbred and to contain no children of consanguineous spouse pairs. We considered all the predicting variables that were reported to be risk factors of BE or EAC (13-17), and found 13 variables available in our data.

Values of the predictor variables for most of the unaffected individuals were missing in both the training and validation datasets (Supplementary Table S1). Nevertheless, we were able to impute age (for the missing 40%) using the family structure according to the method of Schnell et al (24). Moreover, observing that the sum of known date of birth (DOB) and age at examination are mostly between 1999 and 2010, we could to a good approximation assume that the time of diagnosis (DOB + age at examination) is equal for all members in a family, and hence impute age on 98.5% of the individuals (Supplementary Methods, Supplementary Figures S1 - S2). The training dataset of 787 pedigrees that we used initially to estimate the parameters for the prediction model has missing values on the predictor variables for most of the unaffected individuals (94% missing, Supplementary Table S1). Because the missingness depends on the affection status, not on the predictor variable values, it should not influence the estimation of regression coefficients on the covariates; but it will influence the estimation of the baseline risk, which is related to the genetic parameters (the allele frequencies at an assumed trait locus and the genotypic penetrances). However, because the ascertainment of the BE pedigrees would also influence estimation of the genetic parameters, we estimated the covariate and genetic parameters in the MLM model in two separate steps (figure 1).

Step 1. Determine and estimate the covariate regression coefficients. In this step, we fitted multivariate logistic models without adjusting for ascertainment using the training data of 787 pedigrees. We initially included all the thirteen prediction variables (ln(age), sex, education level, parental status, years of smoking, smoking packs per day, use of alcohol, heartburn frequency, age of onset of heartburn, regurgitation frequency, age of onset of regurgitation, BMI, and use of acid suppressant) as covariates of a single baseline susceptibility in the MLM model. The clinical predictors were coded to make the variables jointly linear on the susceptibility scale (logit scale) (Supplementary Methods, Supplementary Figures S3, S4). We then stepwise removed covariates on the basis of likelihood ratio tests (LRT) and Akaike's information criterion (AIC). This analysis is identical to multivariable logistic regression. We also evaluated the covariates of susceptibility by fitting MLM models with two genetic susceptibilities of a latent two-allele locus (Supplementary Table S2). This analysis, in which the penetrance of the heterozygote could equal that of either homozygote, indicated incomplete dominance of the disease-susceptibility allele. Eight covariates, ln(age), sex, education level, parental status, years of smoking, heartburn frequency, regurgitation frequency, and use of acid suppressant, were finally determined as the variables to use as covariates in the prediction model (Supplementary Table S3). Their regression coefficients estimated in the MLM model that assumed a mixture of two (latent) genetic susceptibilities determined by a dominant one-locus model were used in the prediction.

Step 2. Estimate the genetic parameters. In order to estimate the genetic parameters of the MLM model, which are the genotypic susceptibilities on the logit scale and the population allele frequency at the trait locus, we refitted by maximum likelihood the two-allele model, but now adjusting the likelihood for ascertainment. In doing this the estimates of the 8 covariates were fixed at the estimates from the first step The 787 BE pedigrees in the training dataset were taken to be singly ascertained, so in estimating the genetic parameters we adjusted for single ascertainment in the likelihood function by conditioning the likelihood function on the phenotype of the proband. Once again the result indicated incomplete dominance of a disease-susceptibility allele.

The additional multiplex 92 BE pedigrees, which were collected later, were not singly ascertained. We combined them with the 787 singly ascertained pedigrees, and also estimated the genetic parameters using these 879 pedigrees. Because the 879 pedigrees were not all singly ascertained, we both used an adjustment for single ascertainment and additionally added a prevalence constraint on the model when maximizing the likelihood function (25), the population prevalence being constrained to that estimated from the 787 pedigrees.

With the estimated parameters for the prediction models, the probability that a random family member *i* will have BE (by a particular age and at defined covariate values, omitted below for clarity), given the family information available, is then obtained from the pedigree likelihoods according to Bayes' theorem:

$$\begin{aligned} p_i &= \text{prob}(\text{individual i is affected}|\text{other family members' information}) \\ &= \frac{\text{prob (i is affected, other family members' information)}}{\text{prob(other family members' information)}} \\ &= \frac{\text{L(i is affected, other family members' information)}}{\text{L(i is affected, other family members' information)}+\text{L(i is unaffected, other family members' information)}} \\ &= \frac{L_A}{L_A+L_U} \end{aligned} \quad (1)$$

where L_A=L(i is affected, other family members' information) is the pedigree likelihood with individual *i* assumed to be affected and L_U=L(i is unaffected, other family members' information) is the pedigree likelihood with individual *i* assumed to be unaffected. Both pedigree likelihoods can be outputs of SEGREG.

Evaluating the prediction accuracy of the model

Two assessments were used to evaluate the accuracy of the prediction model. One used the calibration criterion, which measures how well the average predicted probabilities agree with the proportion of individuals who actually developed disease. The other used the discrimination criterion, which measures how well the model can separate cases (affected) from controls (unaffected). The calibration assessment we used is the ratio of the observed count to the expected count of BE subjects, O/E (26,27). The expected count is the sum of the predicted probabilities of being affected with BE over all individuals in the sample, and the observed count is the actual number of affected individuals. A 95% confidence interval for this ratio is between a lower limit of $(O/E)\exp(-1.96\times O^{-1/2})$ and an upper limit of $(O/E)\exp(1.96\times O^{-1/2})$ (26,27). A *P*-value for departure from goodness-of-fit can be obtained from the χ^2 distributed statistic $(O-E)^2/E$ with one degree of freedom (26,27). The discrimination assessment we used is the c statistic, which is the area under the receiver operating characteristic (ROC) curve (AUC). The c statistic is a function of sensitivity (true positive rate) and specificity (true negative rate); it measures the probability that predicting the outcome is better than chance (50%).

With the estimated prediction model, we can predict the absolute risk for an individual in a randomly sampled pedigree using formula (1), by calculating two unconditional pedigree likelihoods while fixing the parameter estimates in the MLM model. The prediction accuracy of the prediction model was evaluated in the independent validation dataset of 643 BE pedigrees, as well as in the 787 pedigrees of the training dataset. However, neither dataset was randomly sampled from the population: the data on the 787 BE pedigrees were singly ascertained through a proband, the independent data on the 643 BE pedigrees were multiplex pedigrees.

In order to predict the risk for individuals in singly ascertained pedigrees, we need to calculate the pedigree likelihoods conditioned on the appropriate subset C, which includes the probands (23). The risk for such an individual is:

$$\begin{aligned} p_{ic} &= \text{prob(individual i is affected}|\text{other family members' information, and the proband's information and affected} \\ &= \frac{\text{prob(i is affected, other family members' information | the proband's information and affected status)}}{\text{prob(other family members' information | the proband's information and affected status)}} \\ &= \frac{\text{L(i is affected, other family members' information | the proband's information and affect}}{\text{L(i is affected, other family members' information | the proband's information and affected status)+L(i is unaffected, other family members'}} \quad \text{(2} \\ &= \frac{L_{AC}}{L_{AC}+L_{UC}} \end{aligned}$$

where L_{AC} and L_{UC} are the two conditional pedigree likelihoods output by SEGREG on adjusting for single ascertainment.

In SEGREG,

$$\begin{aligned} L_{AC} &= \text{L(i is affected, other family members' information}|\text{the proband's information and affected status)} \\ &= \frac{L_A}{L_C} \end{aligned} \tag{3}$$

and

$$\begin{aligned} L_{UC} &= \text{L(i is unaffected, other family members' information}|\text{the proband's information and affected status)} \\ &= \frac{L_U}{L_C} \end{aligned} \tag{4}$$

where L_A and L_U are as given in formula (1) and L_C is the likelihood function for the proband. C is the subset of pedigree members to be conditioned on, which includes only the probands (singly ascertained pedigrees have one proband per pedigree). L_C is taken to be the pedigree likelihood L computed as though all individuals not in C are missing (23).

(1) For a non-proband in a singly ascertained pedigrees, according to formulae (2) - (4), the probability of being affected given his/her family members' information and the proband's status is $p_{ic}=\frac{L_{AC}}{L_{AC}+L_{UC}}=\frac{\frac{L_A}{L_C}}{\frac{L_A}{L_C}+\frac{L_U}{L_C}}=\frac{L_A}{L_A+L_U}=p_i$, which means his/her risk equals the risk for an individual in a randomly sampled pedigree. This equality is intuitive, because for a non-proband in a singly ascertained pedigree, the probability of being affected p_{ic} = prob(individual i is affected|other family members' information, and the proband' s information and affected status), and because the proband is one of the "other family members", p_{ic} = prob(individual i is affected|other family members' information) = p_i. This equality also indicates that single ascertainment can be automatically adjusted by predicting in a family.

(2) For a proband in a singly ascertained pedigree, the probability of being affected is (see Supplementary Methods)

$$p_{ic}=\text{prob(individual i is affected|other family members' information, and the proband's information and affected} \\ =\begin{cases} 1, & \text{if the proband is affected} \\ 0, & \text{if the proband is unaffected} \end{cases} \qquad \text{(5}$$

This shows that because the proband's affection status in a singly ascertained pedigree is already known, the risk of a proband being affected is either 1 or 0 depending on affection status.

In evaluating the prediction model using our ascertained pedigrees, we used formula (1) or (2) to predict the BE risk for non-probands (they produce exactly the same risk), and used formula (5) to predict the risk for probands. For any individual from a random family or for an unrelated individual in the population, we can predict his/her BE risk by formula (1). The prediction accuracy for all individuals (probands and non-probands) and for the non-probands alone were respectively evaluated.

Estimating the variance due to genetic factors and the variance due to environmental, demographic and clinical factors

In order to study how much the prediction is improved by using family information, we estimated the variance due to genetic factors and the variance due to other factors using the training dataset that the model was estimated from, because it had many more non-probands than did the validation dataset. We predicted the risk for individual i in a family (denoted by $R(G_i,xi)$), and estimated the predicted risk for any individual i assuming all individuals are unrelated, which means that everyone has the same genotypic frequencies $\overline{G}$, and has the corresponding risk $R(\overline{G},x_i)$. We also estimated the risk in families but assuming that every individual has the same covariate value $\overline{x}$ ($\overline{x}$ = E(x)), and thus estimated the corresponding risk $R(G_i,\overline{x})$. Whether on the logit scale or on the probability scale, the genetic factors (genotypic frequencies G) and the environmental factors (covariates x) are not linear in the risk, and therefore among the non-probands (as well among all individuals), mean($R(G_i,x_i)$) ≠ mean($R(\overline{G},x_i)$) ≠ mean($R(G_i,\overline{x})$). In order to roughly estimate the variance explained by the genetic and environmental factors, we made the three means equal in the following way. In estimating $R(\overline{G},x_i)$, we found the allele frequency (q = 0.096) that made mean($R(\overline{G},x_i)$) = mean($R(G_i,x_i)$) (where mean($R(G_i,x_i)$) = 0.113); Similarly, in estimating $R(G_i,\overline{x})$, we found the value k (k = −0.433) that made the mean covariate value $\overline{x}$ ($\overline{x}$=E(x)+k×SD(x)×sign(β_x)) such that ($R(G_i,\overline{x})$) = mean($R(G_i,x_i)$) = 0.113, where sign(β_x) is the sign of the estimated regression on covariate x.

Results

The estimated parameters in the prediction model

Using the 787 singly ascertained pedigrees, we estimated the coefficients of the predictor covariates without adjusting for ascertainment in a multivariable logistic model that assumed a mixture of two (latent) genetic susceptibilities determined by a dominant one-locus model. The estimates of the regression coefficients (Table 2) should be approximately unbiased provided only that the linear logistic model is appropriate for the fixed effects (28).

The genetic parameters (genotypic susceptibilities and allele frequency of the trait locus) were then estimated by maximum likelihood under a dominant model while adjusting the likelihood for ascertainment. In estimating the genetic parameters, the 8 covariates were fixed at the estimates in Table 2. The estimate of the genetic parameters using the 787 singly ascertained pedigrees (model 1) and the one with additional 92 pedigrees (model 2) are listed in Table 3; on the penetrance scale the estimates from the two datasets are close to each other.

Predicting BE risk from the prediction model

As an example, we show in Table 4 the probability that a family member will have BE by age 50 or 70 according to the affection status of one or two older siblings. As we can see, for a 50 year old male, if he has one sister with BE and one sister without BE (the last row), his probability of being affected is 13.8% (by prediction model 1); if he has both sisters affected, his probability of being affected increases to 32.9%; and if he has two affected brothers, his probability of being affected is 26.6%.

Prediction accuracy

The accuracy of the prediction model that was fitted from the training dataset of 787 singly ascertained pedigrees (model 1) was evaluated separately on both the training dataset itself, and the independent validation dataset of 643 pedigrees. We evaluated this model because this model is fitted by the 787 singly ascertained pedigrees and theoretically such single ascertainment has been accurately adjusted for in this model. Moreover, the ascertainment adjustment of model 2 was based on the prevalence from model 1, and the estimates and the predictions of the two models are very close, so we evaluated the fundamental model 1. In these two datasets, the clinical variables used in the prediction model were largely missing, especially on the unaffected individuals. In the training dataset, there were 689 pedigrees with 1170 individuals who were informative for all the covariates (Table 1), and in the validation dataset, only 248 pedigrees with 420 individuals were informative for all the 8 covariates. These are the individuals who were used to predict the risk of Barrett's esophagus.

Because the datasets were not randomly ascertained, predicted BE risks for the probands are known to be either 0 or 1 (formula 5), so prediction for non-probands is more meaningful to evaluate the prediction performance. As we can see in Table 5, prediction using relatives' information (predicting in a pedigree) has better calibration accuracy O/E than prediction without relatives' information (prediction assuming the individuals are unrelated, i.e., each individual is assumed to be in a one-individual pedigree). In the training dataset, the prediction O/E for non-probands is 0.943 (95% CI: 0.722 to 1.231) when predicting with relatives' information, but is 1.666 (95% CI: 1.276 to 2.176) when predicting without relatives' information, indicating that predicting without using relatives' information significantly underestimates the BE risk ($P = 1.48\times10^{-4}$). In the validation dataset, although O/E for prediction with relatives' information is 1.378 (95% CI: 0.931 to 2.039), the overestimate of BE risk is not significant (P = 0.108); while O/E for prediction without relatives' information is 2.564 (95% CI: 1.732 to 3.794), suggesting significant underestimation of BE risk when predicting without family information ($P = 1.05\times10^{-6}$). Prediction with relatives' information also had better discrimination accuracy, AUC, than prediction without relatives' information in the training dataset: for non-probands, AUC was 0.753 vs. 0.741 respectively for prediction with and without relatives' information. However, in the validation dataset, the change in AUC between prediction with and without family information was very small (0.803 vs. 0.806). Our results indicate that, on average, prediction in families has better prediction accuracy than prediction without relatives' information; family information improves the prediction accuracy.

Variances due to genetic factors and to environmental/demographic/clinical factors

After making mean $(R(G_i,x_i))$ = mean $(R(\overline{G},x_i))$ = mean $(R(G_i,\overline{x}))$, the sum of squares (ss) due to genetic factors, the other factors, and the total ss were respectively $ss_{G_i}=\Sigma_i\left(R\left(G_i,x_i\right)-R\left(\overline{G},x_i\right)\right)^2=2.485$; $ss_{x_i}=\Sigma_i(R\left(G_i,x_i\right)-R\left(G_i,\overline{x}\right))^2=8.842$; $ss_{\text{total}}=\Sigma_i\left(R\left(G_i,x_i\right)-R\left(\overline{G},\overline{x}\right)\right)^2=9.775$. Thus, because $ss_{G_i}/(ss_{G_i}+ ss_{x_i} = 11.327 > 9.775 = ss_{total}$, there appeared to be no interaction between G_i and x_i. Estimating $ss_{G_i}/(ss_{G_i}+ ss_{x_i}) = 21.9\%$, the genetic factors contributed about 22% and the other factors about 78% of the total variance in the 787 singly ascertained pedigrees.

Discussion

In this study, we developed a BETRNet model to predict absolute risk of Barrett's esophagus in families by incorporating into the model both clinical and genetic factors. Based on the values of multiple clinical variables for family members, our model can predict BE risk for anyone with family members known to have had, or not have had, BE. It can also predict for unrelated individuals without any relatives' information. Our results indicate that the family information helps to predict BE risk, and predicting in families improves both the prediction calibration and the discrimination accuracy. Our prediction model will lead to effective identification of high risk individuals for BE screening and surveillance, consequently allowing intervention at an early stage leading to a reduction in mortality from esophageal adenocarcinoma.

Compared with other BE and EAC prediction models (13-17), our prediction model has the advantage of incorporating information on relatives. However, if no information on relatives is available, our prediction model can still predict for such "unrelated" individuals in the same way that the other prediction models do. Moreover, if BE causal genes are discovered, it will easily allow incorporation of such causal genes into the model. In addition, the model can adjust for single ascertainment for predicting in a family.

Our definition of “affected” status has included BE and its associated cancers because they are epidemiologically similar and there is strong evidence that nearly all EACs and a substantial proportion of junctional cancers arise in Barrett’s epithelium (4,5,29). Some of the cancers included in this prediction model may not have arisen in BE. However, because the prevalence of BE is much higher than that of cancer, it is unlikely that this misclassification will affect the prediction model. It is important to note that this model should be used only for families in which BE is rigorously defined by confirming the presence of intestinal metaplasia in biopsies obtained from visible columnar mucosa in the tubular esophagus. The prediction model should not be used in families where the diagnosis of BE has not been confirmed in family members.

Clinical data were collected on both affected and unaffected family members using a FBE questionnaire (20) based on the Mayo GERQ (19). However, because the major goal in collecting the pedigrees used in this study was to discover susceptibility genes for BE and EAC, family members affected with BE, EAC, or gastroesophageal junctional adenocarcinomas were much more likely to participate than unaffected family members. The large proportion of missing data on the covariates among the unaffected individuals in the training dataset could result in biased estimates of the genotypic susceptibilities, or of the allele frequencies for the genetic trait locus, but any such bias could not be identified by our validation data, which had a similar missing-value problem. In addition, our result showed that in the training data the calibration assessment (O/E) is close to 1, while in the validation dataset our prediction underestimated BE risk, which may indicate the difference in ascertainment between the training and validation datasets. However, the possible bias should not be serious because the estimated population prevalence of Barrett’s esophagus from our prediction model is 3.0%, close to the reported population prevalence (7). Moreover, the missing data resulted in loss of family information (Supplementary Table S4); 71% of the pedigrees in the training set had only one informative individual, which could be the reason the prediction with relatives’ information did not improve much on discrimination accuracy (AUC) compared with prediction without such information. Despite the missing data, our study shows that our model, which included family information, was able to predict risk of BE.

Note that although we modeled the familial effect as a latent variable in the form of a diallelic susceptibility locus, this does not imply that this is the true genetic model; the true underlying model is undoubtedly more complex, in terms of both genetics and the environment. What we have demonstrated in this paper is the importance of including this familial effect into the prediction model. As shown in the example scenarios in Table 4, for a male at age 50 with an unaffected brother, his risk is 3.2%; but if he has one affected brother, his risk is almost tripled (9.1%); furthermore, if he has two affected brothers, his risk is further tripled (26.6%). These increases are attributable to the genetically modeled familial effect. What may appear to be a peculiarity in Table 4 is that in some instances the probability of being affected is seen to be higher at age 50 than at age 70. This arises because the sibs are assumed to be about the same age as the family member. Consider, for example, a male at age 50 with two affected brothers, for whom the predicted risk is 26.6%; whereas (in the same line of the table) at age 70 his risk is only 20.1%. In the former case we know that the two brothers were already affected by ages 52 and 54, whereas in the latter case they may have become affected when up to 20 years older, implying the possibility of a lower genetic predisposition in the family.

In a future study, as more BE pedigrees and more informative individuals in pedigrees are prospectively and rigorously collected, we could improve our prediction model with updated data, and further validate it against external populations. Extensive simulation studies could help find more appropriate strategies to deal with complex ascertainments and missing data, to improve the estimation of clinical and genetic parameters in the prediction model. Simulation studies could also help determine the value of risk prediction for different genetic models, for example, if the disease is transmitted largely recessively. Furthermore, a web-based application (30) has been developed to allow risk calculations with subsequent endoscopy data, so that the model could be monitored over time to improve its accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support: X. Sun was supported in part by U54 CA163060 grant from the National Cancer Institute; R. Elston was supported in part by U54 CA163060 grant from the National Cancer Institute and NRF-2014S1A2A2028559 grant from the Korean Government; J.S. Barnholtz-Sloan, N.J. Shaheen, P.G. Iyer, J.A. Abrams, J.E. Willis were supported in part by U54 CA163060; S. Markowitz was supported in part by grant P50 CA150964 from the National Cancer Institute; A. Chandar, J.M. Warfe, W. Brock, A. Chak were supported in part by grant U54 CA163060 from the National Cancer Institute.

References

1. Pohl H, Welch HG. The role of overdiagnosis and reclassification in the marked increase of esophageal adenocarcinoma incidence. J Natl Cancer Inst. 2005; 97:142–6. [PubMed: 15657344]
2. Brown LM, Devesa SS, Chow WH. Incidence of adenocarcinoma of the esophagus among white Americans by sex, stage, and age. J Natl Cancer Inst. 2008; 100:1184–7. [PubMed: 18695138]
3. Holmes RS, Vaughan TL. Epidemiology and pathogenesis of esophageal cancer. Semin Radiat Oncol. 2007; 17:2–9. [PubMed: 17185192]
4. Solaymani-Dodaran M, Logan RF, West J, Card T, Coupland C. Risk of oesophageal cancer in Barrett's esophagus and gastro-oesophageal reflux. Gut. 2004; 53:1070–4. [PubMed: 15247170]
5. Hvid-Jensen F, Pedersen L, Drewes AM, Sørensen HT, Funch-Jensen P. Incidence of adenocarcinoma among patients with Barrett's esophagus. N Engl J Med. 2011; 365:1375–83. [PubMed: 21995385]
6. Chak A, Ochs-Balcom H, Falk G, Grady WM, Kinnard M, Willis JE, et al. Familiality in Barrett's esophagus, adenocarcinoma of the esophagus, and adenocarcinoma of the gastroesophageal junction. Cancer Epidemiol Biomarkers Prev. 2006; 15:1668–73. [PubMed: 16985029]
7. Ronkainen J, Aro P, Storskrubb T, Johansson SE, Lind T, Bolling-Sternevald E, et al. Prevalence of Barrett's esophagus in the general population: an endoscopic study. Gastroenterology. 2005; 129:1825–31. [PubMed: 16344051]
8. Zagari RM, Fuccio L, Wallander MA, Johansson S, Fiocca R, Casanova S, et al. Gastro-oesophageal reflux symptoms, oesophagitis and Barrett's oesophagus in the general population: the Loiano-Monghidoro study. Gut. 2008; 57:1354–9. [PubMed: 18424568]

9. Juhasz A, Mittal SK, Lee TH, Deng C, Chak A, Lynch HT. Prevalence of Barrett's Esophagus in first degree relatives of patients with esophageal adenocarcinoma. J Clin Gastroenterol. 2011; 45:867–71. [PubMed: 21617543]

10. Ward EM, Wolfsen HC, Achem SR, Loeb DS, Krishna M, Hemminger LL, et al. Barrett's esophagus is common in older men and women undergoing screening colonoscopy regardless of reflux symptoms. Am J Gastroenterol. 2006; 101:12–7. [PubMed: 16405528]

11. Gerson LB, Shetler K. Triadafilopoulos G. Prevalence of Barrett's esophagus in asymptomatic individuals. Gastroenterology. 2002; 123:461–7. [PubMed: 12145799]

12. Juhasz A, Mittal SK, Lee TH, Deng C, Chak A, Lynch HT. Prevalence of Barrett esophagus in first-degree relatives of patients with esophageal adenocarcinoma. J Clin Gastroenterol. 2011; 45:867–71. [PubMed: 21617543]

13. Thrift AP, Kendall BJ, Pandeya N, Whiteman DC. A model to determine absolute risk for esophageal adenocarcinoma. Clin Gastroenterol Hepatol. 2013; 11:138–44. [PubMed: 23103823]

14. Thrift AP, Kendall BJ, Pandeya N, Vaughan TL, Whiteman DC. Study of Digestive Health. A clinical risk prediction model for Barrett esophagus. Cancer Prev Res. 2012; 5:1115–23.

15. Thrift AP, Kramer JR, Qureshi Z, Richardson PA, El-Serag HB. Age at onset of GERD symptoms predicts risk of Barrett's esophagus. Am J Gastroenterol. 2013; 108:915–22. [PubMed: 23567358]

16. Rubenstein JH, Morgenstern H, Appelman H, Scheiman J, Schoenfeld P, McMahon LF Jr, et al. Prediction of Barrett's esophagus among men. Am J Gastroenterol. 2013; 108:353–62. [PubMed: 23318485]

17. Bhat S, Coleman HG, Yousef F, Johnston BT, McManus DT, Gavin AT, et al. Risk of malignant progression in Barrett's esophagus patients: results from a large population-based study. J Natl Cancer Inst. 2011; 103:1049–57. [PubMed: 21680910]

18. Ek WE, Levine DM, D'Amato M, Pedersen NL, Magnusson PK, Bresso F, et al. Germline genetic contributions to risk for esophageal adenocarcinoma, Barrett's esophagus, and gastroesophageal reflux. J Natl Cancer Inst. 2013; 105:1711–8. [PubMed: 24168968]

19. Locke GR, Talley NJ, Weaver AL, Zinsmeister AR. A new questionnaire for gastroesophageal reflux disease. Mayo Clin Proc. 1994; 69:539–47. [PubMed: 8189759]

20. Chak A, Lee T, Kinnard MF, Brock W, Faulx A, Willis J, et al. Familial aggregation of Barrett's oesophagus, oesophageal adenocarcinoma, and oesophagogastric junctional adenocarcinoma in Caucasian adults. Gut. 2002; 51:323–8. [PubMed: 12171951]

21. Chak A, Faulx A, Kinnard M, Brock W, Willis J, Wiesner GL, et al. Identification of Barrett's esophagus in relatives by endoscopic screening. Am J Gastroenterol. 2004; 99:2107–14. [PubMed: 15554988]

22. Karunaratne PM, Elston RC. A multivariate logistic model (MLM) for analyzing binary family data. Am J Med Genet. 1998; 76:428–37. [PubMed: 9556304]

23. Statistical Analysis for Genetic Epidemiology (S.A.G.E.) Version 6.3.

24. Schnell AH, Elston RC, Hull PR, Lane PR. Major gene segregation of actinic prurigo among North American Indians in Saskatchewan. Am J Med Genet. 2000; 92:212–9. [PubMed: 10817657]

25. Sun X, Vengoechea J, Elston R, Chen Y, Amos CI, Armstrong G, et al. A variable age of onset segregation model for linkage analysis, with correction for ascertainment, applied to glioma. Cancer Epidemiol Biomarkers Prev. 2012; 21:2242–51. [PubMed: 22962404]

26. Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, et al. Projecting individualized absolute invasive breast cancer risk in African American women. J Natl Cancer Inst. 2007; 99:1782–92. [PubMed: 18042936]

27. Matsuno RK, Costantino JP, Ziegler RG, Anderson GL, Li H, Pee D, et al. Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. J Natl Cancer Inst. 2011; 103:951–61. [PubMed: 21562243]

28. McCulloch CE, Neuhaus JM. Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. Statist Sci. 2011; 26:388–402.

29. Cameron AJ, Ott BJ, Payne WS. The incidence of adenocarcinoma in columnar-lined (Barrett's) esophagus. N Engl J Med. 1985; 313:857–58. [PubMed: 4033716]

30. http://som-apps.case.edu/betrnet/predication

Impact

Our prediction model will shed light on effectively identifying high risk individuals for BE screening and surveillance, consequently allowing intervention at an early stage and reducing mortality from esophageal adenocarcinoma.

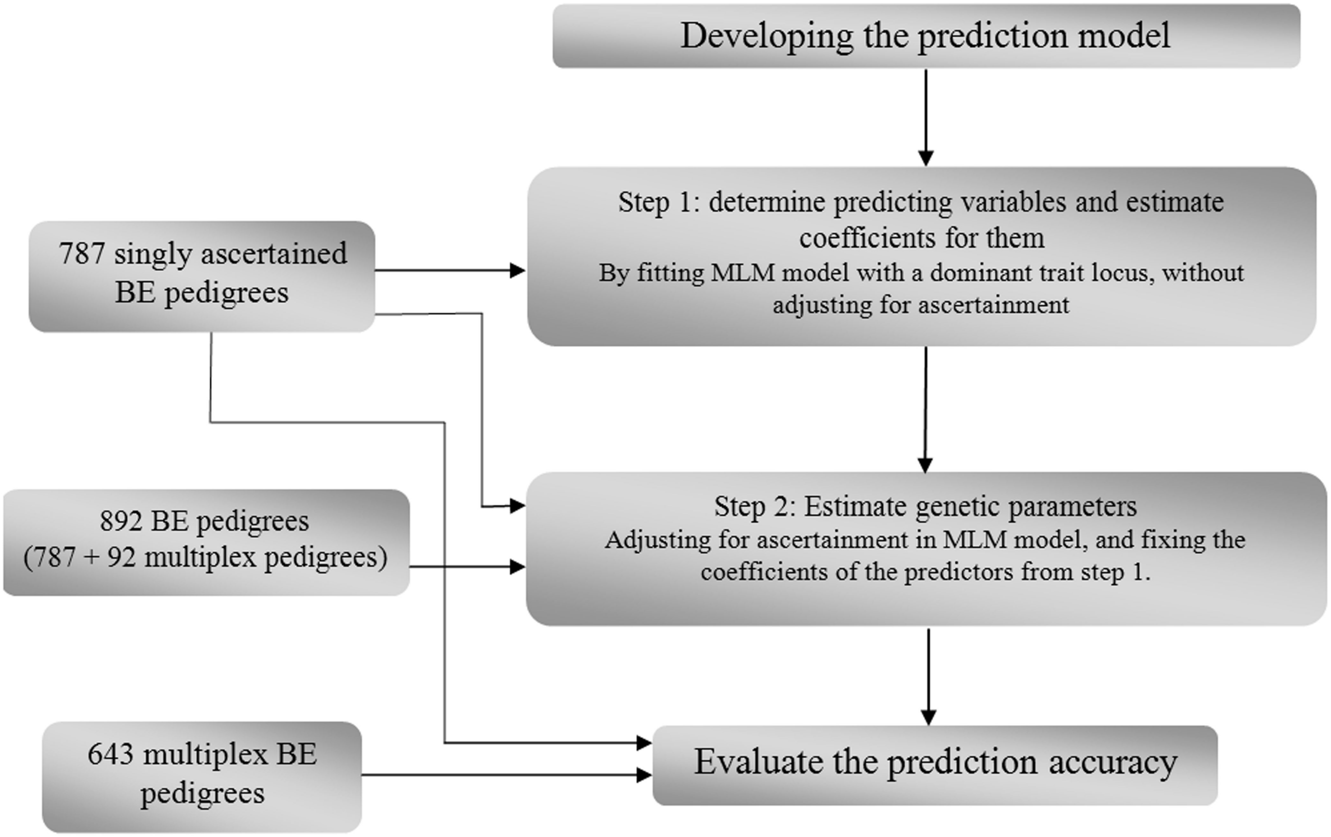
Developing the prediction model
787 singly ascertained BE pedigrees
Step 1: determine predicting variables and estimate coefficients for them
By fitting MLM model with a dominant trait locus, without adjusting for ascertainment
892 BE pedigrees (787 + 92 multiplex pedigrees)
Step 2: Estimate genetic parameters
Adjusting for ascertainment in MLM model, and fixing the coefficients of the predictors from step 1.
643 multiplex BE pedigrees
Evaluate the prediction accuracy

Figure 1.

Flowchart of developing the prediction model

Table 1

Demographic Characteristics of the pedigree members used for prediction and validation, with numbers of individuals (%)

	Training		Validation
	689 pedigrees[a]	**743 pedigrees**[b]	**248 pedigrees**[c]
Affected	716 (61.2)	773 (61.9)	237 (56.4)
Male	576 (80.4)	620 (80.2)	198 (83.5)
Female	140 (19.6)	153 (19.8)	39 (16.5)
Subcategories of diagnosis:			
BE	402 (56.1)	434 (56.1)	97 (40.9)
SSBE	103 (14.4)	192 (24.8)	62 (26.2)
ECA	177 (24.7)	34 (4.4)	68 (28.7)
JCA	34 (4.7)	113 (14.6)	10 (4.2)
Education level:			
< High school	69(9.6)	73(9.8)	16(6.8)
high school	353(49.3)	389(52.4)	117(49.4)
college and beyond	294(41.1)	311(41.9)	104(43.9)
Heartburn frequency:			
None	172(24.0)	172(23.1)	63(26.6)
≤ once a week	319(44.6)	349(47.0)	105(44.3)
several times a week or everyday	225(31.4)	252(33.9)	69(29.1)
Regurgitation frequency:			
None	195(27.2)	196(26.4)	75(31.6)
≤ once a month	295(41.2)	326(43.9)	99(41.8)
weekly or more	226(31.6)	251(33.8)	63(26.6)
Unaffected	454 (38.8)	476 (38.1)	183 (43.6)
Male	197 (43.4)	206 (43.3)	71 (50.4)
Female	257 (56.6)	270 (56.7)	112 (48.7)
Education level:			
< High school	21(4.6)	21(4.4)	5(2.7)
high school	168(37.0)	177(37.2)	57(31.1)
college and beyond	265(58.4)	278(58.4)	121(66.1)
Heartburn frequency:			
None	126(27.8)	128(26.9)	68(37.2)
≤ once a week	214(47.1)	224(47.1)	85(46.4)
several times a week or everyday	114(25.1)	124(26.1)	30(16.4)
Regurgitation frequency:			
None	183(40.3)	185(38.9)	81(44.3)
≤ once a month	200(44.1)	206(43.3)	67(36.6)
weekly or more	71(15.6)	85(17.9)	35(19.1)
Total	1170	1249	420

	Training		Validation
	689 pedigrees[a]	**743 pedigrees**[b]	**248 pedigrees**[c]
Male	773 (66.1)	826 (66.1)	269 (64.0)
Female	397 (33.9)	423 (33.9)	151 (36.0)
Education level:			
< High school	90(7.7)	94(7.5)	21(5.0)
high school	521(44.5)	566(45.3)	174(41.4)
college and beyond	559(47.8)	589(47.2)	225(53.6)
Heartburn frequency:			
None	298(25.5)	300(24.0)	131(31.2)
≤ once a week	533(45.6)	573(45.9)	190(45.2)
several times a week or everyday	339(29.0)	376(30.1)	99(23.6)
Regurgitation frequency:			
None	378(32.3)	381(30.5)	156(37.1)
≤ once a month	495(42.3)	532(42.6)	166(39.5)
weekly or more	297(25.4)	336(26.9)	98(23.3)

[a] A subset of 787 singly ascertained pedigrees with members who are informative on all the clinical variables

[b] A subset of 879 pedigrees (787pedigrees+92multiplex pedigrees) with members who are informative on all the clinical variables

[c] A subset of 643 validation pedigrees with members who are informative on all the clinical variables

Table 2

Estimated effects of covariates using the data on 787 singly ascertained pedigrees

Covariates	**Estimate**	**S.E.**	**OR**	**95% CI of OR**
Sex	−2.101	0.257	0.122	(0.074, 0.202)
Parent	1.015	0.216	2.759	(1.807, 4.214)
Log(age)	3.057	0.416	21.264	(9.409, 48.055)
Years of Smoking	0.021	0.0007	1.021	(1.020, 1.023)
HeartburnFreq	−0.245	0.156	0.783	(0.577, 1.063)
RegurgFreq	0.879	0.167	2.408	(1.736, 3.341)
Education	−0.418	0.166	0.658	(0.476, 0.912)
Use of acid suppressant	1.506	0.304	4.509	(2.485, 8.181)

Table 3

Estimated population susceptibility parameters and allele frequency from 787 BE pedigrees and 879 BE pedigrees, on adjusting for ascertainment

	Model 1[a]				Model 2[b]			
Parameters	**Estimate**	**S.E.**	**OR**	**95% CI of OR**	**Estimate**	**S.E.**	**OR**	**95% CI of OR**
Susceptibility[c] AA, AB	−12.961	0.628	81.70	(13.73, 486.17)	*−12.715*	*0.871*	*90.56*	*(5.81, 1410.79)*
Susceptibility BB[c]	−17.075	0.420	1		*−16.854*	*0.363*	*1*	
Frequency of A	0.027	0.016			*0.021*	*0.142*		
Prevalence (non-parent at age 50)	0.030				*0.030*			

Note: OR: Odds Ratio. S.E.: Standard error. CI: Confidence Interval.

[a] Adjusting for single ascertainment using 787 pedigrees, finding prevalence = 3.0%

[b] Estimated with 879 pedigrees by adjusting for single ascertainment and using prevalence constraint from model 1 (3.0%)

[c] The estimates are on the logistic scale

Table 4

Predicted probability (%) of a family member having BE by (Model 1, *Model 2*) using the parameter values in tables 2 and 3, the values of covariates other than sex, parent and log(age) are at the mean values shown in Supplementary Table S3. Parents of the family members are assumed to be unaffected with BE

1st sib 4 years older than the Family Member	**2nd sib 2 years older than the Family Member**	**Sex, age of Family Member**			
		Male, 50	**Female, 50**	**Male, 70**	**Female, 70**
Male without BE	N/A	3.2 *3.7*	0.5 *0.5*	7.7 *9.2*	1.1 *1.3*
Female without BE	N/A	3.4 *3.9*	0.5 *0.6*	7.8 *9.3*	1.2 *1.3*
Male with BE	N/A	9.1 *7.6*	2.1 *1.8*	10.3 *10.7*	2.4 *2.1*
Female with BE	N/A	14.6 *12.5*	3.7 *3.3*	15.8 *14.5*	5.0 *4.2*
Male with BE	Male with BE	26.6 *23.4*	7.1 *6.8*	20.1 *16.2*	7.1 *5.1*
Male with BE	Male without BE	7.2 *6.2*	1.6 *1.3*	9.1 *10.1*	1.8 *1.7*
Female with BE	Female with BE	32.9 *34.3*	8.9 *10.3*	41.0 *38.9*	17.1 *17.5*
Female with BE	Female without BE	13.8 *11.7*	3.5 *3.1*	14.4 *13.3*	4.3 *3.5*

NOTE: N/A: no affection status known for a second sibling

Table 5

Evaluating the prediction performance using the training and validation datasets

	Training Dataset [b]		Validation Dataset [c]	
Prediction in pedigrees	**All individuals [a]**	**Non-probands**	**All individuals[a]**	**Non-probands**
O	716	54	237	25
E	719.260	57.260	230.148	18.148
O/E (95% CI)	0.995 (0.925 to 1.071)	0.943 (0.722 to 1.231)	1.030 (0.907 to 1.170)	1.378 (0.931 to 2.039)
P value	0.903	0.667	0.652	0.108
AUC	0.981	0.753	0.981	0.803
Prediction assuming individuals are unrelated	All individuals[a]	Non-probands	All individuals[a]	Non-probands
O	716	54	237	25
E	694.404	32.404	221.752	9.752
O/E(95% CI)	1.031 (0.958 to 1.109)	1.666 (1.276 to 2.176)	1.069 (0.941 to 1.214)	2.564 (1.732 to 3.794)
P value	0.412	1.48×10^{-4}	0.306	1.05×10^{-6}
AUC	0.980	0.741	0.981	0.806

[a] Prediction for the probands using formula (5)

[b] the training dataset comprises 787 singly ascertained BE pedigrees

[c] the validation dataset comprises 643 BE pedigrees