

Article

Predictors of the Healthy Eating Index and Glycemic Index in Multi-Ethnic Colorectal Cancer Families

S. Pamela K. Shiao ^{1,*} , James Grayson ², Amanda Lie ³ and Chong Ho Yu ⁴¹ College of Nursing and Medical College of Georgia, Augusta University, Augusta, GA 30912, USA² Hull College of Business, Augusta University, Augusta, GA 30912, USA; jgrayson@augusta.edu³ Citrus Valley Health Partners, Foothill Presbyterian Hospital, Glendora, CA 91741, USA; amandalie413@gmail.com⁴ School of Business, University of Phoenix, Pasadena, CA 91101, USA; alexyu@email.phoenix.edu

* Correspondence: pshiao@msn.com; Tel.: 1-(818)-233-6112

Received: 3 May 2018; Accepted: 24 May 2018; Published: 26 May 2018



Abstract: For personalized nutrition in preparation for precision healthcare, we examined the predictors of healthy eating, using the healthy eating index (HEI) and glycemic index (GI), in family-based multi-ethnic colorectal cancer (CRC) families. A total of 106 participants, 53 CRC cases and 53 family members from multi-ethnic families participated in the study. Machine learning validation procedures, including the ensemble method and generalized regression prediction, Elastic Net with Akaike's Information Criterion with correction and Leave-One-Out cross validation methods, were applied to validate the results for enhanced prediction and reproducibility. Models were compared based on HEI scales for the scores of 77 versus 80 as the status of healthy eating, predicted from individual dietary parameters and health outcomes. Gender and CRC status were interactive as additional predictors of HEI based on the HEI score of 77. Predictors of HEI 80 as the criterion score of a good diet included five significant dietary parameters (with intake amount): whole fruit (1 cup), milk or milk alternative such as soy drinks (6 oz), whole grain (1 oz), saturated fat (15 g), and oil and nuts (1 oz). Compared to the GI models, HEI models presented more accurate and fitted models. Milk or a milk alternative such as soy drink (6 oz) is the common significant parameter across HEI and GI predictive models. These results point to the importance of healthy eating, with the appropriate amount of healthy foods, as modifiable factors for cancer prevention.

Keywords: healthy eating; glycemic index; colorectal cancer; generalized regression elastic net; diverse ethnic groups

1. Introduction

Colorectal cancer (CRC) is recognized as the most preventable cancer worldwide [1]. Unhealthy dietary habits with excess caloric intake and weight gain, smoking, and over-consuming alcohol can increase the risk of developing CRC [2–6] through inflammatory oxidative stress pathways [7–12]. Furthermore, based on the strong evidence defined by The American Institute for Cancer Research (AICR), plant-based foods and healthy weight, reducing red meat and alcohol intake [1,13] can prevent CRC. Fortunately, dietary habits can be improved as one of the modifiable lifestyle factors to prevent CRC and cancer progression [6–12].

A healthy diet has been associated with decreased CRC risk, examined by using the healthy eating index (HEI) [14–18]. Elements of a healthy diet include adequate intakes of vegetables and dark green vegetables, fruits and whole fruits, grains and whole grains, nuts and legumes, proteins including fish and other seafood, milk or alternative dairy products for lactose intolerance; and limiting salt, saturated fat, and empty calories from sugar and alcohol [2–5,7,8]. Higher HEI scores

are associated with decreased CRC risk [19–21]. Additionally, diets rich in fiber, folate, calcium, limiting pro-inflammatory fatty acids are protective against CRC [22–24]. The glycemic index (GI) has been used to assess healthy eating in association with CRC, to manage hyperinsulinemia and insulin resistance [25]. A low-GI with low glycemic load (GL) diet may decrease inflammation and CRC risk [26–30]. For prevention, the risk of CRC was reduced by half when participants followed 4–6 recommendations of healthy eating components over 8 years [31].

In summary, dietary habits are formed over time within families, and the family units can share both dietary habits as part of lifestyle [22–24,32] and the heredity of genome and epigenetics of CRC [33–35]. Family-based studies can provide potential insights into developing prevention strategies for cancer prevention. Therefore, the aim of this study, following a previous report on gene-environment interactions in a family-based study involving CRC patients and their family members [32], was to investigate and predict healthy eating practices by HEI and GI from various dietary and demographic factors of the multi-ethnic CRC families. In this study, we used machine learning validation procedures including the ensemble method [36–39] and generalized regression prediction, Elastic Net with Akaike's Information Criterion with correction and Leave-One-Out cross validation methods [40–43].

2. Materials and Methods

2.1. Study Population and Setting

The study methods were reported before [32] and are summarized in the following. We included 106 participants, 53 CRC cases and 53 family members by accessing the California Cancer Registry (CCR) database and other cases through referrals from the community that the study was conducted. The designated Human Subjects Institutional Review Boards (IRB) from the local educational institutions and the California State Committee for the Protection of Human Subjects approved the project [32]. With the approved study procedures, the qualified participants were recruited. The participants were interviewed on campus or in their homes.

2.2. Demographic Data

Demographic data included lifestyle and dietary status [32,44], family history, functional capacities using the items included in the 1999–2012 National Health Interview Survey [45] and the family pedigrees from the Coalition for Health Professional Education in Genetics ([46], www.nchpeg.org).

2.3. Dietary Indexes

We assessed healthy eating by using dietary measurements including HEI (HEI-2015) [16,18], GI [47,48] and recommended daily intakes (RDI) [17], collected with Food Frequency Questionnaire [49,50] and data processed through the Nutrition Data Systems for Research [51,52]. HEI was developed to assess diet quality issued by the US Department of agriculture (USDA) based on the standards of a healthy lifestyle in association with health outcomes. HEI is composed of 12 scored components which include 5 major food groups: fruit (total and whole), vegetable (total and greens/beans), grains (total and whole), dairy or alternative dairy and protein, oils and nuts; in addition to limiting saturated fats, sodium, and empty calories. The total HEI score is the sum of the components, with a range of 0 to 100. A score between 0–50 indicates a poor diet; 51–80, a moderate diet quality that needs improvement; and a score greater than 80, a good diet [16,53].

GI is a measure of carbohydrates in foods on a scale of 0–100, based on how the foods affect the levels of blood sugar. Foods with a high GI (score of 70 or more) are quickly digested, absorbed and metabolized, causing a quick spike in blood sugar and insulin levels. A low GI diet (score of 55 or less) includes whole grains or carbohydrates that lead to a slow and steady release of blood sugar and insulin [54]. Examples of foods with high GI include white bread, pretzels, potatoes, corn flakes, and foods with lower GI include whole wheat bread, rolled or steel-cut oatmeal, sweet

potatoes, legumes, non-starchy vegetables. One study systematically organized GI values for over 1000 foods [55]. GL takes into consideration the GI in foods (<http://lpi.oregonstate.edu/mic/food-beverages/glycemic-index-glycemic-load>). GL is calculated by multiplying the GI by the quantity (grams) of carbohydrates in a serving of a food divided by 100 (≤ 10 : low, 11–19: medium, ≥ 20 : high [56]).

The recommended daily intake (RDI) is issued by the Food and Nutrition Board of the Institute of Medicine, which recommends the sufficient required daily intake of nutrients for healthy people based on gender and age [17]. Macronutrients include carbohydrates, protein, total fat, saturated fat, cholesterol; B vitamins—B9 (folate), B1 (thiamine), B2 (riboflavin), B3 (niacin) B6 and B12; and other micronutrients—Vitamin A, C, D and E, calcium, magnesium, iron, zinc, methionine, and choline [57]

2.4. Data Analysis

Machine learning based analytics were employed in JMP Pro 13 ([58–60], SAS Institute, Cary, NC, USA). The analytics and rationales have been reported earlier [32] and are summarized in the following. We included ensemble methods [36–39], for a well-known remedy in small-sample studies [61] with random subsets of repeated analysis to correct bias [62], which is superior to conventional regression modeling for a best fit model [63,64]. We used generalized regression (GR) with machine learning validation to obtain a smaller prediction error [43]. It is important to point out that GR eliminates certain predictors to avoid over-fitting. For example, when there are several collinear predictors, LASSO selects only one and ignore the others or zeroes out some regression coefficients. The Ridge method counteracts against collinearity and variance inflation by shrinking the regression coefficients towards zero, but not exactly zero. The Elastic Net method combines the penalties of the LASSO and Ridge approaches. Unlike linear least squares in estimating the unknown parameters in a linear regression model, GR could simply zero out certain unused predictors [60]. In traditional statistics, usually one model is used to fit the data, and thus the probability is nothing more than an approximation based on sampling distributions, which are open-ended (the two-tails never touch the x-axis). In this case, the p value at most could only be 0.9999, but not exactly one. However, when all permutations are exhausted, such as what was done in an exact test, the probability could be exactly one. In a similar vein, GR exhausts different paths to find the best model. When the full model has a mixture of important and unused predictors, the p value cannot be one. However, when the data could be perfectly described by the restricted model resulting from path searching, the probability of observing the data could be 1.

When developing a GR model for a predictive model, the first type of model presented in JMP Pro 13 is a logistic regression (LR) model because the default estimation method is an LR. After this default method, other model launches can be pursued by choosing a variety of estimation methods (lasso, Elastic Net and others) and associated validation methods (a validation column, minimum AICc, leave-one-out (LOO) validation and others, [65]). Both AICc validation and LOO cross-validation methods are effective methods for small data sets [66]. In effect, the default LR method could be characterized as an explanatory model whereas the other GR estimation methods might best be characterized as a predictive model. An explanatory model is typically used to explain the association between the model parameters and the model response to test causal hypotheses, using a predictive model, for predicting future observations [67]. The nature of the model objectives (causal versus predictive) directly influence the underlying algorithms which can result in different results of models using the same set of initial parameters. Typically, using an explanatory model, a set of statistically significant parameters is identified for a final model. The predictive model using GR will pursue methods to shrink coefficients towards zero in part to guard against overfitting the model. For model prediction in GR analysis, continuous variables are recoded into new dichotomous variables grouped by either median distribution or a known score criterion of healthy eating. The prediction profiler and interactive profiler can be used to visualize the direction of association between two parameters (a predictor or factor with the outcome variable of healthy eating status in profiler) or among three parameters (set of interactive variables with non-parallel distribution in addition to the outcome status

of healthy eating in the interactive profiler). The visualization of the profiler and interactive profiler will enable the analyst to visualize and account for the interactions of various factors. The index of showing the fitness of the model over complexity is AIC or AICc [64–70], with a smaller AIC suggesting a more optimal model for model quality [68,71,72]. We examined model quality using the misclassification rate (smaller is better), AICc, and the area under the receiver operating characteristic (ROC) curve (AUC).

3. Results

3.1. Characteristics of Study Participants

Table 1 presents the key demographic characteristics of the 106 participants. There were more women than men in the sample, with racial compositions of about one-third Asians, one-third Caucasians, and one-third Hispanic and African Americans combined. About 25% of the sample presented as obese based on body mass index (BMI), more than half of the sample drank alcohol, and 8.5% were smokers.

Table 1. Demographic characteristics of the sample.

| Parameters | | Total (N = 106) n (%) |
|-----------------|------------------|-----------------------|
| Gender | Male | 39 (37%) |
| | Female | 67 (63%) |
| Age, Years | M ± SD | 54 ± 16 |
| Ethnicity | Asian | 40 (38%) |
| | Caucasian | 34 (32%) |
| | Hispanic | 23 (22%) |
| | African American | 9 (9%) |
| BMI status | Obese | 26 (25%) |
| Alcohol drinker | Yes | 57 (54%) |
| Smoker | Yes | 9 (9%) |

Note: BMI: body mass index.

3.2. Dietary Parameters

The distribution on the dietary parameters was organized and presented for HEI and GI parameters in Table 2. Overall, this sample presented a healthy eating profile based on the average recommended intake for the HEI parameters; however, the average of limiting parameters (saturated fat, salt, and empty calorie) was higher than recommended levels, with less than half of the sample (38.7%) receiving a good HEI score of greater than 80. The median HEI score for this sample was 77, with 51% of the sample above the median score. The average GI was 53.8, which presents as a low GI diet (good GI), with 62% of the sample scoring less than 55 (Table 2).

RDI parameters are presented in Table 3. For carbohydrates, 71% of the sample consumed more than 45% of the RDI. For protein, 36% of the sample ate more than 20% of the RDI. For saturated fat, 48% of the sample consumed less than 10% of RDI. On average, the sample consumed more than the RDI for all parameters except cholesterol, fiber, total folate, calcium and magnesium. For cholesterol, the mean intake was 259 mg, and 74% of the sample ate less than 100% of the RDI (<300 mg). For fiber, the mean intake was 19 g; 15% of the sample ate more than 100% of the RDI (≥ 25 g). In addition, only 32% of the sample had more than 100% of the RDI for total folate (>400 μ g), with an average mean intake of 365.5 μ g. Less than half of the sample consumed more than 75% of the RDI for calcium (1000 mg) and magnesium (320 mg), with mean intakes of 837 mg and 295 mg, respectively. For sodium, the mean intake was 2950 mg, which was greater than the RDI of <2300 mg, and only 38% of the sample ate less than 100% of the RDI (Table 3).

Table 2. Healthy Eating Index and parameters for the sample ($N = 106$).

| Parameters (Amount, Maximum Score) | Intake M \pm SD | Score M \pm SD | Maximum Score n (%) |
|--|-------------------|------------------|------------------------|
| Calorie (per day) | 1600 \pm 850 | – | – |
| Total Fruit (≥ 0.8 cup, 5 points) | 1.6 \pm 1.5 | 4.1 \pm 1.5 | 69 (65%) |
| Whole Fruit (≥ 0.4 cup, 5 points) | 1.2 \pm 1.1 | 4.2 \pm 1.5 | 78 (74%) |
| Vegetables (≥ 1.1 cup, 5 points) | 1.5 \pm 1.2 | 4.1 \pm 1.4 | 62 (59%) |
| Dark greens (≥ 0.4 cup, 5 points) | 0.9 \pm 0.7 | 4.4 \pm 1.2 | 77 (73%) |
| Total Grain (≥ 3 oz, 5 points) | 4.6 \pm 3.0 | 4.3 \pm 1.2 | 66 (62%) |
| Whole Grain (≥ 1.5 oz, 5 points) | 1.7 \pm 1.7 | 3.4 \pm 1.7 | 45 (43%) |
| Dairy (≥ 1.3 cup, 10 points) | 1.4 \pm 3.2 | 5.4 \pm 3.6 | 24 (23%) |
| Protein (≥ 2.5 oz, 10 points) | 5.8 \pm 4.2 | 9.2 \pm 2.0 | 86 (54%) |
| Oil and nuts (≥ 12 g, 10 points) | 36 \pm 22 | 9.9 \pm 0.4 | 103 (97%) |
| Saturated Fat (g, $\leq 8\%$ energy) | 18.5 \pm 11.4 | 7.0 \pm 2.7 | 18 (17%) |
| Sodium (≤ 1.1 g, 10 points) | 3.2 \pm 1.9 | 2.1 \pm 3.5 | 6 (6%) |
| Empty Calorie ($\leq 19\%$ energy, 20 points) | 350 \pm 230 | 17.6 \pm 3.4 | 47 (44%) |
| Healthy Eating Index score (>80 , good) | 76 \pm 9 | – | >80 : 41 (39%) |
| ≥ 77 (median distribution) | | | ≥ 77 : 54 (51%) |
| Glycemic Load | 96 \pm 59 | – | – |
| Glycemic Index (≤ 55 , low and good) | 54 \pm 4.2 | – | ≤ 55 : 66 (62%) |
| ≤ 53.8 (median distribution) | | | ≤ 53.8 : 53 (50%) |

Note: M: mean, SD: standard deviation, oz: ounce, g: gram.

Table 3. Recommended dietary daily intake for the sample ($N = 106$).

| Parameters, Unit, RDI | M \pm SD | Intake % | n (%) |
|------------------------------------|-----------------|--------------|----------|
| Carbohydrates, g, 45–65% calorie | 200 \pm 110 | $\geq 45\%$ | 75 (71%) |
| Protein, g, 10–35% calorie | 77 \pm 44 | $\geq 20\%$ | 38 (36%) |
| Total Fat, g, 20–35% calorie | 370 \pm 220 | $< 35\%$ | 69 (65%) |
| Saturated Fat, g, $< 10\%$ calorie | 19 \pm 11 | $< 10\%$ | 51 (48%) |
| Cholesterol, < 300 mg | 260 \pm 170 | $< 100\%$ | 78 (74%) |
| Sodium, < 2300 mg | 3000 \pm 1700 | $< 100\%$ | 40 (38%) |
| Fiber, ≥ 25 g | 19 \pm 10 | $\geq 100\%$ | 16 (15%) |
| Total Folate, 400 μ g | 370 \pm 220 | $\geq 100\%$ | 34 (32%) |
| Vitamin B1 (Thiamine), 1.1 mg | 1.4 \pm 0.8 | $\geq 100\%$ | 65 (61%) |
| Vitamin B2 (Riboflavin), 1.1 mg | 1.9 \pm 1.3 | $\geq 100\%$ | 78 (74%) |
| Vitamin B6, 1.3 mg | 1.8 \pm 1.0 | $\geq 100\%$ | 68 (64%) |
| Vitamin B12, 2.4 μ g | 6.1 \pm 8.2 | $< 150\%$ | 44 (42%) |
| Niacin, 14 mg | 21 \pm 12 | $\geq 100\%$ | 72 (68%) |
| Calcium, 1000 mg | 840 \pm 620 | $\geq 75\%$ | 46 (43%) |
| Magnesium, 320 mg | 300 \pm 160 | $\geq 75\%$ | 52 (49%) |
| Iron, 8 mg | 13 \pm 7.6 | $\geq 100\%$ | 44 (42%) |
| Zinc, 8 mg | 11 \pm 6.9 | $\geq 100\%$ | 53 (50%) |
| Methionine, 13 mg/Kg | 1.8 \pm 1.0 | $< 150\%$ | 45 (43%) |

Note: RDI: recommended daily intake, g: gram; mg: milligram, μ g: microgram, Kg: Kilogram.

3.3. Predictive Modeling for Healthy Eating—Generalized Regression Analysis

Four sets of models were tested for prediction of healthy eating based on HEI and GI scores: an HEI score greater than 80 (HEI 80) is a good HEI score, an HEI score of 77 and higher (HEI 77) is the median score for this sample, GI of 55 and lower (low and good GI), and GI of 53.8 (median score for this sample). All individual dietary parameters under HEI and RDI categories and demographic parameters were tested for variables of importance and predictive models. Eleven common parameters across the four scoring criteria (HEI 80, HEI 77, GI 55, and GI 53.8) were identified for the prediction of healthy eating. These 11 parameters include in sequence of presentation in these analyses: whole fruit (1 cup), milk or dairy alternative such as a soy drink (6 oz), whole grain (1 oz), saturated fat (15 g), oils and nuts (1 oz), empty calories (300), fiber (19 g), gender, gender interacting with cancer status (Group Ca), Group Ca, and dark greens (6 oz) (Supplementary Tables S1–S4). We presented the testing on all 11 common parameters in addition to the models with significant parameters to illustrate the differences between the models with misclassification rates for accuracy of prediction, AICc for fitness, and AUC for coverage.

Table 4 presents significant individual parameters for HEI 80 prediction. A baseline LR model with validation was constructed with five significant individual parameters; all five parameters are HEI items: whole fruit (1 cup), milk or soy drink (6 oz), whole grain (1 oz), saturated fat (15 g), and oil nut intakes (1 oz) (all $p < 0.05$, amount per component representing the medians of this sample), with no significant parameters from other categories of demographic or RDI parameters. The results of baseline LR with validation are shown in the left panel of Table 4. Then, two GR models were developed using Adaptive Elastic Net with AICc validation and LOO cross validation methods to predict the probability of healthy eating with HEI 80 (the middle and right panels of Table 4). In both GR validation models, oil and nut intake did not present statistical significance. The GR AICc validation model presented as the best model with lowest misclassification rate and highest AUC, but higher AICc than the baseline LR model. The AUC as shown in Figure 1 with the baseline LR model presented 0.8333 and the GR Elastic Net AICc model and LOO model with AUC of 0.8674 and 0.8671, respectively.

Compared to the 11-parameter model that included all significant parameters for all models combined (Supplementary Table S1, Supplementary Figure S1), the 5-parameter model in Table 4 presented better model quality with smaller AICc (better) with fewer parameters (58 versus 75 for LR and 105 versus 113 for the GR AICc validation) and lower misclassification rate (better) (0.30 versus 0.32 for LR). The 5- and 11-parameter models presented similar AUC across the LR and GR models, with increased (better) AUC for the LR model.

Table 4. Predictors of Healthy Eating Index (80): Baseline logistic regression and generalized regression Elastic Net models.

| | Logistic Regression with Validation | | Generalized Regression Elastic Net | | | |
|------------------------|-------------------------------------|----------|------------------------------------|----------|--------------------------|----------|
| | Estimate | $p(X^2)$ | AICc Validation | | Leave-One-Out Validation | |
| Parameters | Estimate | $p(X^2)$ | Estimate | $p(X^2)$ | Estimate | $p(X^2)$ |
| (Intercept) | 2.64 | 0.002 | 1.61 | 0.001 | 1.58 | 0.002 |
| Whole Fruit 1 cup | −2.51 | 0.002 | −1.90 | 0.0004 | −1.86 | 0.0006 |
| Milk Soy 6 oz | −2.62 | 0.002 | −1.86 | 0.0002 | −1.84 | 0.0002 |
| Whole Grain 1 oz | −2.44 | 0.007 | −2.28 | 0.001 | −2.26 | 0.002 |
| Sat Fat 15 g | 3.81 | 0.008 | 2.31 | 0.010 | 2.55 | 0.010 |
| Oil Nut 1 oz | −2.57 | 0.02 | −1.29 | 0.12 | −1.56 | 0.09 |
| Misclassification Rate | 0.30 | | 0.23 | | 0.23 | |
| AICc | 58 | | 105 | | n/a | |
| Area under the curve | 0.83 | | 0.87 | | 0.87 | |

Note. AICc: Akaike’s information criterion with corrections.

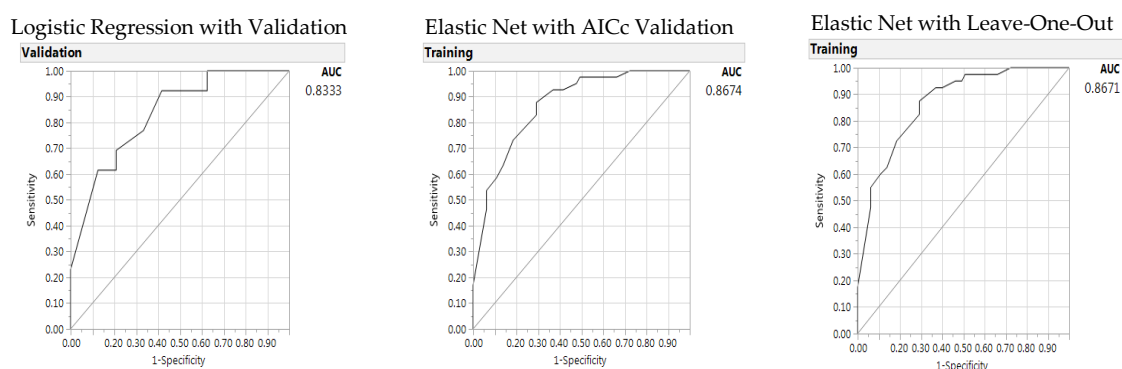


Figure 1. Predictors of the Healthy Eating Index (80): Area under the receiver operating characteristic curve (AUC) for logistic regression (left), Elastic Net with Akaike’s information criteria with correction (AICc) validation (middle) and Leave-One-Out validation models (right).

These models are then tested with the HEI score of 77 (HEI 77) as the median score of HEI for this study sample (Table 5). There is one significant interaction in addition to the six individual parameters

in the model for HEI 77 (Table 5): milk or soy drink (6 oz), whole grain (1 oz), empty calories (300), and fiber (19 g) as dietary parameters; gender and cancer/control status, and interaction of gender and cancer/control status. While cancer/control status as an individual parameter is not significant with respect to the *p* value, it must be included in the model because of its significant interaction with the gender status. The GR LOO validation model presents as the best model with the highest number of significant parameters, lowest misclassification rate for accuracy and highest AUC (Figure 2).

Table 5. Predictors of Healthy Eating Index (77): Baseline logistic regression and generalized regression Elastic Net models.

| Logistic Regression with Validation | | | Generalized Regression Elastic Net | | | |
|-------------------------------------|----------|--------------------|------------------------------------|--------------------|--------------------------|--------------------|
| | | | AICc Validation | | Leave-One-Out Validation | |
| Parameters | Estimate | <i>p</i> (X^2) | Estimate | <i>p</i> (X^2) | Estimate | <i>p</i> (X^2) |
| (Intercept) | 1.11 | 0.18 | 0.27 | 0.65 | 0.31 | 0.61 |
| Milk Soy 6 oz | −2.23 | 0.0008 | −1.82 | 0.0003 | −1.71 | 0.0006 |
| Whole Grain 1 oz | −1.23 | 0.10 | −1.30 | 0.02 | −1.37 | 0.01 |
| Empty Calories 300 | 1.21 | 0.11 | 1.21 | 0.03 | 1.10 | 0.048 |
| Fiber 19 g | 1.01 | 0.21 | 1.36 | 0.03 | 1.38 | 0.03 |
| Gender | −2.04 | 0.11 | −1.83 | 0.06 | −2.60 | 0.003 |
| GroupCa * Gender | 1.88 | 0.23 | 1.63 | 0.17 | 2.73 | 0.03 |
| GroupCa | −0.54 | 0.47 | 0.37 | 0.55 | 0.29 | 0.63 |
| Misclassification Rate | 0.27 | | 0.25 | | 0.23 | |
| AICc | 63 | | 122 | | n/a | |
| Area under the curve | 0.80 | | 0.83 | | 0.84 | |

Note. AICc: Akaike’s information criterion with corrections * interaction.

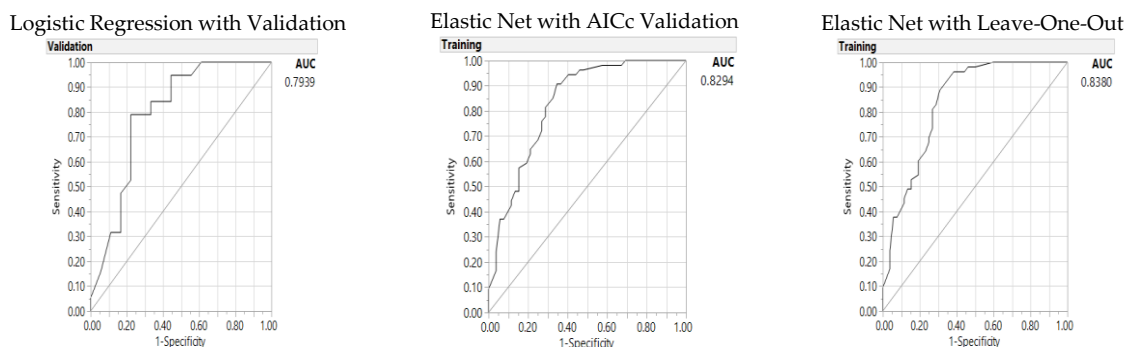


Figure 2. Predictors of the Healthy Eating Index (77): Area under the receiver operating characteristic curve (AUC) for baseline logistic regression (left), Elastic Net with Akaike’s information criteria with correction (AICc) validation (middle) and Leave-One-Out validation models (right).

In comparison to the 11-parameter model (Supplementary Table S2 and Figure S2), the significance model in Table 5 presents better fitness with lower AICc (63 versus 69 for LR); while the 11-parameter models present lower misclassification rates for both GR models (0.1604 versus 0.25 for the GR AICc validation and 0.1524 versus 0.23 for the GR LOO validation) and higher AUCs (0.86 versus 0.79 for LR, 0.90 versus 0.83 for GR AICc, and 0.92 versus 0.84 for GR LOO models). In comparison to the HEI 80, HEI 77 presented with lower misclassification rates, but higher AICc and lower AUC across LR and GR models.

The JMP profiler, shown in Figure 3a, and the interaction profiler shown in Figure 3b, are illustrative of how to interpret the interaction results. To illustrate, the excerpt of the interaction profiler depicts interactions between milk soy and gender, gender and cancer/control group status (group Ca), milk soy and cancer/control group status. Visually, the more non-parallel the two levels, the more likely there is a significant interaction between the two parameters. For example, we see in the milk soy and gender cell the lines or levels are almost parallel, indicating likely no-significant

interaction. However, for the gender with group Ca, there is a crossing of the two lines, indicating there is likely a statistically significant interaction effect between these parameters; a significant finding in the GR LOO validation ($p < 0.05$).

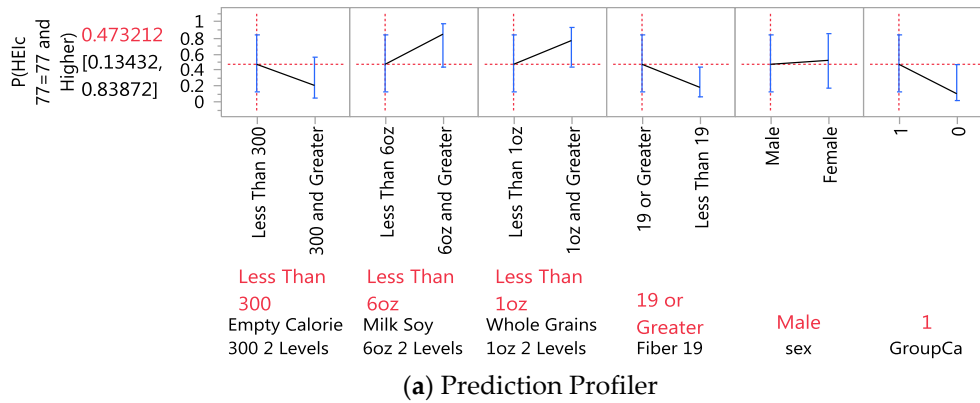


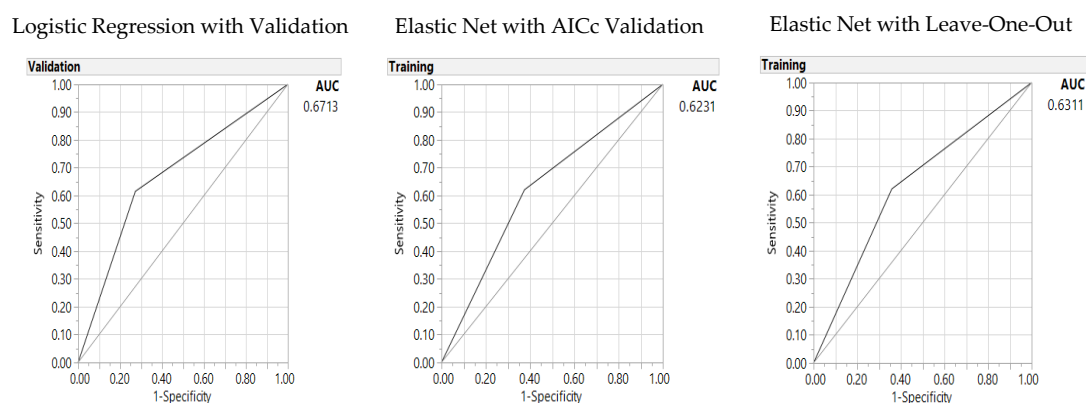
Figure 3. Prediction profiler (a) for significant predictors of health eating (score 77) and (b) interaction of gender with cancer/control group (non-parallel and crossing lines) when compared to another parameter (dairy or soy drink intake) without interaction (parallel lines).

The models are then tested with the GI score of 55 (GI 55), as the good GI score (Table 6). There is only one significant parameter: milk or soy drink in this model. LR outperformed two GR validation models for this one significant parameter model with the lowest misclassification rate, lower AICc, and highest AUC (Figure 4).

Table 6. Predictors of the Glycemic Index (55): Baseline logistic regression and generalized regression Elastic Net models.

| | Logistic Regression I with Validation | | Generalized Regression Elastic Net | | | |
|------------------------|---------------------------------------|----------|------------------------------------|----------|--------------------------|----------|
| | Estimate | $p(X^2)$ | AICc Validation | | Leave-One-Out Validation | |
| Parameters | Estimate | $p(X^2)$ | Estimate | $p(X^2)$ | Estimate | $p(X^2)$ |
| (Intercept) | 0.73 | 0.04 | 1.07 | 0.0005 | 1.01 | 0.0009 |
| Milk Soy 6 oz | −0.86 | 0.09 | −1.07 | 0.01 | −1.01 | 0.02 |
| Misclassification Rate | 0.35 | | 0.37 | | 0.38 | |
| AICc | 49 | | 139 | | n/a | |
| Area Under Curve | 0.67 | | 0.62 | | 0.63 | |

Note. AICc: Akaike's information criterion with corrections.

**Figure 4.** Predictors of the Glycemic Index (55): Area under the receiver operating characteristic curve (AUC) for baseline logistic regression (left), Elastic Net with Akaike's information criteria with correction (AICc) validation (middle) and Leave-One-Out validation models (right).

In comparison to the 11-parameter model (Supplementary Table S3 and Figure S3), the significance model in Table 6 presents better fitness with lower AICc (49 versus 87 for LR, and 139 versus 153 for GR AICc validation); while the 11-parameter models present lower misclassification rates for both GR models (0.16 versus 0.25 for the GR AICc validation and 0.15 versus 0.23 for the GR LOO validation) and higher AUCs (0.86 versus 0.79 for LR, 0.90 versus 0.83 for GR AICc, and 0.92 versus 0.84 for GR LOO models) and AUC for LR (0.67 versus 0.55).

Finally the models are then tested with the GI score of 53.8 (GI 53.8), as the median GI score (Table 7). Three dietary parameters under the HEI domain categories were significant parameters for GI 53.8: milk or soy drink empty calories, and dark greens. GR validation outperformed the LR model with lower misclassification rates and higher AUC (Figure 5).

In comparison to the 11-parameter model (Supplementary Table S4 and Figure S5), the significance model in Table 7 presents better fitness with lower AICc with fewer parameters in the model (62 versus 91 for LR, and 141 versus 149 for GR AICc validation) and slightly higher AUC for GR LOO model (0.717 versus 0.715; while the 11-parameter models present slightly higher AUC (0.63 versus 0.58 for LR, 0.72 versus 0.70 for GR AICc). In comparison to the GI 55 prediction, GI 53.8 predictive models present lower misclassification rates across LR and GR models and higher AUC for GR models. However, GI 55 models present lower AICc in both LR and GR AICc validation models with fewer parameters. In comparison with the two HEI models of HEI 80 and HEI 77, two GI models of GI 55 and GI 53.8 presented higher misclassification rates, higher AICc, and lower AUC across all LR and GR models; hence, the HEI models presented better quality models than the GI models.

Table 7. Predictors of Glycemic Index (53.8): Baseline logistic regression and generalized regression Elastic Net models.

| Logistic Regression with Validation | | | Generalized Regression Elastic Net | | | |
|-------------------------------------|----------|---------------|------------------------------------|---------------|--------------------------|---------------|
| | | | AICc Validation | | Leave-One-Out Validation | |
| Parameters | Estimate | p (X^2) | Estimate | p (X^2) | Estimate | p (X^2) |
| (Intercept) | 0.51 | 0.34 | 0.79 | 0.04 | 0.87 | 0.03 |
| Milk Soy 6 oz | −1.36 | 0.02 | −1.29 | 0.003 | −1.41 | 0.002 |
| Empty Calories 300 | 1.37 | 0.02 | 0.62 | 0.16 | 0.74 | 0.10 |
| Dark Green 6 oz | −1.18 | 0.04 | −0.94 | 0.03 | −1.05 | 0.02 |
| Misclassification Rate | 0.38 | | 0.34 | | 0.33 | |
| AICc | 62 | | 141 | | n/a | |
| Area Under Curve | 0.58 | | 0.70 | | 0.72 | |

Note. AICc: Akaike’s information criterion with corrections.

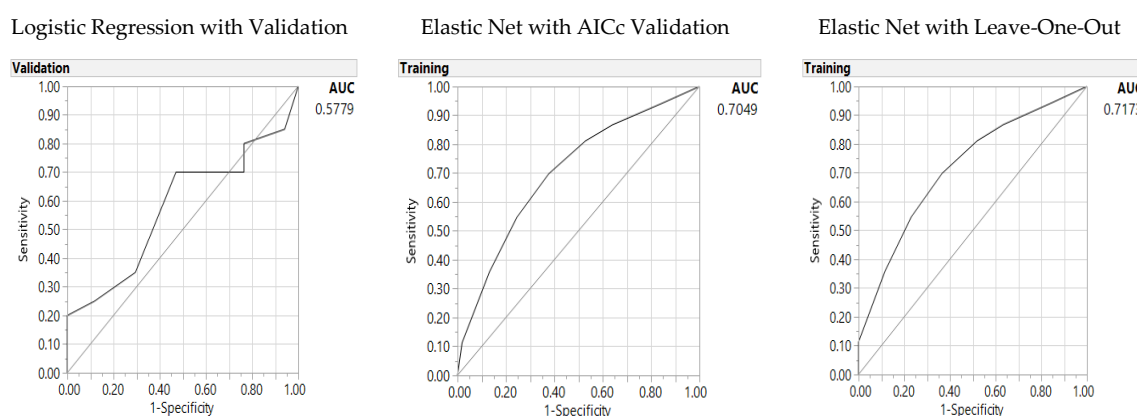


Figure 5. Predictors of the Glycemic Index (53.8): Area under the receiver operating characteristic curve (AUC) for baseline logistic regression (left), Elastic Net with Akaike’s information criteria with correction (AICc) validation (middle) and Leave-One-Out validation models (right).

4. Discussion

We presented a ground-breaking study, to cross-validate the results using both conventional LR statistics, with machine learning-based analytics, including the ensemble method and GR validation methods to predict healthy eating in diverse multi-ethnic families with CRC patients. While previous studies presented higher HEI scores in association with lower risks of CRC [19–21,73], we further documented the sensitivity of the HEI scale with median split distribution (a score of 77 versus 80) for predictive testing of healthy eating in association with CRC risk. Predictors of HEI 80 as the criterion score of a good diet included five significant dietary parameters (with intake amount): whole fruit (1 cup), milk or alternative-soy drinks for lactose intolerance (6 oz) [74], whole grain (1 oz), saturated fat (15 g), and oil and nuts (1 oz) for the diverse multi-ethnic sample of CRC families. Compared to the GI models, HEI models presented more accurate, fitted models, and greater coverage. Milk or alternative dairy for lactose intolerance [74] such as soy drinks (6 oz) is the common significant parameter across four HEI and GI predictive models.

Using SAS JMP programming (SAS Institute, Cary, NC, USA), we identified significant parameters of healthy eating in the diverse groups of families of CRC patients with their family members. As dietary habits can be modified, specific domain parameters for healthy eating can be helpful for these families to focus on key food items, with specific amounts for minimum intake levels or restricted intake levels. For a demonstration study of future dietary interventions, we used machine learning-based analytics, including ensemble methods and GR AICc and LOO validation models, for small-sample studies to validate the analyses by the random subsets of samples [75]. We further presented an

interaction profiler including 3-way interactions (interaction profile includes bi-variate interactions in association with the outcome) for the best quality and optimal model.

As part of prevention efforts, healthy eating is essential in personalized nutrition for nutrigenetics in providing methyl-donors to prevent CRC. Family members share dietary habits and lifestyles that affect epigenetics and nutrigenomics pathways affecting health outcomes [76]. For sustainable improvement of dietary modifications, as part of healthy lifestyles, the involvement of family members is vital to provide an essential support system within the families with heightened awareness of healthy eating within the family units [32,76]. Further studies with larger datasets and diverse samples are needed to further examine these findings in diverse groups for personalized nutrition in preparation for precision-based healthcare.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2072-6643/10/6/674/s1>, Table S1: Table 1. Predictors of Healthy Eating Index (80): Baseline logistic regression and generalized regression Elastic Net models including 11 common parameters, Table S2: Predictors of Healthy Eating Index (77): Baseline logistic regression and generalized regression Elastic Net models including 11 common parameters, Table S3: Predictors of Glycemic Index (55): Baseline logistic regression and generalized regression Elastic Net models including 11 common parameters, Table S4: Predictors of Glycemic Index (53.8): Baseline logistic regression and generalized regression Elastic Net models including 11 common parameters, Figure S1: Predictors of Healthy Eating Index (80), including 11 common parameters: Area under the receiver operating characteristic curve (AUC) for baseline logistic regression model (left panel), Elastic Net with Akaike's information criteria with correction (AICc) validation model (middle) and Leave-One-Out validation model (right panel), Figure S2: Predictors of Healthy Eating Index (77), including 11 common parameters: Area under the receiver operating characteristic curve (AUC) for baseline logistic regression model (left panel), Elastic Net with Akaike's information criteria with correction (AICc) validation model (middle) and Leave-One-Out validation model (right panel), Figure S3: Predictors of Glycemic Index (55) including 11 common parameters: Area under the receiver operating characteristic curve (AUC) for baseline logistic regression model (left panel), Elastic Net with Akaike's information criteria with correction (AICc) validation model (middle) and Leave-One-Out validation model (right panel), Figure S4: Predictors of Glycemic Index (53.8) including 11 common parameters: Area under the receiver operating characteristic curve (AUC) for baseline logistic regression model (left panel), Elastic Net with Akaike's information criteria with correction (AICc) validation model (middle) and Leave-One-Out validation model (right panel).

Author Contributions: Conceived the concepts and study design: S.P.K.S.; acquisition and search of the literature: A.L. and S.P.K.S.; data entry and verification of data accuracy: A.L., S.P.K.S.; analysis and interpretation of data: S.P.K.S., A.L., J.G., and C.H.Y.; wrote the first draft of the manuscript: S.P.K.S. and A.L. Agreed with manuscript results and conclusions: all authors reviewed and approved the final manuscript, ensuring integrity and accuracy.

Funding: Funding support included the Doctoral Research Council Grants, Azusa Pacific University and Research Start-up fund from Augusta University awarded to the corresponding author.

Acknowledgments: The authors acknowledge assistance from Haiyan Xiao, who retrieved a portion of the literature; and Joyce D. Kusuma, who helped with coding the Healthy Eating Index.

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Diet, Nutrition, Physical Activity and Colorectal Cancer. Available online: <http://www.wcrf.org/colorectal-cancer-2017> (accessed on 1 April 2018).
2. Ryan-Harshman, M.; Aldoori, W. Diet and colorectal cancer: Review of the evidence. *Can. Fam. Phys.* **2007**, *53*, 1913–1920.
3. Turati, F.; Bravi, F.; Di Maso, M.; Bosetti, C.; Polesel, J.; Serraino, D.; Dalmartello, M.; Giacosa, A.; Montella, M.; Tavani, A.; et al. Adherence to the World Cancer Research Fund/American Institute for Cancer Research recommendations and colorectal cancer risk. *Eur. J. Cancer* **2017**, *85*, 86–94. [[CrossRef](#)] [[PubMed](#)]
4. Tabung, F.K.; Brown, L.S.; Fung, T.T. Dietary Patterns and Colorectal Cancer Risk: A Review of 17 Years of Evidence (2000–2016). *Curr. Colorectal Cancer Rep.* **2017**, *13*, 440–454. [[CrossRef](#)] [[PubMed](#)]
5. Vieira, A.R.; Abar, L.; Chan, D.S.M.; Vingeliene, S.; Polemiti, E.; Stevens, C.; Greenwood, D.; Norat, T. Foods and beverages and colorectal cancer risk: A systematic review and meta-analysis of cohort studies, an update of the evidence of the WCRF-AICR Continuous Update Project. *Ann. Oncol.* **2017**, *28*, 1788–1802. [[CrossRef](#)] [[PubMed](#)]

6. Shiao, S.P.K.; Lie, A.; Chong, H.Y. Meta-analysis of homocysteine-related factors on the risk of colorectal cancer. *Oncotarget* **2018**, 925681–925697. Available online: <http://www.oncotarget.com/index.php?journal=oncotarget&page=article&op=view&path%5B%5D=25355&path%5B%5D=79428> (accessed on 25 May 2018). [[CrossRef](#)]
7. Magalhães, B.; Peleteiro, B.; Lunet, N. Dietary patterns and colorectal cancer: Systematic review and meta-analysis. *Eur. J. Cancer Prev.* **2012**, *21*, 15–23. [[CrossRef](#)] [[PubMed](#)]
8. Fan, Y.; Jin, X.; Man, C.; Gao, Z.; Wang, X. Meta-analysis of the association between the inflammatory potential of diet and colorectal cancer risk. *Oncotarget* **2017**, *8*, 59592–59600. [[CrossRef](#)] [[PubMed](#)]
9. Tárraga López, P.J.; Albero, J.S.; Rodríguez-Montes, J.A. Primary and secondary prevention of colorectal cancer. *Clin. Med. Insights Gastroenterol.* **2014**, *7*, 33–46. [[CrossRef](#)] [[PubMed](#)]
10. Cavicchia, P.P.; Steck, S.E.; Hurley, T.G.; Hussey, J.R.; Ma, Y.; Ockene, I.S.; Hebert, J.R. A new dietary inflammatory index predicts interval changes in serum high-sensitivity C-reactive protein. *J. Nutr.* **2009**, *139*, 2365–2372. [[CrossRef](#)] [[PubMed](#)]
11. Shivappa, N.; Steck, S.E.; Hurley, T.G.; Hussey, J.R.; Hebert, J.R. Designing and developing a literature-derived, population-based dietary inflammatory index. *Public Health Nutr.* **2014**, *17*, 1689–1696. [[CrossRef](#)] [[PubMed](#)]
12. Johnson, C.M.; Wei, C.; Ensor, J.E.; Smolenski, D.J.; Amos, C.I.; Levin, B.; Berry, D.A. Meta-analyses of colorectal cancer risk factors. *Cancer Causes Control* **2013**, *24*, 1207–1222. [[CrossRef](#)] [[PubMed](#)]
13. American Institute for Cancer Research. How AICR Recommendations Cuts Colorectal Cancer Risk for Both Men and Women. Available online: http://www.aicr.org/cancer-research-update/2016/11_02/cru-how-AICR-recommendations-cuts-colorectal-cancer-risk-for-men-and-women.html (accessed on 4 April 2018).
14. Yuan, Y.; Li, F.; Dong, R.-H.; Chen, J.-S.; He, G.-S.; Li, S.-J.; Chen, B. The Development of a Chinese Healthy Eating Index and Its Application in the General Population. *Nutrients* **2017**, *9*, 977. [[CrossRef](#)] [[PubMed](#)]
15. United States (U.S.) Department of Health and Human Services and U.S. Department of Agriculture. 2015–2020 Dietary Guidelines for Americans. Available online: <http://health.gov/dietaryguidelines/2015/guidelines/> (accessed on 1 April 2018).
16. United States Department of Agriculture (USDA). Healthy Eating Index (HEI). (2016, November). Available online: <https://www.cnpp.usda.gov/healthyeatingindex> (accessed on 31 March 2018).
17. National Institute of Health (NIH). Nutrient Recommendations: Dietary Reference Intakes (DRI). Available online: https://ods.od.nih.gov/Health_Information/Dietary_Reference_Intakes.aspx (accessed on 1 April 2018).
18. Panizza, C.E.; Shvetsov, Y.B.; Harmon, B.E.; Wilkens, L.R.; Le Marchand, L.; Haiman, C.; Reedy, J.; Boushey, C.J. Testing the Predictive Validity of the Healthy Eating Index-2015 in the Multiethnic Cohort: Is the Score Associated with a Reduced Risk of All-Cause and Cause-Specific Mortality? *Nutrients* **2018**, *5*, 452. [[CrossRef](#)] [[PubMed](#)]
19. Djuric, Z.; Severson, R.K.; Kato, I. Association of dietary quercetin with reduced risk of proximal colon cancer. *Nutr. Cancer* **2012**, *64*, 351–360. [[CrossRef](#)] [[PubMed](#)]
20. Miller, P.E.; Lazarus, P.; Lesko, S.M.; Muscat, J.E.; Harper, G.; Cross, A.J.; Sinha, R.; Ryczak, K.; Escobar, G.; Mauger, D.T.; et al. Diet index-based and empirically derived dietary patterns are associated with colorectal cancer risk. *J. Nutr.* **2010**, *140*, 1267–1273. [[CrossRef](#)] [[PubMed](#)]
21. Reedy, J.; Wirfält, E.; Flood, A.; Mitrou, P.N.; Krebs-Smith, S.M.; Kipnis, V.; Midthune, D.; Leitzmann, M.; Hollenbeck, A.; Schatzkin, A.; et al. Comparing 3 dietary pattern methods—cluster analysis, factor analysis, and index analysis—With colorectal cancer risk: The NIH-AARP Diet and Health Study. *Am. J. Epidemiol.* **2010**, *171*, 479–487. [[CrossRef](#)] [[PubMed](#)]
22. Lucock, M.; Yates, Z.; Martin, C.; Choi, J.H.; Beckett, E.; Boyd, L.; LeGras, K.; Ng, X.; Skinner, V.; Wai, R.; et al. Methylation diet and methyl group genetics in risk for adenomatous polyp occurrence. *BBA Clin.* **2015**, *3*, 107–112. [[CrossRef](#)] [[PubMed](#)]
23. de Vogel, S.; Wouters, K.A.; Gottschalk, R.W.; van Schooten, F.J.; de Goeij, A.F.; de Bruïne, A.P.; Goldbohm, R.A.; van den Brandt, P.A.; van Engeland, M.; Weijnenberg, M.P. Dietary methyl donors, methyl metabolizing enzymes, and epigenetic regulators: Diet-gene interactions and promoter CpG island hypermethylation in colorectal cancer. *Cancer Causes Control* **2011**, *22*, 1–12. [[CrossRef](#)] [[PubMed](#)]

24. Sharp, L.; Little, J.; Brockton, N.T.; Cotton, S.C.; Masson, L.F.; Haites, N.E.; Cassidy, J. Polymorphisms in the methylenetetrahydrofolate reductase (MTHFR) gene, intakes of folate and related B vitamins and colorectal cancer: A case-control study in a population with relatively low folate intake. *Br. J. Nutr.* **2008**, *99*, 379–389. [[CrossRef](#)] [[PubMed](#)]
25. Oh, K.; Willett, W.C.; Fuchs, C.S.; Giovannucci, E.L. Glycemic index, glycemic load, and carbohydrate intake in relation to risk of distal colorectal adenoma in women. *Cancer Epidemiol. Biomarkers Prev.* **2004**, *13*, 1192–1198. [[PubMed](#)]
26. Neuhouser, M.L.; Schwarz, Y.; Wang, C.; Breymeyer, K.; Coronado, G.; Wang, C.Y.; Noar, K.; Song, X.; Lampe, J.W. A low-glycemic load diet reduces serum C-reactive protein and modestly increases adiponectin in overweight and obese adults. *J. Nutr.* **2012**, *142*, 369–374. [[CrossRef](#)] [[PubMed](#)]
27. Li, H.L.; Yang, G.; Shu, X.O.; Xiang, Y.B.; Chow, W.H.; Ji, B.T.; Zhang, X.; Cai, H.; Gao, J.; Gao, Y.T.; et al. Dietary glycemic load and risk of colorectal cancer in Chinese women. *Am. J. Clin. Nutr.* **2011**, *93*, 101–107. [[CrossRef](#)] [[PubMed](#)]
28. Kabat, G.C.; Shikany, J.M.; Beresford, S.A.; Caan, B.; Neuhouser, M.L.; Tinker, L.F.; Rohan, T.E. Dietary carbohydrate, glycemic index, and glycemic load in relation to colorectal cancer risk in the Women’s Health Initiative. *Cancer Causes Control* **2008**, *19*, 1291–1298. [[CrossRef](#)] [[PubMed](#)]
29. Sieri, S.; Krogh, V.; Agnoli, C.; Ricceri, F.; Palli, D.; Masala, G.; Panico, S.; Mattiello, A.; Tumino, R.; Giurdanella, M.C.; et al. Dietary glycemic index and glycemic load and risk of colorectal cancer: Results from the EPIC-Italy study. *Int. J. Cancer* **2015**, *136*, 2923–2931. [[CrossRef](#)] [[PubMed](#)]
30. Chang, K.T.; Lampe, J.W.; Schwarz, Y.; Breymeyer, K.L.; Noar, K.A.; Song, X.; Neuhouser, M.L. Low glycemic load experimental diet more satiating than high glycemic load diet. *Nutr. Cancer* **2012**, *64*, 666–673. [[CrossRef](#)] [[PubMed](#)]
31. Hastert, T.A.; White, E. Association between meeting the WCRF/AICR cancer prevention recommendations and colorectal cancer incidence: Results from the VITAL cohort. *Cancer Causes Control* **2016**, *27*, 1347–1359. [[CrossRef](#)] [[PubMed](#)]
32. Shiao, S.P.K.; Grayson, J.; Yu, C.H.; Wasek, B.; Bottiglieri, T. Gene Environment Interactions and Predictors of Colorectal Cancer in Family-Based, Multi-Ethnic Groups. *J. Personal. Med.* **2018**, *8*, 10. [[CrossRef](#)] [[PubMed](#)]
33. Holden, D.J.; Harris, R.; Porterfield, D.S.; Jones, D.E.; Morgan, L.C.; Reuland, D.; Gilehrst, M.; Viswanathan, M.; Lohr, K.N.; Lynda-McDonald, B. Enhancing the use of quality of colorectal cancer screening. *Evid. Rep. Technol. Assess.* **2010**, *190*, 1–195.
34. Visser, A.; Vrieling, A.; Murugesu, L.; Hoogerbrugge, N.; Kampman, E.; Hoedjes, M. Determinants of adherence to recommendations for cancer prevention among Lynch Syndrome mutation carriers: A qualitative exploration. *PLoS ONE* **2017**, *12*, e0178205. [[CrossRef](#)] [[PubMed](#)]
35. Campbell, P.T.; Curtin, K.; Ulrich, C.M.; Samowitz, W.S.; Bigler, J.; Velicer, C.M.; Caan, B.; Potter, J.D.; Slattery, M.L. Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle factors. *Gut* **2009**, *58*, 661–667. [[CrossRef](#)] [[PubMed](#)]
36. Simidjievski, N.; Todorovski, L.; Džeroski, S. Modeling dynamic systems with efficient ensembles of process-based models. *PLoS ONE* **2016**, *11*, e0153507. [[CrossRef](#)] [[PubMed](#)]
37. Khalilia, M.; Chakraborty, S.; Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inform. Decis. Mak.* **2011**, *11*, 51. [[CrossRef](#)] [[PubMed](#)]
38. Islam, M.M.; Yao, X.; Shahriar Nirjon, S.M.; Islam, M.A.; Murase, K. Bagging and boosting negatively correlated neural networks. *IEEE Trans. Syst. Man. Cybern. B Cybern.* **2008**, *38*, 771–784. [[CrossRef](#)] [[PubMed](#)]
39. Wang, C.W. New ensemble machine learning method for classification and prediction on gene expression data. In Proceedings of the EMBS ’06, 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 30 August–3 September 2006; Volume 1, pp. 3478–3481.
40. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [[CrossRef](#)] [[PubMed](#)]
41. Song, L.; Langfelder, P.; Horvath, S. Random generalized linear model: A highly accurate and interpretable ensemble predictor. *BMC Bioinform.* **2013**, *14*, 5. [[CrossRef](#)] [[PubMed](#)]
42. Witten, D.M.; Tibshirani, R. Covariance-regularized regression and classification for high-dimensional problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **2009**, *71*, 615–636. [[CrossRef](#)] [[PubMed](#)]
43. Wu, Y. Elastic Net for Cox’s proportional hazards model with a solution path algorithm. *Stat. Sin.* **2012**, *22*, 27–294. [[CrossRef](#)] [[PubMed](#)]

44. Krist, A.H.; Glenn, B.A.; Glasgow, R.E.; Balasubramanian, B.A.; Chambers, D.A.; Fernandez, M.E.; Heurtin-Roberts, S.; Kessler, R.; Ory, M.G.; Phillips, S.M.; et al. Designing a valid randomized pragmatic primary care implementation trial: The my own health report (MOHR) project. *Implement. Sci.* **2013**, *8*, 73. [CrossRef] [PubMed]
45. CDC. National Health and Nutrition Examination Survey. Center for Disease Control and Prevention. 2012. Available online: http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm (accessed on 15 December 2012).
46. National Coalition for Health Professional Education in Genetics. Family History Educational Aids. NCHPEG. Available online: http://www.nchpeg.org/index.php?option=com_content&view=article&id=145&Itemid=64 (accessed on 10 October 2016).
47. About Glycemic Index. Available online: <http://www.glycemicindex.com/about.php> (accessed on 1 April 2018).
48. Glycemic Index for 60+ foods. Available online: <https://www.health.harvard.edu/diseases-and-conditions/glycemic-index-and-glycemic-load-for-100-foods> (accessed on 1 April 2018).
49. Neuhouser, M.L.; Kristal, A.R.; McLerran, D.; Patterson, R.E.; Atkinson, J. Validity of short food frequency questionnaires used in cancer chemoprevention trials: Results from the Prostate Cancer Prevention Trial. *Cancer Epidemiol. Biomarkers Prev.* **1999**, *8*, 721–725. [PubMed]
50. Patterson, R.E.; Kristal, A.R.; Tinker, L.F.; Carter, R.A.; Bolton, M.P.; Agurs-Collins, T. Measurement characteristics of the Women’s Health Initiative food frequency questionnaire. *Ann. Epidemiol.* **1999**, *9*, 178–187. [CrossRef]
51. Schakel, S.F.; Sievert, Y.A.; Buzzard, I.M. Sources of data for developing and maintaining a nutrient database. *J. Am. Diet. Assoc.* **1988**, *88*, 1268–1271. [PubMed]
52. Harnack, L.; Lee, S.; Schakel, S.F.; Duval, S.; Luepker, R.V.; Arnett, D.K. Trends in the trans-fatty acid composition of the diet in a metropolitan area: The Minnesota Heart Survey. *J. Am. Diet. Assoc.* **2003**, *103*, 1160–1166. [CrossRef]
53. Bowman, S.A.; Lino, M.; Gerrior, S.A.; Basiotis, P.P. The Healthy Eating Index: 1994–96. Available online: https://www.cnpp.usda.gov/sites/default/files/healthy_eating_index/hei94-96report.PDF (accessed on 1 April 2018).
54. Glycemic Index and Diabetes. Available online: <http://www.diabetes.org/food-and-fitness/food/what-can-i-eat/understanding-carbohydrates/glycemic-index-and-diabetes.html> (accessed on 1 April 2018).
55. Atkinson, F.S.; Foster-Powell, K.; Brand-Miller, J.C. International tables of glycemic index and glycemic load values: 2008. *Diabetes Care* **2008**, *31*, 2281–2283. [CrossRef] [PubMed]
56. David, M. Revised International Table of Glycemic Index (GI) and Glycemic Load (GL) Values–2008. Available online: <http://www.mendosa.com/gilists.htm> (accessed on 1 April 2018).
57. Nutrient Recommendations: Dietary Reference Intakes (DRI). Available online: https://ods.od.nih.gov/Health_Information/Dietary_Reference_Intakes.aspx (accessed on 1 April 2018).
58. Grayson, J.; Gardner, S.; Stephens, M. *Building Better Models with JMP®Pro*. 2015; SAS Press: Cary, NC, USA, 2015.
59. Klimberg, R.; McCullough, B.D. *Fundamentals of Predictive Analytics with JMP*, 2nd ed.; SAS Press: Cary, NC, USA, 2016.
60. SAS Institute. Overview of the Generalized Regression Personality. 2017. Available online: <https://www.jmp.com/support/help/14/overview-of-the-generalized-regression-personali.shtml> (accessed on 1 April 2018).
61. Yu, C.H. Resampling: A Conceptual and Procedural Introduction. In *Best Practices in Quantitative Methods*; Osborne, J., Ed.; Sage Publications: Thousand Oaks, CA, USA, 2007; pp. 283–298.
62. Faraway, J.J. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models (Texts in Statistical Science)*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2005.
63. Meir, R.; Rätsch, G. An introduction to boosting and leveraging. In *Advanced Lectures on Machine Learning*; Mendelson, S., Smola, A.J., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2003; Volume 2600, pp. 118–183. [CrossRef]
64. Zaman, M.F.; Hirose, H. Classification performance of bagging and boosting type ensemble methods with small training sets. *New Gener. Comput.* **2011**, *29*, 277–292. [CrossRef]
65. SAS Institute Inc. *JMP 13 Fitting Linear Models*, 2nd ed.; SAS Institute Inc.: Cary, NC, USA, 2016.

66. Cheng, H.; Garrick, D.J.; Fernando, R.L. Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *J. Anim. Sci. Biotechnol.* **2017**, *8*, 38. [[CrossRef](#)] [[PubMed](#)]
67. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **2010**, *25*, 289–310. [[CrossRef](#)]
68. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer: New York, NY, USA, 2002.
69. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
70. Akaike, H. A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Stat. Math.* **1978**, *30*, 9–14. [[CrossRef](#)]
71. Burnham, K.P.; Anderson, D.R. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [[CrossRef](#)]
72. Yang, Y. Can the strengths of AIC and BIC be shared? *Biometrika* **2005**, *92*, 937–950. [[CrossRef](#)]
73. Park, S.Y.; Boushey, C.J.; Wilkens, L.R.; Haiman, C.A.; Le Marchand, L. High-quality diets associate with reduced risk of colorectal cancer: Analyses of diet quality indexes in the multiethnic cohort. *Gastroenterology* **2017**, *153*, 386–394. [[CrossRef](#)] [[PubMed](#)]
74. Andrzej, P.; Piotr, M.; Borun, P.; Skrzypczak-Zielinska, M.; Wojciechowska-Lacka, A.; Godlewski, D.; Banasiewicz, T. Influence of lactose intolerance on colorectal cancer incidence in the Polish population. *Hered. Cancer Clin. Pract.* **2015**, *13*, 7. [[CrossRef](#)]
75. Diaconis, P.; Efron, B. Computer-intensive methods in statistics. *Sci. Am.* **1983**, *248*, 116–130. [[CrossRef](#)]
76. Cenit, M.C.; Olivares, M.; Codoñer-Franch, P.; Sanz, Y. Intestinal microbiota and celiac disease: Cause, consequence or co-evolution? *Nutrients* **2015**, *7*, 6900–6923. [[CrossRef](#)] [[PubMed](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).