**Supplementary Methods**

*Imputation of age*

We imputed age (for the missing 40%) using the family structure according to the method of Schnell et al (1). Moreover, observing that the sum of the known of date of birth (DOB) and age at examination are mostly between 1999 and 2010 (Supplementary Figure S1), 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

We used the leave-one-out cross validation to evaluate the age imputation. For 1512 individuals whose ages are known, we made one of the 1512 individuals’ age (and date of birth) missing each time, and used the other individuals to impute the missing information of age and date of birth; we did this 1512 times for the 1512 individuals. In this way, we evaluated the imputing accuracy under a similar missing rate and pedigree structure information of our data. The imputed ages are highly consistent with the true age (Supplementary Figure S2) with correlation coefficient = 0.91, and mean±SD of the absolute prediction error (absolute value of true age – imputed age) = 4.40±5.19.

*Estimating the parameters of the prediction model*

We estimated the prediction model using a pedigree likelihood based on a multivariate logistic model (MLM) (20) using the training dataset of 787 singly ascertained pedigrees. The prediction model includes segregation at a major trait locus, and other clinical predictors, such as demographic factors (age, sex, education level, parental status), and environmental factors (smoking, alcohol consumption, use of acid suppressant medications, gastroesophageal reflux symptoms such as GERD and heartburn). The prediction model was obtained from a generalized and modified version of the SEGREG program in the S.A.G.E. package (2). In SEGREG, although there is no restriction on the size of the family, the family is assumed to be non-inbred and to contain no children of consanguineous spouse pairs. In estimating the parameters of the MLM model, we assumed no residual associations between family members because of the theoretical difficulty this entails (3). Note that the MLM (3) is a **multivariate** logistic regression model that assumes the familial correlation arise, in addition to segregation at a single genetic locus, from a polygenic component – which we did not included in any of our analyses. If no single major locus is included in the likelihood, the model is identical to ordinary **multivariable** logistic regression.

The training dataset (787 pedigrees) has missing values on the predictor variables for most of the unaffected individuals. This should not influence the estimation of regression coefficients of 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). Because including ascertainment in the likelihood for the BE pedigrees would influence both the genetic parameters and the regression coefficients, we estimated the coefficients of the prediction covariates and the genetic parameters of the model in two separate steps.

In the first step, we estimated the predictor coefficients, and selected which predictors to keep in the model, by fitting MLM models without adjusting for ascertainment. We selected the predictors by separately fitting MLM models with one and two genetic susceptibilities of a latent genetic locus, starting with a full model that included the all 12 reported predictors (4-8) and then stepwise eliminating predictors on the basis of likelihood ratio tests and Akaike’s AICs (supplementary Table S2). We finally determined 8 covariates in the model (Supplementary Table S3), among which three are categorical variables, which are respectively education level, heartburn frequency, and regurgitation frequency. They were coded as in shown Supplementary Table S3 to make them jointly linear in the susceptibility (Supplementary Figure S3). For the continuous clinical variables, we check the linearity in the logit scale (Supplementary Figure S4) and did the curvature test for each of them, finding no significant nonlinearity*.* The coefficients of the eight predictor covariates were estimated in a MLM model that assumed a mixture of two genetic susceptibilities determined by a dominant model. In this way, the regression coefficients (Table 2) should be approximately unbiased provided only that the linear logistic model is appropriate for the fixed effects (9).

In the second step, using the 787 singly ascertained pedigrees, we re-estimated the genetic parameters with the same MLM model but adjusting for single ascertainment while fixing the coefficients of the covariate coefficients that were estimated in the first step.

*The probability of being affected for a proband: proof of formula (5)*

For a proband in a singly ascertained pedigree, the probability of being affected is

$$p\_{ic}= prob\left(other family members^{'}information, and the proband's information and affected status\right)$$

$$=\frac{P\left(the proband's information and affected status\right)}{\begin{array}{c}P\left(the proband^{'}s information and affected status\right)+\\ P\left(the proband's information and affected status\right)\end{array}}$$

$=\left\{\begin{array}{c}\frac{P\left(the proband is affected\right)}{\begin{array}{c}P\left(the proband is affected\right)+\\P(the proband is unaffected,otherfamily members^{'}information|the proband is affected) \end{array}}, \& if the proband is affected\\\frac{P\left(the proband is unaffected\right)}{\begin{array}{c}P\left(the proband is unaffected\right)+\\P\left(the proband is unaffected\right) \end{array}}, \& if the proband is unaffected\end{array}\right.$

$=\left\{\begin{array}{c}\frac{P\left(the proband is affected\right)}{P\left(the proband is affected\right)+0}, \& if the proband is affected\\\frac{0}{0+P\left(the proband is unaffected\right)}, \& if the proband is unaffected\end{array}\right.$

$=\left\{\begin{array}{c}1, \& if the proband is affected\\0, \& if the proband is unaffected\end{array}\right.$

References

1. 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.
2. Statistical Analysis for Genetic Epidemiology (S.A.G.E.) Version 6.3.
3. Karunaratne PM, Elston RC. A multivariate logistic model (MLM) for analyzing binary family data. Am J Med Genet 1998;76:428-37.
4. Thrift AP, Kendall BJ, Pandeya N, Whiteman DC. A model to determine absolute risk for esophageal adenocarcinoma. Clin Gastroenterol Hepatol 2013;11:138-44.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.

Supplementary Table S1 Missing rates on the 8 covariates in the training and validation datasets

|  |  |  |
| --- | --- | --- |
|  | Training data | Validation Data |
|  | BE unknown |  BE Unaffected | BE Affected | total | BE unknown | BE Unaffected | BE Affected | total |
| Missing  | 64 | 7049 | 191 | 7304 | 0 | 2692 | 230 | 2922 |
| None missing | 0 | 454 | 716 | 1170 | 0 | 183 | 237 | 420 |
| Sum  | 64 | 7503 | 907 | 8474 | 0 | 2875 | 467 | 3342 |
| Missing rate | 1 | 0.939 | 0.211 | 0.862 | 0 | 0.936 | 0.493 | 0.874 |

Supplementary Table S2 Stepwise removal of covariates in the MLM model on the basis of LRT tests and AIC

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | The multivariate logistic model with one susceptibility | The multivariate logistic model with two susceptibilities |
| Stepwise removal of covariates | Number of parameters | -2ln(L) | Akaike's AIC | P value a | P value b | -2ln(L) | Akaike's AIC | P value a | P value b |
| Full model c | 13 | 765.72 | 793.73 |  |  | 749.42 | 781.42 |  |  |
| - alcohol | 12 | 766.38 | 792.38 | 0.42 | 0.42 | 750.55 | 780.55 | 0.29 | 0.29 |
| - age of onset of heartburn | 11 | 767.65 | 791.65 | 0.38 | 0.26 | 751.66 | 779.66 | 0.33 | 0.29 |
| - age of onset of regurgitation | 10 | 767.95 | 789.95 | 0.53 | 0.58 | 752.39 | 778.39 | 0.40 | 0.39 |
| - BMI | 9 | 770.71 | 790.71 | 0.29 | 0.10 | 755.27 | 779.27 | 0.21 | 0.09 |
| - packs per day | 8 | 771.50 | 789.50 | 0.33 | 0.37 | 757.32 | 779.32 | 0.16 | 0.15 |

a LRT test comparing with the full model

b LRT test comparing with the model previous to it – the one without removal of that covariate

c Thirteen predicting covariates were included in the full model, which were 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

Supplementary Table S3. Description of the eight predictor covariates

|  |  |  |  |
| --- | --- | --- | --- |
| covariate | Type | Coding | Mean in the datab |
| Sex | Binary | Male (0), female (1) | 0.49 |
| Parent | Binary | Parent (1), non-parent (0) | 0.46 |
| ln(age) | Continuous |  | 3.95 (corresponds to 52.06 years) |
| Years of Smoking | Continuous |  | 15.41 |
| Heartburn frequency | Categoricala | None (0), less than or equal to once a week (1), several times a week or everyday (2) | 1.00 |
| Regurgitation frequency | Categoricala  | None (0), less than or equal to once a month (1), weekly or more (2) | 0.90 |
| Education level | Categoricala | < High school (0), high school (1), college and beyond (2) | 1.40 |
| Acid suppressant | Binary | Yes (1), no (0) | 0.85 |

a The linearity of these coding schemes is illustrated in figure S3

b The training data of 787 pedigrees

Supplementary Table S4 Number of informative pedigrees without missing covariate data

|  |  |
| --- | --- |
| Number of informative individuals in pedigree | Number of Pedigrees |
| Training data | Validation Data |
| 4+ | 57 | 23 |
| 3 | 50 | 16 |
| 2 | 96 | 38 |
| 1 | 486 | 171 |
| Total | 689 | 248 |