Published in final edited form as:
Ann Epidemiol. 2014 January ; 24(1): 50–57. doi:10.1016/j.annepidem.2013.10.009.

Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancer: classification tree analysis

Stephanie A Navarro Silvera, PhD,
Montclair State University, Department of Health and Nutrition Sciences, Montclair, NJ

Susan T Mayne, PhD,
Yale School of Public Health and Yale Cancer Center, New Haven CT

Marilie D Gammon, PhD,
University of North Carolina, Department of Epidemiology, Chapel Hill, NC

Thomas L Vaughan, MD, MPH,
Fred Hutchinson Cancer Research Center, Program in Epidemiology, and University of Washington, School of Public Health and Community Medicine, Department of Epidemiology, Seattle, WA

Wong-Ho Chow, PhD,
MD Anderson Cancer Center, Houston, TX

Joel A Dubin, PhD,
University of Waterloo, Department of Statistics & Actuarial Science, and School of Public Health and Health Systems, Waterloo, Ontario

Robert Dubrow, MD, PhD,
Yale School of Public Health and Yale Cancer Center, New Haven CT

Janet L Stanford, PhD, MPH,
Fred Hutchinson Cancer Research Center, Program in Epidemiology, and University of Washington, School of Public Health and Community Medicine, Department of Epidemiology, Seattle, WA

A Brian West, MD,
Yale School of Medicine and Yale Cancer Center, New Haven, CT

Heidrun Rotterdam, MD,
Columbia University, Department of Pathology, New York, NY

William J Blot, PhD, and

Please address correspondence and reprint requests to Stephanie A. Navarro Silvera, Montclair State University, Department of Health and Nutrition Sciences, 1 Normal Ave, Montclair, NJ 07043. Telephone 973-655-2125; FAX 973-655-5461 silveras@mail.montclair.edu.

No authors had any conflicts of interest.

International Epidemiology Institute, Rockville, MD

Harvey A Risch, MD, PhD.
Yale School of Public Health and Yale Cancer Center, New Haven CT

Abstract

Purpose—Although risk factors for squamous cell carcinoma of the esophagus (SCC) and adenocarcinomas of the esophagus (EA), gastric cardia (GC) and other (non-cardia) gastric sites (OG) have been identified, little is known about interactions among risk factors. We sought to examine interactions of diet, other lifestyle, and medical factors with risks of subtypes of esophageal and gastric cancer.

Methods—We used classification tree analysis to analyze data from a population-based case-control study (1,095 cases, 687 controls) conducted in Connecticut, New Jersey, and western Washington State.

Results—Frequency of reported gastroesophageal reflux (GERD) symptoms was the most important risk stratification factor for EA, GC, and OG, with dietary factors (EA, OG), smoking (EA, GC), wine intake (GC, OG), age (OG), and income (OG) appearing to modify risk of these cancer sites. For SCC, smoking was the most important risk stratification factor, with GERD, income, race, non-citrus fruit, and energy intakes further modifying risk.

Conclusion—Various combinations of risk factors appear to interact to affect risk of each cancer subtype. Replication of these data-mining analyses are required before suggesting causal pathways; however, the classification tree results are useful in partitioning risk and mapping multi-level interactions among risk variables.

Keywords

esophageal adenocarcinoma; gastric cardia adenocarcinoma; diet; gastroesophageal reflux; classification tree; CART

Introduction

Adenocarcinoma of the esophagus and, to a lesser extent of the gastric cardia, has been increasing in incidence [1–3]. It has been reported that the annual incidence of esophageal adenocarcinoma increased 350% between 1976 and 1994 [2], and Holmes and colleagues reported a 6.75-fold increase among white men between 1973 and 2002 [4]. Increases have been found in the United States [1–5], Great Britain, Australia, The Netherlands, Denmark, and other western nations [6]. Etiologic studies in the United States and elsewhere have identified several important risk factors, including cigarette smoking [7–10], obesity [11–14], and gastroesophageal reflux (GERD) [15–17]. It has also been shown that *Helicobacter pylori* colonization may be protective for esophageal and gastric cardia adenocarcinoma, particularly so for CagA-positive strains [18–20].

Epidemiologic studies have reported that fruit and vegetable consumption may be inversely associated with risks of esophageal and gastric cancer without regard to subsite or histologic type [21]. A review of the literature conducted by Thrift et al. [22] indicated that a moderate

to substantially decreased risk of esophageal adenocarcinoma is associated with regular fruit and vegetable intake. There is limited evidence, however, examining the role of dietary factors on subtypes of these cancers in combination with other factors. We have previously reported significant inverse associations between intake of nutrients found primarily in plant-based foods and the risk of esophageal and gastric cardia adenocarcinomas [23, 24]. In addition, Steevens et al. [25] noted a statistically significant reduced risk of esophageal adenocarcinoma associated with raw vegetable consumption and a significant inverse association between brassica vegetables and gastric cardia adenocarcinoma among a cohort of Dutch men and women. However, in the AARP cohort, Freedman et al. found a significant inverse association between fruit intake and risk of esophageal squamous cell carcinoma, but not of esophageal adenocarcinoma [26]. While we observed a significant positive association between intake of meat and animal protein and risk of adenocarcinomas of the esophagus and gastric cardia [23, 24], Keszei and colleagues [27], in an analysis of data from The Netherlands Cohort Study, did not find any association between red or processed meat and esophageal or gastric adenocarcinomas. They did, however, find a significantly elevated risk of esophageal squamous cell carcinoma associated with both red and processed meats among men [27]. According to the World Cancer Research Fund expert panel report, the available evidence suggests a positive association between non-cardia gastric adenocarcinoma and nitrite-related foods in western countries and salted or preserved foods in Asian countries [21]. There is also evidence of an inverse association between dietary fiber intake and risks of adenocarcinoma of the esophagus [11, 23] and of the gastric cardia [23, 28].

Dietary behaviors are complex. For example, consumption of fruit and vegetables is associated both positively and negatively with consumption of other food groups [29]. In addition, dietary behaviors correlate with other health behaviors and demographic factors [30–32]. Recursive partitioning techniques, including classification trees, have been used as a means of examining the complex interactions or patterns of risk factors in a variety of diseases [33, 34] including colon [35] and lung [36] cancer. Classification tree analysis is agnostic in evaluating interactions that do not need to be pre-specified, in contrast to standard regression models, in which interactions are generally pre-specified [37]. Given that most cancers are multifactorial in nature, often involving combinations of both host and genetic factors in determining risk, classification tree models can give clues to important interactions by sorting through the complex, multi-level nature of risk factors associated with these cancers. Thus, the purpose of this analysis was to explore a variety of dietary and lifestyle variables as predictors of risk of subtypes of esophageal and gastric cancer, to understand better which of these correlated variables appears to be most important for risk stratification, and to examine multi-level interactions involving these same variables.

Subjects and methods

Study population

Subject recruitment and data collection methods have been reported in detail [9]. Briefly, a multi-center, population-based case-control study of esophageal adenocarcinoma, gastric cardia adenocarcinoma, esophageal squamous cell carcinoma, and adenocarcinoma of other

anatomic sites of the stomach was conducted. Because the original motivation for the study was to discover risk factors for esophageal adenocarcinoma and gastric cardia adenocarcinoma due to their increasing incidence, these cancer types were termed the target cases, whereas esophageal squamous cell carcinoma and non-cardia gastric adenocarcinoma were termed the comparison cases (with declining incidence rates). The project sought to enroll case groups of approximately equal size, using population-based tumor registries, within the entire state of Connecticut, a 15-county area of New Jersey, and a 3-county area of western Washington State, along with controls. Institutional review board approval was obtained from all participating centers and from the Connecticut Department of Public Health.

Participants were English-speaking men and women between 30–79 years of age who had been newly diagnosed between 1993 and 1995 with one of the cancer types noted above, along with population-based controls. Attempts were made to recruit all target cases and a random sample of comparison cases, frequency matched to target cases on 5-year age group, gender (in New Jersey and Washington State), race (in New Jersey) and geographic area. Cases were identified via rapid reporting systems in each of the three areas. Pathology reports were sought for all potentially eligible cases; slides and medical records were systematically reviewed by two study pathologists in order to determine final eligibilty.

Population based controls were randomly selected from the general population of each study area and were frequency matched to the expected distribution of target cases by 5-year age group, gender, and geographic area. Controls aged 30–64 years were recruited using Waksberg random digit dialing methods [38]; those aged 65–79 years were identified by random selection from rosters maintained by the Health Care Financing Administration.

In-person interviews were completed for 81% of eligible target cases, 74% of eligible comparison cases, and 70% of eligible controls. The mean time between case diagnosis and interview was 3.7 months. Thirty-four subjects were seriously ill and unable to complete the dietary portion of the questionnaire and were therefore were excluded from analyses. After additional exclusion of 23 persons because of implausible energy intakes (< 600 Kcal/d, n = 20 or >5000 Kcal/d, n = 3), the dietary analyses included 1,782 subjects: 687 controls, 282 cases with esophageal adenocarcinoma, 255 with gastric cardia adenocarcinoma, 206 with esophageal squamous cell carcinoma, and 352 with non-cardia gastric adenocarcinoma. Proxy interviews for deceased subjects were more common among cases (esophageal adenocarcinoma =31%, adenocarcinoma of the gastric cardia = 26%, esophageal squamous cell carcinoma = 35%, and non-cardia gastric adenocarcinoma = 30%) than among controls (3.4%).

Data collection

Study participants, or if necessary, close relatives who served as proxy respondents, were administered structured in-person questionnaires by trained interviewers after informed consent was obtained. The questionnaire contained questions on demographics, diet and lifestyle variables including tobacco, alcohol, other beverage use (e.g., coffee, tea), medical history, use of medications, and occupational history. A previously validated [39] food frequency questionnaire (although not validated for proxy respondents) developed by

investigators at the Fred Hutchinson Cancer Research Center (FHCRC) was adapted to assess usual consumption in the period 3–5 years before diagnosis (cases) or interview (controls). After food frequency questionnaire data were entered and verified, they were sent to the FHCRC for processing and were initially linked with the University of Minnesota Nutrition Coding Center Nutrient Data system for estimation of nutrient intake. Average daily intake of dietary nitrite (mg/day) was estimated separately, through software and databases developed and maintained by the authors [40]. Food subgroup variables were defined as previously described and represent servings per day [24]. The following medical and lifestyle variables were also included in the classification tree analysis: body mass index calculated using self-reported usual adult height and weight (continuous), average number of cigarettes smoked per day (in the year prior to interview; continuous), average number of drinks of beer, wine and liquor per day (each separately; in the year prior to interview; continuous), and reported frequency of GERD symptoms per year (continuous).

Statistical analysis

For each cancer subtype separately, we performed classification tree analysis [37, 41] to relate risk factors and their interactions with cancer risk. This method uses binary recursive partitioning whereby observations are repeatedly bifurcated into “nodes,” based on the considered risk factors for predicting outcome, which, in this study is case vs. control status. For each of the cancers under study, 35 risk variables, as described in Table 1, were considered for selection in the tree building process. Both continuous and categorical predictors were included in the models. Node splits in continuous variables can occur at any value and were not predetermined.

The root, or parent, node included all cases of a given cancer subtype and all controls. We used CART™ software to evaluate all potential variables, and selected the best variable on which to split. The goal was to create two child nodes that are as different as possible in terms of representation of case versus control risk. That is, ideally, one of the subsequent child nodes would contain 100% cases and the other would contain 100% controls, thus yielding maximum separation between the nodes and minimizing the variation within each node. In our analysis we utilized a numerical criterion called *Gini diversity*, available within the CART™ software, to maximize node separation [41]. This method of selecting the best splitter is repeated for each of the two child nodes resulting from the initial split. This process is further repeated until no additional beneficial splits can be made.

The resulting tree perfectly fits the data set. That is, it will pick up residual noise in the data analogous to saturated over-fitting in traditional regression models. While this tree will perfectly fit the data, it would likely result in high misclassification rates if applied to an independent data set. The next step in our modeling was therefore to *prune* the tree in order to obtain the smallest number of nodes without sacrificing goodness-of-fit [41]. Classification tree analysis uses a method of *cost-complexity* pruning in which child nodes are pruned and the predicted misclassification cost is compared to the change in tree complexity, yielding a number of smaller, nested trees [41]. The program then utilizes a cross-validation technique, splitting the dataset into 10 random but mutually exclusive subsets, each representing 10% of the study population, stratified by the outcome of interest.

The first time the model is run, the first subset of 10% is set aside as a test dataset and the remaining 90% is used to generate (i.e., train) a tree. This training tree is then compared to the test data in order to calculate the misclassification rate. This process is then repeated 10 times, with a different subset of the data reserved as the test dataset each time. Misclassification rates for each potential tree are calculated and used to select the optimal tree, which yields the most complex subtree while minimizing misclassification. These steps are described in detail in Breiman et al. [41] and are part of the CART™ software package used for this analysis.

The final tree for our analysis of each subtype of esophageal and gastric cancer included information on splitting nodes and terminal nodes. Each node contains information on number of controls (top number of uppermost box or oval) and number of cases (bottom number of uppermost box or oval). Terminal subsets are represented by square boxes and are identified by letter in the lower left corner. The proportion in the bottom right corner of each terminal subset gives the probability of being a case in that group. Given that the number of cases with a particular subtype was lower than the number of controls (e.g., 282 esophageal adenocarcinoma cases versus 687 controls in Figure 1 or approximately 29% cases), terminal subsets enriched in cases (e.g., > 0.29 for Figure 1) are considered higher risk nodes and have been bolded and italicized accordingly as has been done by others using this analytic approach [42].

We conducted sensitivity analyses in which we excluded proxy respondents.

Results

Esophageal adenocarcinoma

The classification tree for esophageal adenocarcinoma is presented in Figure 1 (n = 969). Six of the potential 35 variables from the candidate list (Table 1) remained in building the tree. The sample initially split on number of GERD symptoms/year. Those who reported experiencing symptoms 6.5 times/year or less comprised the lower risk group representing primarily controls (cases = 15/176 or 8% of the lower risk group). Those who reported experiencing symptoms more than 6.5 times/year were classified as the higher risk group and went on to further subdivisions. As an illustration of the classification of risk using this tree, among individuals suffering from GERD symptoms more than 6.5 times/year, the data split occurred such that those who consumed slightly more than ½ serving of red meat per day were classified as higher risk compared to those who consumed red meat less frequently. Additionally, this risk appeared to be modified by non-citrus fruit consumption, with those with lower intake (data split at ≤ 1.77 servings/day) more likely to be cases (52%, compared to the overall study population of 29% cases). Overall, the case misclassification rate was 31%.

Gastric cardia adenocarcinoma

The classification tree, based on the classification of gastric cardia adenocarcinoma case status, is presented in Figure 2 (n = 942). Four of the potential 35 variables from the candidate list (Table 1) remained in building this tree, and frequency of GERD symptoms

split four times within the tree, potentially suggesting a dose-response relationship between frequency of GERD symptoms and risk (Figure 2). The optimal split of the entire sample involved frequency of GERD symptoms. According to these analyses, individuals reporting experiencing infrequent symptoms (data split at 5.5 times/year or less) represent the lower risk of gastric cardia adenocarcinoma, primarily controls (e.g., 13/159 or 8% cases). Those who reported experiencing symptoms more than 5.5 times/year went on to further subdivisions. Based on this tree, individuals experiencing GERD symptoms at least daily (data split at ≥ 408 times/year) were among those at greatest risk of gastric cardia adenocarcinoma (e.g., 28/61 or 46% cases). After frequency of GERD symptoms, number of cigarettes smoked per day also entered the model, with those smoking somewhat more than half a pack per day (data split at ≥ 13.5 cigarettes per day) more likely to be cases than controls, though this risk appeared to be moderated somewhat by frequency of GERD symptoms. Overall, the case misclassification rate was 38%.

Esophageal squamous cell carcinoma

The classification tree, based on the classification of esophageal squamous cell carcinoma case status, is presented in Figure 3 (n = 893). Six of the potential 35 variables from the candidate list (Table 1) were selected in building this tree, including variables not selected in the other tumor site classification trees. The optimal split for this outcome involved number of cigarettes smoked/day, with values less than or equal to 25.5 (including never smokers, who made up 35.2% of controls and 9.9% of esophageal squamous cell carcinoma cases) indicating a lower risk group, and those with values greater than 25.5 representing a higher risk group. The lowest risk group was those who smoked 25.5 or fewer cigarettes/day who also reported experiencing low or no GERD symptoms (5.5 times/year or less; e.g., 3/131 or 2% cases). Based on this tree, subjects who smoked greater than 25.5 cigarettes/day whose income was in the lowest three categories (e.g., <$49,999/yr) were much more likely to be cases compared to smokers with higher incomes. Similarly, among subjects in the lower smoking category (25.5 cigarettes per day or less) who reported experiencing GERD symptoms and who had low non-citrus fruit consumption, non-white subjects were classified as being at higher risk (11/16 or 69% cases) than their white counterparts (23/153 or 15% cases). Only one food group (non-citrus fruit) was selected for this cancer site. Overall, the case misclassification rate was 18%.

Non-cardia gastric adenocarcinoma

The final classification tree, based on the classification of non-cardia gastric adenocarcinoma case status, is presented in Figure 4 (n = 1,039). Eight of the potential 35 variables from the candidate list (Table 1) remained in building this tree. The optimal split on the entire sample involved frequency of GERD symptoms per year. The lowest risk group included those who reported experiencing very infrequent GERD symptoms (4.5 or fewer times/year; 26/164 or 16% cases). Based on this tree, subjects who experienced GERD symptoms (data split at > 4.5 times/year) and who were over 68.5 years of age according to the data split were classified as higher risk. For subjects 68.5 years of age or younger, risk appears to be affected by consumption of nitrites, refined grains, dark green vegetables, whole grains, wine, and income, with sometimes complex interactions (e.g., subjects with low wine consumption and low income were at higher risk, but subjects with

high wine consumption, high nitrate consumption, and low dark green vegetable consumption were also at higher risk). Overall, the case misclassification rate was 25%.

Sensitivity analyses excluding proxy respondents

As in the main analyses that included proxy respondents, when we excluded proxy respondents we found that GERD was the most important risk stratification factor for esophageal adenocarcinoma, gastric cardia adenocarcinoma, and non-cardia gastric adenocarcinoma, and that smoking was the most important risk stratification factor for esophageal squamous cell carcinoma. The number of variables in the classification tree for the proxy-excluded analysis compared with the main analysis was four versus six for esophageal adenocarcinoma, three versus four for gastric cardia adenocarcinoma, three versus six for esophageal squamous cell carcinoma, and eight versus eight for non-cardia gastric adenocarcinoma. The following variables were selected as risk stratification factors in both analyses and were thus the most robust predictors: GERD and non-citrus fruits for esophageal adenocarcinoma; GERD and education for gastric cardia adenocarcinoma; cigarettes and income for esophageal squamous cell carcinoma; and GERD, age, whole grains, and nitrite for non-cardia gastric adenocarcinoma.

In the proxy-excluded analyses the case misclassification rates were 28% for esophageal adenocarcinoma (compared with 31% in the main analysis), 29% for gastric cardia adenocarcinoma (compared with 38% in the main analysis), 19% for esophageal squamous cell carcinoma (compared with 18% in the main analysis), and 26% for non-cardia gastric adenocarcinoma (compared with 25% in the main analysis).

Discussion

In this large population-based case-control study of men and women in the United States, we applied agnostic recursive partitioning to our data and generated subsets of subjects that were relatively homogeneous with respect to important risk variables. Because the risk variables in each of the classification trees had been previously identified as risk factors [7–17, 22–24, 43–52], this analysis provides some insight into how these risk factors interact to increase risk of esophageal and gastric cancer, and their relative importance in risk stratification.

The trees generated were able to classify correctly between 62–82% of all cases, with the tree for esophageal squamous cell carcinoma performing the best (18% misclassification rate) and the tree for gastric cardia adenocarcinoma exhibiting the highest misclassification error (38%). The final trees for esophageal adenocarcinoma and non-cardia gastric adenocarcinoma contained factor splits on both dietary and lifestyle variables, whereas the trees for gastric cardia adenocarcinoma and esophageal squamous cell carcinoma indicated that dietary variables were relatively unimportant in risk stratification for the majority of subjects.

The analysis performed here supports the literature, particularly with respect to the critical role of reported GERD symptoms in risk of adenocarcinoma of the esophagus and gastric cardia [7, 15, 17, 51]. While our previous analysis of food groups found a significantly

increased risk of gastric cardia adenocarcinoma associated with consumption of high-fat dairy foods [24], our classification tree did not select this variable as a split. This is not surprising given our finding that dairy products tracked closely with GERD in a separate principal components analysis [53] and the strong association between this cancer and GERD.

The inclusion of dietary variables such as consumption of red meat and dark green and raw vegetables as splitting factors for esophageal adenocarcinoma is consistent with our previous analyses of these data using unconditional multivariate logistic regression analysis [24], as is the inclusion of refined grains and nitrites as bifurcators for non-cardia gastric adenocarcinoma. As well, although refined grains were found to track with the meat/nitrite group in previous analyses [53], refined grains entered into the tree model at two nodes for non-cardia gastric adenocarcinoma, supporting our previous findings, which suggested higher consumption of refined grains is a significant, independent, risk factor for non-cardia gastric adenocarcinoma.

Likewise, the importance of cigarette smoking, income, and race in the esophageal squamous cell carcinoma tree supports a large body of literature implicating smoking as a primary risk factor [9, 14]. Given literature regarding alcohol consumption as a risk factor for esophageal squamous cell carcinoma [9, 14, 43, 54, 55], it may be surprising that alcohol was not included in this tree. However, in our previous principal components analysis of these data, we found that cigarette smoking and alcohol consumption tracked so closely that little residual information was provided by alcohol intake once smoking was included [53].

One limitation of the study was the relatively high prevalence of proxy interviews among the cases (ranging from 26% for gastric cardia adenocarcinoma to 35% for esophageal squamous cell carcinoma) combined with the fact that our food frequency questionnaire has not been validated for proxy interviews. However, in a previous logistic regression analysis of nutrient intakes in relation to cancer risk in the same study sample, Mayne et al. [23] found the proxy-included and proxy-excluded results to be nearly identical. In the current classification tree study, in our sensitivity analyses we found results of the proxy-included and proxy-excluded analyses to be broadly similar, but with appreciable differences.

Nevertheless, we decided to include the proxy interviews in our main analyses for several reasons. First, with the exception of non-cardia gastric adenocarcinoma, the proxy-excluded trees included fewer variables than the proxy-included trees. This difference was likely due to the decreased size of the case group and resultant decreased statistical power in the proxy-excluded analyses. Given the similar case misclassification rates in the proxy-included and proxy-excluded analyses (with the exception of gastric cardia adenocarcinoma), it seemed reasonable to opt for the greater statistical power.

Second, given that the study population was largely male, and most proxy interviews were done with spouses [9], the proxies (mostly wives) were likely able to validly report on the dietary habits, smoking and other behaviors of their spouses (mostly husbands). Lindstead & Kuzma [56] found that spouses tended to eat the same diet, suggesting that spouses can report on their partner’s diet with reasonable accuracy. Furthermore, several small studies

have found diets reported by subjects to be moderately correlated with diets reported by their spouse [57].

One of the strengths of recursive partitioning is the ability to illustrate that among some subgroups of individuals with higher risk behaviors, risk can potentially be attenuated by other 'protective' behaviors. For example, among individuals with more frequent GERD symptoms who ate more than 0.68 servings of red meat per day, those who ate more non-citrus fruit had a lower risk of esophageal adenocarcinoma than those who ate less (20% cases vs. 52% cases in terminal nodes). Likewise, among those who consumed larger amounts of wine and dietary nitrites, we observed a lower risk for non-cardia gastric adenocarcinoma among those who consumed more versus less dark green vegetables.

The literature has not consistently identified an association between GERD and non-cardia gastric adenocarcinoma, and it is therefore surprising that GERD was selected as the primary split for this cancer. However, Haber et al. [58], in their analysis of reflux symptoms and gastritis, concluded that all *H. pylori* positive subjects with GERD symptoms show some type of gastritis, which has been associated with the development of gastric neoplasms [59]. While this association is dependent on whether the strain in question is CagA-positive or CagA-negative, it could, in part, explain our finding.

Overall, these findings are consistent with and extend our own earlier findings that utilized unconditional multivariate logistic regression techniques to analyze these variables separately, suggesting that tree-based methods may be useful in partitioning risk for these cancers. These trees also provide some amount of internal validity by their concurrence with risk factors previously identified by our studies, which utilized more traditional analytic methodologies.

One of the advantages to using the classification tree analysis method is that logistic regression methods can only evaluate interactions that are selected *a priori*, whereas recursive partitioning techniques have the ability to reveal interactions that may not have been otherwise considered. Thus, future investigations of potential risk factors can take into account variables that only exert influence when in the presence of additional risk factors. Another important strength of classification tree analysis is its use of splitting algorithms to select cut points for continuous risk variables, which need not be normally distributed [41]. As well, this type of analysis allows for the inclusion of the same risk variable at different levels of the tree using different cut points. However, while variables can enter the tree repeatedly at different levels, thus implying a dose-response relationship, classification tree analyses do not allow for measuring dose-response in a straightforward manner.

The classification tree analyses presented here are the first to attempt to model multi-level interactions in esophageal and gastric cancers. A limitation of this type of analysis is that trees are grown based on maximizing magnitudes of odds ratios across splits and given the large number of splits considered, interpretation becomes increasingly challenging. Given that measurement error is inherent in dietary assessment methods, the interpretation of lower nodes may become suspect, particularly when these nodes represent a small subsample of the study population and may therefore be less clinically important. This problem is

fortunately mitigated during the pruning phase of the CART tree development. It is important, however, that these precise nodes and cut points not be over-interpreted, but rather, they should be viewed as identifying which variables are potentially most important for risk stratification. Further, because frequency of reflux symptoms was the primary split for esophageal and gastric adenocarcinomas, as was cigarette smoking for esophageal squamous cell carcinoma, our results suggest reasonable points to start partitioning risk from a clinical standpoint as well. Nevertheless, validation of pruned trees in separate studies will assist in confirming that such results do not overfit the data or represent chance findings.

Acknowledgments

We thank the following: study managers Sarah Greene and Linda Lannom (Westat), data management Shelley Niwa (Westat), and field supervisors Patricia Owen (Connecticut), Tom English (New Jersey), and Berta Nicol-Blades (Washington) for data collection and processing; Dr. Alan Kristal for assistance in designing and processing the dietary questionnaires; the Yale Cancer Center Rapid Case Ascertainment Shared Resource, the 178 hospitals in Connecticut, New Jersey, and Washington for their participation in the study; and the study participants. Certain data used in this study were obtained from the Connecticut Tumor Registry, located in the Connecticut Department of Public Health. The authors assume full responsibility for analyses and interpretation of these data.

Funding

United States Public Health Service (U01-CA57983, U01-CA57949, U01-CA57923, P30-ES10126); National Cancer Institute, National Institutes of Health, Department of Health and Human Services (N02-CP40501, N01-CN05230).

Abbreviations

FHCRC	Fred Hutchinson Cancer Research Center
GERD	gastroesophageal reflux

References

1. Blot WJ, Devesa SS, Kneller RW, Fraumeni JF Jr. Rising incidence of adenocarcinoma of the esophagus and gastric cardia. JAMA. 1991; 265:1287–1289. [PubMed: 1995976]
2. Devesa SS, Blot WJ, Fraumeni JF Jr. Changing patterns in the incidence of esophageal and gastric carcinoma in the United States. Cancer. 1998; 83:2049–2053. [PubMed: 9827707]
3. Simard EP, Ward EM, Siegel R, Jemal A. Cancers with increasing incidence trends in the United States: 1999 through 2009. Ca Cancer J Clin. 1999; 62:118–128.
4. Holmes RS, Vaughan TL. Epidemiology and pathogenesis of esophageal cancer. Semin Radiat Oncol. 2007; 17(1):2–9. [PubMed: 17185192]
5. Brown LM, Swanson CA, Gridley G, Swanson GM, Silverman DT, Greenberg RS, et al. Dietary factors and risk of squamous cell esophageal cancer among black and white men in the United States. Cancer Causes Control. 1998; 9:467–474. [PubMed: 9934713]
6. Bollschweiler E, Wolfgarten E, Gutschow C, Holscher AH. Demographic variations in the rising incidence of esophageal adenocarcinoma in white males. Cancer. 2001; 3:549–555. [PubMed: 11505399]
7. Anderson LA, Watson RGP, Murphy SJ, Johnston BT, Comber H, Mc Guigan J, et al. Risk factors for Barrett's oesophagus and oesophageal adenocarcinoma: results from the FINBAR study. World J Gastroenterol. 2007; 13(10):1585–1594. [PubMed: 17461453]
8. Chow WH, Swanson CA, Lissowska J, Groves FD, Sobin LH, Nsierowska-Guttmejer A, et al. Risk of stomach cancer in relation to consumption fo cigarettes, alcohol, tea and coffee in Warsaw, Poland. Int J Cancer. 1999; 81:871–876. [PubMed: 10362132]

9. Gammon MD, Schoenberg JB, Ahsan H, Risch HA, Vaughan TL, Chow WH, et al. Tobacco, alcohol, and socioeconomic status and adenocarcinomas of the esophagus and gastric cardia. J Natl Cancer Inst. 1997; 89:1277–1284. [PubMed: 9293918]
10. Lagergren J, Bergtrom R, Lindgren A, Nyren O. The role of tobacco, snuff and alcohol use in the aetiology of cancer of the oesophagus and gastric cardia. Int J Cancer. 2000; 85:340–346. [PubMed: 10652424]
11. Brown LM, Swanson CA, Gridley G, Swanson GM, Schoenberg JB, Greenberg RS, et al. Adenocarcinoma of the esophagus: Role of obesity and diet. J Natl Cancer Inst. 1995; 87:104–109. [PubMed: 7707381]
12. Chow WH, Blot WJ, Vaughan TL, Risch HA, Gammon MD, Stanford JL, et al. Body mass index and risk of adenocarcinomas of the esophagus and gastric cardia. J Natl Cancer Inst. 1998; 90:150–155. [PubMed: 9450576]
13. Lagergren J, Bergtrom R, Nyren O. Association between body mass and adenocarcinoma of the esophagus and gastric cardia. Ann Intern Med. 1999; 130:883–890. [PubMed: 10375336]
14. Vaughan TL, Davis S, Kristal A, Thomas DB. Obesity, alcohol, and tobacco as risk factors for cancers of the esophagus and gastric cardia: Adenocarcinoma versus squamous cell carcinoma. Cancer Epidemiol Biomark Prev. 1995; 4:85–92.
15. Chow WH, Finkle W, McLaughlin JK, Frankl H, Ziel HK, Fraumeni JF Jr. The relation of gastroesophageal reflux disease and its treatment to adenocarcinomas of the esophagus and gastric cardia. JAMA. 1995; 274:474–477. [PubMed: 7629956]
16. Farrow DC, Vaughan TL, Sweeney C, Gammon M, Chow WH, Risch HA, et al. Gastroesophageal reflux disease, use of H2 receptor antagonists, and risk of esophageal and gastric cancer. Cancer Causes Control. 2000; 11:231–238. [PubMed: 10782657]
17. Lagergren J, Bergtrom R, Lindgren A, Nyren O. Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. New England J Med. 1999; 240:825–831. [PubMed: 10080844]
18. Chow WH, Blaser MJ, Blot WJ, Gammon MD, Vaughan TL, Risch HA, et al. An inverse relation between cagA+ strains of Helicobacter pylori infection and risk of esophageal and gastric cardia adenocarcinoma. Cancer Res. 1998; 58(4):588–590. [PubMed: 9485003]
19. Kamangar F, Dawsey SM, Blaser MJ, Perez-Perez GI, Pietinen P, Newschaffer CJ, et al. Opposing risks of gastric cardia and noncardia gastric adenocarcinomas associated with Helicobacter pylori seropositivity. J Natl Cancer Inst. 2006; 98(20):1445–1452. [PubMed: 17047193]
20. Wong A, Fitzgerald RC. Epidemiologic risk factors for Barrett's esophagus and associated adenocarcinoma. Clin Gastroenterol Hepatol. 2005; 3:1–10. [PubMed: 15645398]
21. World Cancer Research Fund. Food, physical activity and the prevention of cancer: a global perspective. Washington DC: American Institute for Cancer Research; 2007.
22. Thrift AP, Pandeya N, Whiteman DC. Current status and future perspectives on the etiology of esophageal adenocarcinoma. Frontiers in Oncology. 2012; 2:1–7.
23. Mayne ST, Risch R, Dubrow R, Chow WH, Gammon MD, Vaughan T, et al. Nutrient intake and risk of subtypes of esophageal and gastric cancer. Cancer Epidemiol Biomark Prev. 2001; 10:1055–1062.
24. Navarro-Silvera SA, Mayne ST, Risch HA, Gammon MD, Vaughan TL, Chow WH, et al. Food group intake and risk of subtypes of esophageal adenocarcinoma. Int J Cancer. 2008; 123:852–860. [PubMed: 18537156]
25. Steevens J, Schouten LJ, Goldbohm RA, van den Brandt PA. Vegetables and fruits consumption and risk of esophageal and gastric cancer subtypes in the Netherlands Cohort Study. Int J Cancer. 2011; 129(11):2681–2693. [PubMed: 21960262]
26. Freedman ND, Park Y, Subar AF, Hollenbeck AR, Lieitzmann MF, Schatskin A, et al. Fruit and vegetable intake and esophageal cancer in a large prospective cohort study. Int J Cancer. 2007; 121:2753–2760. [PubMed: 17691111]
27. Keszei AP, Schouten LJ, Goldbohm RA, van den Brandt PA. Red and processed meat consumption and the risk of esophageal and gastric cancer subtypes in The Netherlands Cohort Study. Ann Oncol. 2012; 23(9):2319–2326. [PubMed: 22351741]

28. Terry P, Lagergren J, Ye W, Wolk A, Nyren O. Inverse association between intake of cereal fiber and risk of gastric cardia cancer. Gastroenterol. 2001; 120:387–391.
29. Gordon T, Fisher M, Rifkind BM. Some difficulties inherent in the interpretation of dietary data from free-living populations. Am J Clin Nutr. 1984; 39:152–156. [PubMed: 6606975]
30. Kant AK, Schatzkin A, Graubard BI, Schairer C. A prospective study of diet quality and mortality in women. JAMA. 2000; 283(16):2109–2115. [PubMed: 10791502]
31. Randall E, Marshal JR, Graham S, Brasure J. High-risk health behaviors associated with various dietary patterns. Nutr Cancer. 1991; 16(2):135–151. [PubMed: 1796009]
32. Wirfalt AK, Jeffery RW. Using cluster analysis to examine dietary patterns: nutrient intakes, gender, and weight status differ across food pattern clusters. J Am Diet Assoc. 1997; 97(3):272–279. [PubMed: 9060944]
33. Germanson TP, Lanzino G, Kongable GL, Torner JC, Kassell NF. Risk classification after aneurysmal subarachnoid hemorrhage. Surg Neurol. 1998; 49:155–163. [PubMed: 9457265]
34. Hess KR, Abbruzzese MC, Lenzi R, Raber MN, Abbruzzese JL. Classification and regression tree analysis of 100 consecutive patients with unknown primary carcinoma. Clin Cancer Res. 1999; 5:3403–3410. [PubMed: 10589751]
35. Camp NJ, Slattery ML. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). Cancer Causes Control. 2002; 13:813–823. [PubMed: 12462546]
36. Papathomas M, Molitor J, Richardson S, Riboli E, Vineis P. Examining the joint effect of multiple risk factors using exposure risk profiles: Lung cancer in nonsmokers. Environ Health Perspect. 2011; 119(1)10.1289/ehp.1002118
37. Zhang, H.; Singer, BH. Recursive Partitioning and Applications. 2. New York: Springer; 2010.
38. Waksberg J. Sampling methods for random digit dialing. J Am Stat Assoc. 1978; 73:40–46.
39. Kristal AR, Feng Z, Coates FJ, Oberman A, George V. Associations of race, ethnicity, education and dietary intervention on the validity and reliability of a food frequency questionnaire in the Women's Health Trial Feasibility Study in Minority Populations. Am J Epidemiol. 1997; 146:856–869. [PubMed: 9384206]
40. Risch HA, Jain M, Choi NW, Fodor JG, Pfieffer CJ, How GR, et al. Dietary factors and the incidence of cancer of the stomach. Am J Epidemiol. 1985; 122:947–959. [PubMed: 2998182]
41. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Belmont, CA: Wadsworth; 1984.
42. Nelson LM, Bloch DA, Longstreth WTJ, Shi W. Recursive partitioning for the identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. J Clin Epidemiol. 1998; 51(3):199–209. [PubMed: 9495685]
43. Bahmanyar S, Ye W. Dietary patterns and risk of squamous-cell carcinoma and adenocarcinoma of the esophagus and adenocarcinoma of the gastric cardia: A population-based case-control study in Sweden. Nutr Cancer. 2006; 54(2):171–178. [PubMed: 16898861]
44. Chen H, Ward MH, Graubard BI, Heineman EF, Markin RM, Potischman NA, et al. Dietary patterns and adenocarcinoma of the esophagus and distal stomach. Am J Clin Nutr. 2002; 75:137–144. [PubMed: 11756071]
45. Cheng KK, Sharp L, McKinney PA, Logan RFA, Chilvers CED, Cook-Mozaffari P, et al. A case-control study of oesophageal adenocarcinoma in women: a preventable disease. Br J Cancer. 2000; 83(1):127–132. [PubMed: 10883680]
46. Gonzalez CA, Jakszyn P, Pera G, Agudo A, Bingham S, Palli D, et al. Meat intake and risk of stomach and esophageal adenocarcinoma within the European Prospective Investigation Into Cancer and Nutrition (EPIC). J Natl Cancer Inst. 2006; 98(5):345–354. [PubMed: 16507831]
47. Levi F, Pasche C, Lucchini F, Bosetti C, Franceschi S, Monnier P, et al. Food groups and oesophageal cancer risk in Vaud, Switzerland. Eur J Cancer Prev. 2000; 9(4):257–263. [PubMed: 10958328]
48. Terry P, Lagergren J, Hansen H, Wolk A, Nyren O. Fruit and vegetable consumption in the prevention of oesophageal and cardia cancers. Eur J Cancer Prev. 2001; 10:365–369. [PubMed: 11535879]

49. Tuyns AJ, Riboli E, Doornbos G, Pequignot G. Diet and esophageal cancer in Calvados (France). Nutr Cancer. 1987; 9(2–3):81–92. [PubMed: 3562297]
50. World Cancer Research Fund. Food, nutrition and the prevention of cancer: a global perspective. Washington DC: American Institute for Cancer Research; 1997.
51. Ye W, Chow WH, Lagergren J, Yin L, Nyren O. Risk of adenocarcinomas of the esophagus and gastric cardia in patients with gastroesophageal reflux diseases and after antireflux surgery. Gastroenterol. 2001; 121:1286–1293.
52. Zhang ZF, Kurtz RC, Tu GP, Sun M, Gargon N, Karpeh M Jr, et al. Adenocarcinomas of the esophagus and gastric cardia: the role of diet. Nutr Cancer. 1997; 27:298–309. [PubMed: 9101561]
53. Navarro Silvera SA, Mayne ST, Risch HA, Gammon MD, Vaughan T, Chow WH, et al. Principal component analysis of dietary and lifestyle patterns in relation to risk of subtypes of esophageal and gastric cancer. Ann Epidemiol. 2011; 21(7):543–550. [PubMed: 21435900]
54. Blot, WJ.; McLaughlin, JK.; Fraumeni, JF., Jr, editors. Esophageal Cancer. 3. Oxford: Oxford University Press; 2006.
55. Garidou A, Tzonou A, Lipworth L, Signorello LB, Kalapothaki V, Trichopoulos D. Lifestyle factors and medical conditions in relation to esophageal cancer by histologic type in a low-risk population. Int J Cancer. 1996; 68:295–299. [PubMed: 8903469]
56. Lindsted K, Kuzma JW. Husband-wife diet concordance and changes in dietary practices by surviving spouses of cancer cases. Nutr Cancer. 1990; 13(3):175–187. [PubMed: 2308873]
57. Willett, W. Nutritional Epidemiology. 3. Oxford: Oxford University Press; 2013.
58. Haber MM, Lopez L. Reflux gastritis in gastroesophageal reflux disease: A histopathological study. Ann Diagn Pathol. 1999; 3(5):281–286. [PubMed: 10556474]
59. Genta RM. The gastritis connection: prevention and early detection of gastric neoplasms. J Clin Gastroenterol. 2003; 36(5Suppl):S44–49. [PubMed: 12702965]

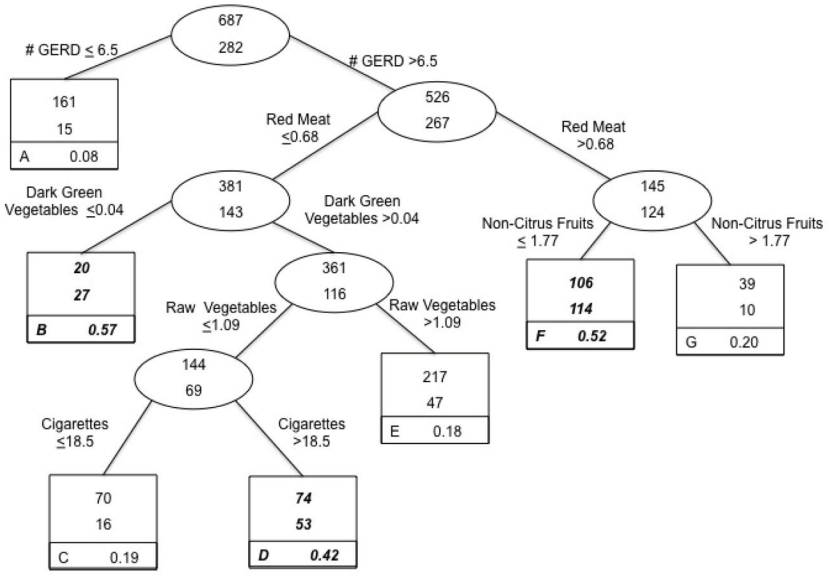

Figure 1.

Classification Tree: Lifestyle factors and risk of esophageal adenocarcinoma, from United States multicenter, population-based study (1993–1995).

Each group contains the number of controls (top number of uppermost box) and the number of cases (bottom number of uppermost box). Terminal subsets are represented by square boxes and are identified by letter in the lower left corner. The proportion in the bottom right corner of each terminal subset gives the probability of being a case in that group. Because the prevalence of cases in the total sample was 29%, terminal subsets comprised of more than 29% cases are considered higher risk groups for classification purposes and are highlighted in bold italics.

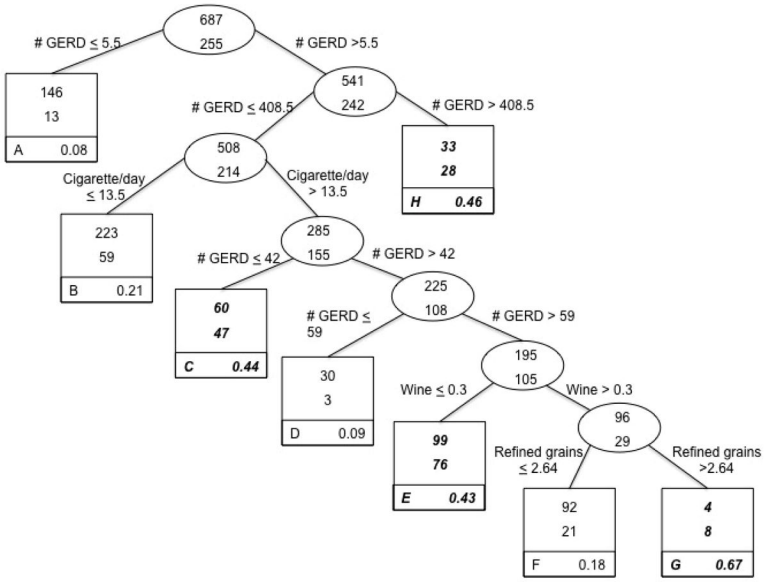

Figure 2.

Classification Tree: Lifestyle factors and risk of gastric cardia adenocarcinoma, from United States multicenter, population-based study (1993–1995).

Each group contains the number of controls (top number of uppermost box) and the number of cases (bottom number of uppermost box). Terminal subsets are represented by square boxes and are identified by letter in the lower left corner. The proportion in the bottom right corner of each terminal subset gives the probability of being a case in that group. Because the prevalence of cases in the total sample was 27%, terminal subsets comprised of more than 27% cases are considered higher risk groups for classification purposes and are highlighted in bold italics.

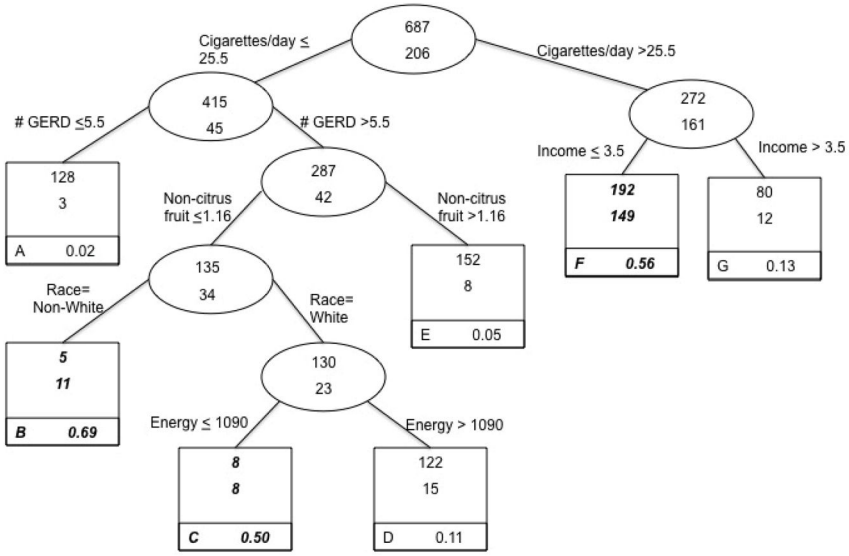

Figure 3.

Classification Tree: Lifestyle factors and risk of esophageal squamous cell carcinoma, from United States multicenter, population-based study (1993–1995).

Each group contains the number of controls (top number of uppermost box) and the number of cases (bottom number of uppermost box). Terminal subsets are represented by square boxes and are identified by letter in the lower left corner. The proportion in the bottom right corner of each terminal subset gives the probability of being a case in that group. Because the prevalence of cases in the total sample was 23%, terminal subsets comprised of more than 23% cases are considered higher risk groups for classification purposes and are highlighted in bold italics.

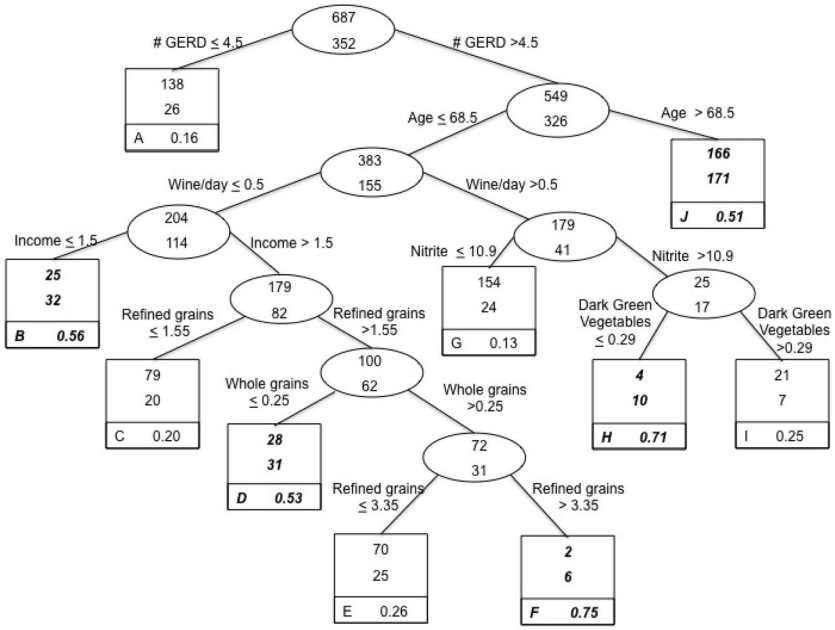

Figure 4.

Classification Tree: Lifestyle factors and risk of non-cardia gastric adenocarcinoma, from United States multicenter, population-based study (1993–1995).

Each group contains the number of controls (top number of uppermost box) and the number of cases (bottom number of uppermost box). Terminal subsets are represented by square boxes and are identified by letter in the lower left corner. The proportion in the bottom right corner of each terminal subset gives the probability of being a case in that group. Because the prevalence of cases in the total sample was 34%, terminal subsets comprised of more than 34% cases are considered higher risk groups for classification purposes and are highlighted in bold italics.

Table 1

Potential explanatory variables in a United States multi-center, population-based case-control study of esophageal and gastric cancer (1993–1995)

Demographic Variables	
Gender	Education (categorical, 7 levels)
Age (continuous)	Income (categorical, 6 levels)
Race (White/non-White)	
Dietary Variables†	
Fruit and other Juices	High-fat Dairy products
Citrus Fruit	Whole Grains
Non-citrus Fruit	Refined Grains
Cruciferous Vegetables	Fish
Deep yellow Vegetables	Poultry
Low Vitamin A cruciferous Vegetables	High-nitrite Meats
Dark green leafy Vegetables	Red Meats
Starchy Vegetables	Meat Alternates
Raw Vegetables	Nitrites (mg/day)
Dry beans and peas (legumes)	Dietary Fiber (mg/day)
Tomato products	Vitamin C (mg/day)
Low-fat Dairy products	Energy intake (Kcal/day)
Medical variables	
Body mass index (kg/m^2) (continuous)	Gastroesophageal reflux symptoms (per year)
Lifestyle variables	
Beer (drinks/day)	Liquor (drinks/day)
Wine (drinks/day)	Cigarette smoking (cigarettes/day)

† Servings/day unless otherwise noted