

Article

Logistic LASSO Regression for Dietary Intakes and Breast Cancer

Archana J. McEligot^{1,*}, Valerie Poynor², Rishabh Sharma³ and Anand Panangadan⁴ ¹ Department of Public Health, California State University, Fullerton, CA 92834, USA² Department of Mathematics, California State University, Fullerton, CA 92834, USA; vpoynor@fullerton.edu³ Department of Computer Science, University of Houston, Houston, TX 77004, USA; rsharma26@uh.edu⁴ Department of Computer Science, California State University, Fullerton, CA 92834, USA; apanangadan@fullerton.edu

* Correspondence: amceligot@fullerton.edu; Tel.: +1-657-278-3822

Received: 30 July 2020; Accepted: 27 August 2020; Published: 31 August 2020



Abstract: A multitude of dietary factors from dietary fat to macro and micronutrients intakes have been associated with breast cancer, yet data are still equivocal. Therefore, utilizing data from the large, multi-year, cross-sectional National Health and Nutrition Examination Survey (NHANES), we applied a novel, modern statistical shrinkage technique, logistic least absolute shrinkage and selection operator (LASSO) regression, to examine the association between dietary intakes in women, ≥ 50 years, with self-reported breast cancer ($n = 286$) compared with women without self-reported breast cancer (1144) from the 1999–2010 NHANES cycle. Logistic LASSO regression was used to examine the relationship between twenty-nine variables, including dietary variables from food, as well as well-established/known breast cancer risk factors, and to subsequently identify the most relevant variables associated with self-reported breast cancer. We observed that as the penalty factor (λ) increased in the logistic LASSO regression, well-established breast cancer risk factors, including age ($\beta = 0.83$) and parity ($\beta = -0.05$) remained in the model. For dietary macro and micronutrient intakes, only vitamin B12 ($\beta = 0.07$) was positively associated with self-reported breast cancer. Caffeine ($\beta = -0.01$) and alcohol ($\beta = 0.03$) use also continued to remain in the model. These data suggest that a diet high in vitamin B12, as well as alcohol use may be associated with self-reported breast cancer. Nonetheless, additional prospective studies should apply more recent statistical techniques to dietary data and cancer outcomes to replicate and confirm the present findings.

Keywords: diet; LASSO; breast cancer; NHANES

1. Introduction

In 2019, existing/prevalent cases of breast cancer in the United States reached more than 3.8 million, and approximately 42,000 women are expected to die from the disease in 2019 [1]. Although breast cancer incidence has been declining an average of 2.3% per year since 1990, 268,600 new cases of invasive breast cancer will be diagnosed in 2019, and breast cancer is still the second leading cause of cancer deaths for U.S. women. Nonetheless, it is only within the last 50 years that we have begun to investigate factors, including metabolic, dietary and other behavioral factors, that may be associated with breast cancer outcomes [2].

Modifiable risk factors such as dietary intakes, weight/obesity and physical activity, as well as alcohol consumption may independently influence a woman's risk of breast cancer diagnoses [3]. Many epidemiologic studies have examined the relationship between dietary intakes and cancer risk/incidence. Overall, studies suggest that a largely plant-based diet, high in vegetables, whole fruits and fiber, and low in calories, as well as lower obesity have been shown to be protective against

some types of cancers, particularly hormone-mediated cancers, including breast; but nonetheless, data are limited and inconclusive [3–9]. Further, due to the complexity and multitude of dietary variables, including whole dietary patterns of vegetable, fruit, fiber, fat, processed meats, sugary foods, lean meats, and a multitude of macro and micronutrients, most studies can and only have assessed either dietary patterns and/or single nutrients in relation to disease outcomes. Further, not only the volume of these modifiable factors, but also the multicollinearity between dietary macro and micronutrients may contribute to statistical challenges, specifically limiting the number and type of dietary variables included in the models, potentially resulting in conflicting findings and/or null results, and therefore warrant further investigation to elucidate the role of dietary intakes, both macro and micronutrients, with breast cancer.

The advent of big data science (BDs) has generated enormous amounts, varieties, and sources of complex data, and together with the availability of large open-source datasets and modern statistical techniques, has the vast potential for the creation of new knowledge, particularly in relation to primary and secondary disease prevention [10]. Beginning in 1999, the National Health and Nutrition Examination Survey (NHANES) began to continuously collect dietary, lifestyle, laboratory, physical examination, and other health information, amassing data on nearly 5000 participants per year on a multitude of health variables spanning over a decade [11]. Currently, NHANES is a large, open-source, publicly available dataset, which provides a unique opportunity to examine large, complex dietary, and other health information, including breast cancer diagnoses.

With the availability of vast amounts of diet and health data, and in conjunction with recent statistical techniques, we have an emerging opportunity to elucidate novel associations, patterns and clusters not previously observed with traditional statistical approaches and smaller datasets, potentially contributing to and providing a more robust understanding to the literature on the role of diet and breast cancer. One specific modern technique, the least absolute shrinkage and selection operator (LASSO) has garnered much attention [12]. Traditional regression techniques are limited in the analysis and synthesis of large numbers of covariates, including multicollinear variables, but to date, a majority of the data on diet and breast cancer outcomes have utilized traditional statistical techniques. LASSO is a regression-based methodology permitting for a large number of covariates in the model, and importantly has the unique feature penalizing the absolute value of a regression coefficient; thus, regulating the impact a coefficient may have on the overall regression. The greater the penalization, the greater the shrinkage of coefficients, with some reaching 0, thus automatically removing unnecessary/uninfluential covariates [13–15].

Therefore, for the present study, we aimed to investigate via modern statistical techniques, specifically LASSO regression, the relationship between dietary intakes, obesity, and other risk factors on self-reported breast cancer. To our knowledge and in reviewing the diet and cancer literature, this is the first study to apply LASSO regression techniques to dietary intakes and breast cancer.

2. Materials and Methods

2.1. Study Design

Beginning in 1999, NHANES transitioned to a continuous ongoing cross-sectional survey conducted by the National Center for Health Statistics at the Centers for Disease Control and Prevention (CDC). The CDC, via NHANES, collects data on the health and nutritional status of noninstitutionalized U.S. adults and children using probabilistic, multistage sampling and oversampling to achieve a nationally representative sample of the U.S. population. The interview component of NHANES ascertains information on demographic, socioeconomic, and health-related factors and includes a 24-h dietary recall assessment. Detailed information on survey design and methodology have been previously published [11].

2.2. Sample

The 1999–2010 cycle of NHANES included a nationally representative population of 62,160 with 59,367 participants with dietary data. We limited our analyses to adults ≥ 50 years ($n = 14,770$) who had demographic data, participated in dietary assessment and had medical conditions and reproductive data ($n = 14,770$), and then further honed our analyses to female participants, ≥ 50 years with non-missing data on primary breast cancer diagnoses, reproductive, and dietary data ($n = 7426$). Our final study sample of women with breast cancer included women ≥ 50 years with self-reported breast cancer and no other cancers, as well as complete/non-missing demographic, reproductive, and dietary data ($n = 286$); per the NHANES definition. This required that all relevant variables associated with the 24-h dietary recall are non-missing and have a value. Our initial sample of women without breast cancer included women ≥ 50 years with no self-reported breast cancer and no other cancers, as well as complete/non-missing demographic, reproductive, and dietary data ($n = 5372$). In addition, of the women without self-reported breast cancer for the present analyses ($n = 5372$), we further random-sampled for optimal statistical power, maintaining an approximate optimal ratio of up to 1:4 for cases [16], resulting in a final study sample of women with self-reported breast cancer ($n = 286$) and women without self-reported breast cancer ($n = 1144$) (Figure 1). The CDC Institutional Review Board approved NHANES and all participants provided written informed consent. The study protocol review was conducted and approved by the Internal Review Board (IRB) of the California State University, Fullerton (HSR# 18-19-250).

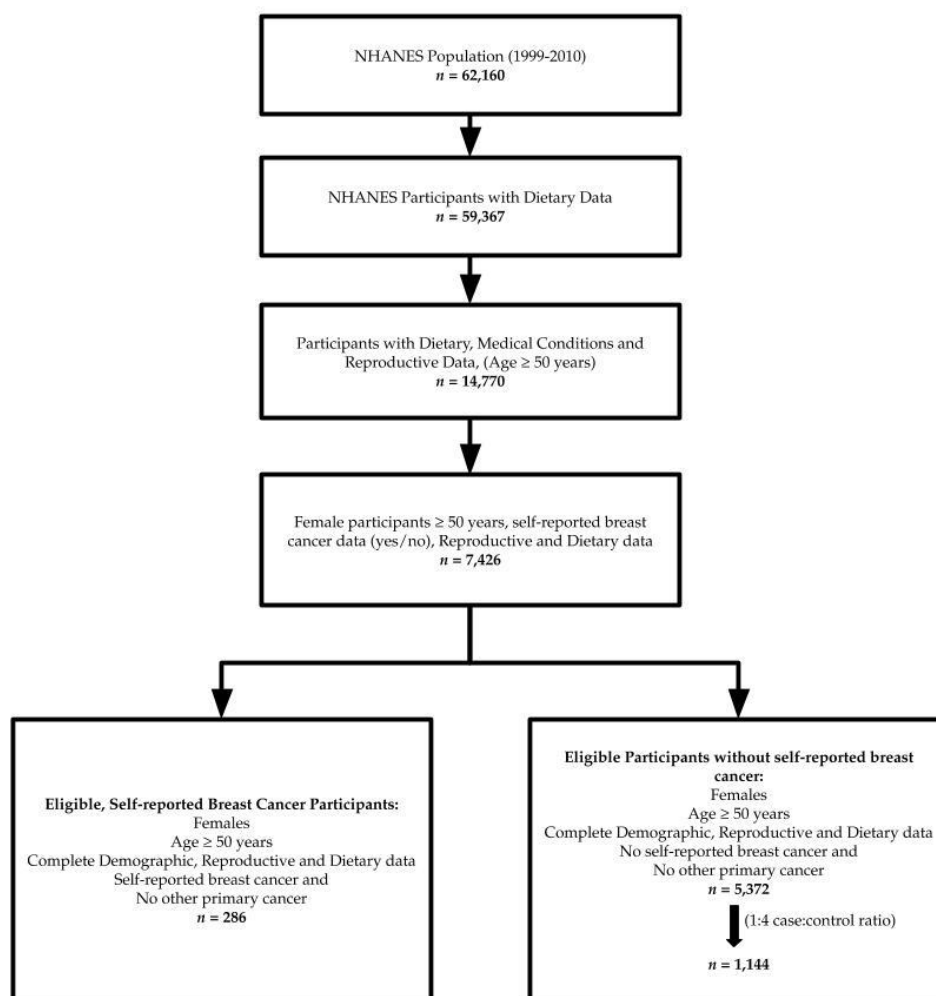


Figure 1. Flowchart of study participants.

2.3. Breast Cancer Data

In the “Medical Conditions” portion of the NHANES interview, health conditions and medical history were collected on adults, including cancer malignancies. Participants were queried on the following: “Have you ever been told by a doctor or other health professional that you had cancer or a malignancy of any kind?” And then followed by, “What kind of cancer was it?” Age at diagnoses was also collected via the following question: “How old were you when breast cancer was first diagnosed?” Women reporting only breast cancer diagnoses were included in the present study.

2.4. Dietary Intake

Dietary macro- and micronutrient intakes, from food only, were obtained from the total nutrient intakes data set. Dietary intakes were reported via a 24-h dietary recall in which respondents reported individual foods (and drinks) consumed during the midnight-to-midnight 24-h period prior to the in-person dietary interview. Coding of interview data and conversion to total nutrient intakes were done by NHANES using the USDA Food and Nutrient Database for Dietary Studies, 5.0 (FNDDS 5.0) [17]. The FNDDS 5.0 nutrient values were based on the USDA National Nutrient Database for Standard Reference, release 24 [18].

2.5. Other Measures

Age and race/ethnicity were obtained from the demographic variables and sample weights data set. NHANES categorizes race/ethnicity into four groups: non-Hispanic white, Hispanic, African American and other/multi-racial. Body mass index [BMI (kg/m^2)] was obtained from the body measures data set. Parity was ascertained via the following NHANES question, “How many times have you been pregnant? (Again, be sure to count all your pregnancies including (current pregnancy), live births, miscarriages, stillbirths, tubal pregnancies or abortions)”. Alcohol and caffeine consumption were expressed as grams and mg per day, respectively, and obtained via 24-h dietary recall data.

2.6. Statistical Analysis

Statistical analyses were conducted in R Statistical Software (version 3.5.2). The “survey” package (version 4.0) in R was utilized for the univariate analyses accounting for the stratified, multistage probability cluster sampling design in NHANES. NHANES provides sampling weights for analytical purposes, which account for oversampling of certain subgroups, differences between the sample and the population due to nonresponse, and population sizes. More specifically, NHANES provides sampling weights to be used for dietary analyses, which also account for the fact that not all participants completed the dietary interview and that different days of the week were represented in the 24-h periods for which dietary intake was assessed. Our study sample consisted of 6 cycles of continuous NHANES data from 1999–2010, thus dietary weights were adjusted to reflect the U.S. population. Due to different reference populations, the 4-year dietary weights for 1999–2002 were adjusted, while the remaining cycles for 2003–2010 adjusted the 2-year dietary weights. The NHANES stratification variable (SDMVSTRA) and primary sampling unit variable (SDMVPSU) were incorporated according to the survey design to appropriately adjust the variance estimates.

Descriptive and dietary variables were tested for normality and were log-transformed as appropriate, including all dietary variables. Specifically, statistical analyses were conducted on log transformed variables, including univariate tests and LASSO regression; however, mean data and 95% confidence intervals shown are on the raw/non-log transformed data. Height and weight data were used to calculate BMI (kg/m^2). Univariate analyses, specifically *t*-test for continuous variables and chi-square for discrete variables, respectively, were performed for demographic data including age and ethnicity, as well as age at first menarche, parity, and BMI. Also, mean intakes for macro- and micronutrients (from food only) were calculated and *t*-test analyses, as well as respective 95% confidence intervals of the mean differences were conducted to examine differences in dietary nutrient intakes between

women with and without self-reported reported breast cancer diagnoses. All statistical tests were two-sided with 0.05 significance levels.

The logistic LASSO model is a shrinkage method that can actively select from a large and potentially multicollinear set of variables in the regression, resulting in a more relevant and interpretable set of predictors [12]. LASSO performs via a continuous shrinking operation, minimizing regression coefficients in order to reduce the likelihood of overfitting, however, the technique is computed so as to shrink the sum of the absolute value of regression coefficients, forcing and producing coefficients that are exactly 0, thus selecting for the nonzero variables to remain in the model.

We utilized the “glmnet” package (version 2.0-16) to fit the logistic LASSO regression. The dietary weights were normalized and incorporated in the same fashion as standard weighted regression [19]. The covariates were not standardized as this would cause the weighting structure to be lost as described previously [20]. But, briefly, McConville, 2011 [19] showed that for survey-weighted LASSO regression analysis, the covariates should not be standardized as the inverse inclusion weights associated with each participant would be lost. The inverse inclusion weights are the normalized sampling weights. Furthermore, our analysis was performed on the log scale of the covariates, which minimizes the range of the covariate values, thus no one covariate dominated in the model due to a larger/wider range. We utilized ten-fold cross-validation to select the penalty term, λ . The binomial deviance was computed for the test data as measures of the predictive performance of the fitted models. The built-in function in R produces two automatic λ 's—one that minimizes the binomial deviance and one representing largest λ that is still within 1 standard error of the minimum binomial deviance. We opted for the latter λ as it results in stricter penalty allowing us to reduce the number of covariates even further than the former λ . For the present analyses, the λ values ranged from 0.00009 to 0.06937 with a minimal binomial deviance achieved at 0.0035 and more stringent value of 0.0108 (Figure 2). The standard errors of the LASSO coefficients were obtained via bootstrapping within the primary sampling unit and strata [21].

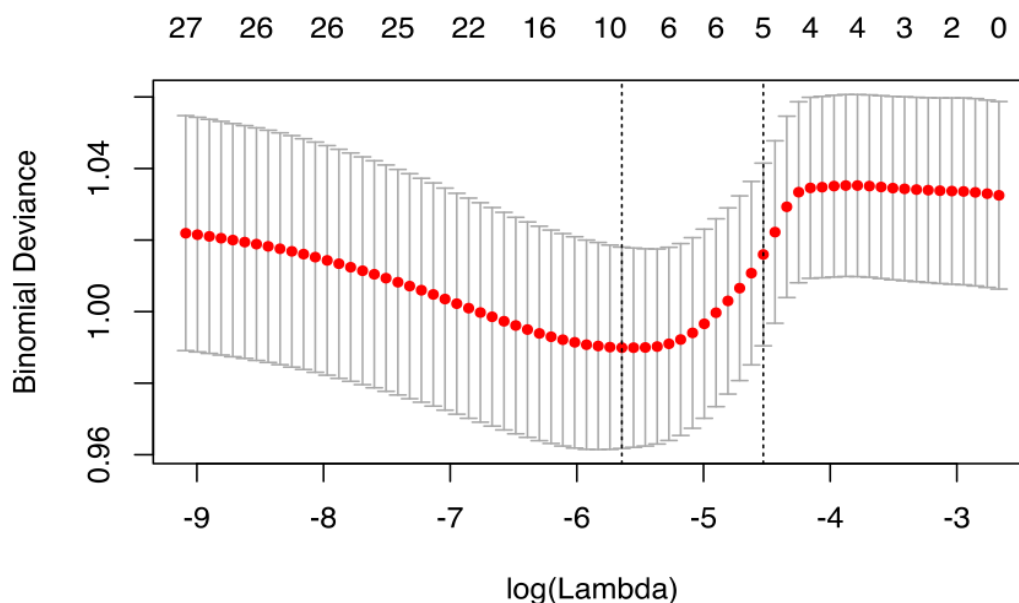


Figure 2. Cross validation plot for the penalty term.

For the logistic LASSO regression, self-reported breast cancer was included as the dependent variable, Y , and coded as 0 for no cancer and 1 for presence of breast cancer. Additionally, we included all 21 dietary variables from food, in addition to alcohol and caffeine consumption, available during the respective years of analyses, via the NHANES 24-h dietary recall data, including: energy (Kcal), % energy from carbohydrate, % energy from fat, % energy from protein, % energy from fat, cholesterol (mg), fiber (g), folate (μg), vitamin B12 (μg), vitamin B6 (mg), thiamin (vitamin B1, mg), riboflavin

(vitamin B2, mg), calcium (mg), phosphorous (mg), magnesium (mg), iron (mg), vitamin A (RE), vitamin C (mg), vitamin E (mg), zinc (mg), sodium (mg), potassium (mg), caffeine (mg), and alcohol (g). Energy dense macronutrients, including dietary fat, carbohydrate, and protein were adjusted for energy intakes and included in the logistic LASSO regression model as % energy of the respective macronutrient. All dietary variables were included as continuous variables in the model. All variables evaluated as potential confounders and specifically those shown to be previously associated with breast cancer risk (based on literature) were also included in the model: age (continuous), age at menarche (continuous), and parity (continuous). We also included BMI in our model, however BMI data were available only on a subset of participants, $n = 279$ (with self-reported breast cancer) and $n = 1116$ (without breast cancer), and analyses for these data were conducted on the respective sample size. We also examined time since breast cancer diagnoses (age at interview minus age at diagnoses), and conducted correlation analyses between the time since diagnoses variable and dietary intakes. Specifically, independent pairwise correlation tests were performed on the log scale of all dietary variables with time since diagnoses.

3. Results

Demographic and well-established breast cancer risk factor data (including respective % (\pm SD) or mean (\pm SD)) are shown in Table 1. Significant differences ($p \leq 0.05$) between women with self-reported breast cancer and women without were observed for age 68.46 (0.74) vs. 63.19 (0.36) years, age at first menarche (12.62 (0.13) vs. 12.89 (0.06) years), and ethnicity, where women with self-reported breast cancer were more likely to be older, had a younger age at menarche, were less parous, and were more likely to be to be non-Hispanic white compared with women without self-reported breast cancer (88% vs. 77%, respectively). Correlation analyses with dietary intakes and time since diagnoses showed no correlation between variables, except for alcohol ($r = 0.797$, $p = 0.013$) (data not shown).

Table 1. Descriptive and other characteristics in participants with and without self-reported breast cancer.

Descriptive Variable	Women with Self-Reported Breast Cancer ($n = 286$)	Women without Self-Reported Breast Cancer ($n = 1144$)	p -Value
Mean age (\pm SD)	68.46, (0.74)	63.19, (0.36)	<0.001
Parity, mean (\pm SD)	2.49, (0.17)	2.70, (0.07)	0.15
Age at first menarche, mean (\pm SD)	12.62, (0.13)	12.89, (0.06)	0.06
Ethnicity, n , (%)			
Non-Hispanic White	203, (88%)	595, (77%)	<0.001
Non-Hispanic Black	44, (7.2%)	219, (10.9%)	
Hispanic	32, (2.6%)	294, (8.1%)	
Unknown/Other	5, (2.3%)	36, (4.5%)	
BMI (kg/m^2) ¹ , mean (\pm SD)	28.89, (0.55)	29.38, (0.33)	0.43

¹ BMI variable, $n = 279$ with breast cancer, $n = 1116$ without breast cancer.

Table 2 presents data on dietary intakes between women with and without self-reported breast cancer. Univariate analyses suggest a statistically significant borderline higher intake of dietary vitamin B12 ($\mu\text{g}/\text{d}$) in women with self-reported breast cancer compared with women without self-reported breast cancer ((5.02 (0.75) vs. 4.17 (0.15), respectively; 95% CI: $(-0.632, 2.34)$; $p = 0.08$). Women with self-reported breast cancer also had higher alcohol (g) consumption ((5.31 (1.01) vs. 3.17 (0.49)) as well as vitamin A (IU) intakes ((685.55 (75.15), 648.52 (18.85)), however these variables did not reach statistical significance ($p = 0.19$).

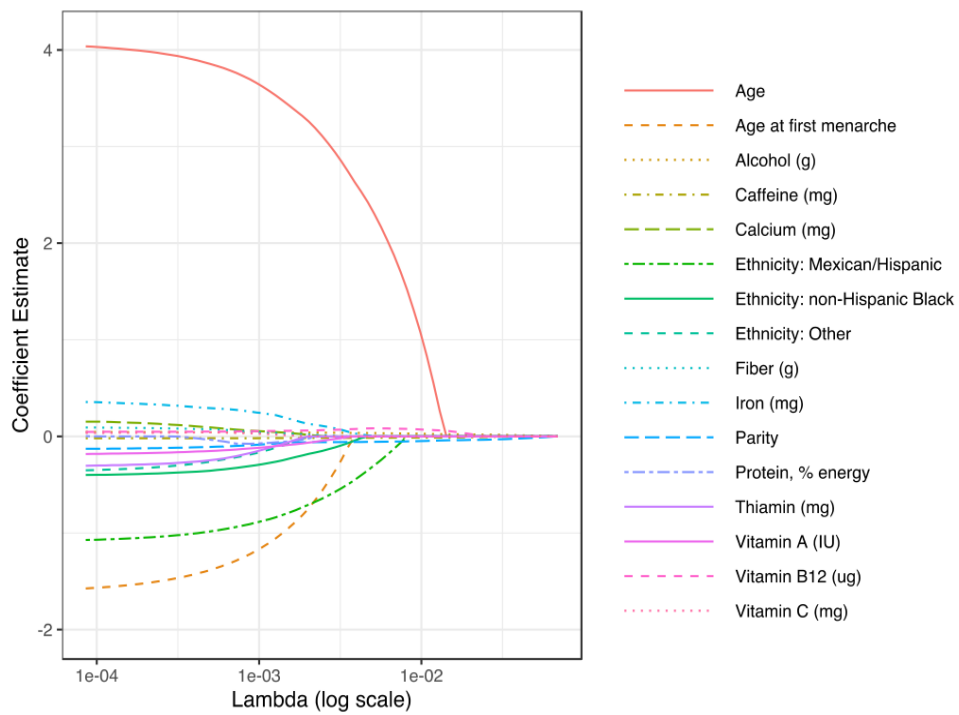
Table 2. Dietary macro- and micronutrient intakes in women with and without self-reported breast cancer.

Descriptive Variable Mean (SD) ¹	Women with Self-Reported Breast Cancer (n = 286)	Women without Self-Reported Breast Cancer (n = 1144)	95% CI (Difference of Means) ¹	p-Value
Energy (Kcal)	1638 (43.60)	1648 (28.23)	(−109.59, 91.01)	0.46
Carbohydrate (g)	205.38 (6.46)	204.75 (3.77)	(−14.34, 15.57)	0.36
Carbohydrate, % energy	50.43 (0.86)	50.44 (0.48)	(−2.00, 1.98)	0.80
Protein (g)	64.20 (2.38)	65.45 (1.31)	(−6.46, 3.97)	0.68
Protein, % energy	15.89 (0.30)	16.10 (0.18)	(−0.944, 0.520)	0.72
Total Fat (g)	61.86 (2.05)	63.99 (1.47)	(−6.83, 2.64)	0.50
Fat, % energy	33.57 (0.60)	33.97 (0.39)	(−1.88, 1.06)	0.78
Cholesterol (mg)	213.80 (10.56)	226.50 (7.70)	(−40.46, 15.05)	0.40
Fiber (g)	15.23 (0.72)	14.85 (0.34)	(−1.25, 2.02)	0.38
Folate (µg)	353.49 (13.81)	347.07 (8.40)	(−26.49, 39.34)	0.38
Vitamin B12 (µg)	5.02 (0.75)	4.17 (0.15)	(−0.632, 2.34)	0.08
Vitamin B6 (mg)	1.61 (0.07)	1.60 (0.04)	(−0.16, 0.19)	0.45
Thiamin (mg)	1.36 (0.07)	1.39 (0.04)	(−0.18, 0.15)	0.73
Riboflavin (mg)	1.89 (0.07)	1.88 (0.04)	(−0.164, 0.16)	0.45
Calcium (mg)	772.74 (27.24)	780.44 (20.44)	(−67.42, 52.03)	0.21
Phosphorous (mg)	1082 (35.87)	1096 (20.13)	(−91.45, 63.47)	0.49
Magnesium (mg)	253.27 (9.13)	256.33 (4.79)	(−24.00, 17.89)	0.65
Iron (mg)	13.24 (0.56)	12.84 (0.29)	(−0.93, 1.72)	0.25
Vitamin A (IU)	685.55 (75.15)	648.52 (18.85)	(−116.40, 190.45)	0.19
Vitamin C (mg)	87.68 (4.14)	92.04 (4.81)	(−16.66, 7.94)	0.28
Vitamin E (mg)	6.66 (0.40)	6.67 (0.19)	(−0.93, 0.91)	0.52
Zinc (mg)	9.76 (0.35)	9.61 (0.24)	(−0.61, 0.92)	0.24
Sodium (mg)	2665 (84.90)	2768 (59.30)	(−289.68, 81.87)	0.80
Potassium (mg)	2452 (61.65)	2476 (39.86)	(−158.56, 109.05)	0.38
Caffeine (mg)	154.56 (14.42)	174.94 (11.62)	(−57.97, 17.21)	0.38
Alcohol (g)	5.31 (1.01)	3.17 (0.49)	(0.08, 4.21) ²	0.19

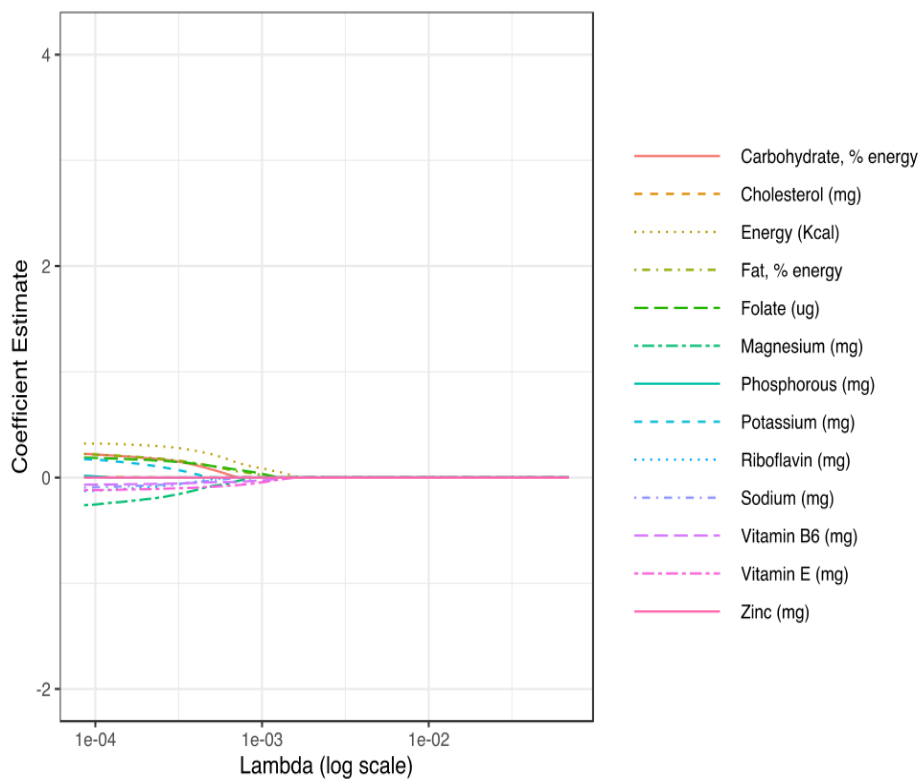
¹ Mean (SD) macro- and micronutrient, and 95% CI data shown are on the raw data; p-value data shown are on log transformed data; ² Log transformed 95% CI = (−0.20, 1.03).

Table 3 shows data adjusted for all the dietary variables, including macro- and micronutrient intakes, as well as well-established variables associated with breast cancer. The logistic LASSO regression results showed that of the well-established breast cancer risk factors, age ($\beta = 0.83$) and parity ($\beta = -0.05$) contributed to self-reported breast cancer. Specifically, age was positively associated with breast cancer, while parity was inversely associated. For dietary macro- and micronutrient intakes, only vitamin B12 ($\beta = 0.07$) was positively associated with self-reported breast cancer. Alcohol ($\beta = 0.03$) use also continued to remain in the model and was positively associated, while caffeine was inversely ($\beta = -0.01$) related to self-reported breast cancer diagnoses.

Figure 3a,b shows results on the 29 variables included in the LASSO regression and their corresponding coefficients for the different values of the penalty parameter. We observed that at $\lambda = 0.00009$, all 29 variables remain in the model (i.e., are nonzero). In Figure 3a, we show that 16 variables (age, age at first menarche, alcohol, caffeine, calcium, ethnicity, fiber, iron, parity, protein, thiamin, vitamin A, vitamin B12, and vitamin C) remained longest in the model as the penalty term increased with the other remaining variables approaching zero more quickly (Figure 3b). As λ increases to 0.01079, only five variables, potentially the most influential on self-reported breast cancer, remain in the model. Specifically, as λ approaches 0.01079, age, vitamin B12, caffeine, alcohol, and parity confer the largest signal in the model.



(a)



(b)

Figure 3. Plots for LASSO regression coefficients over different values of the penalty parameter. In (a), data shown are the sixteen variables that remained in the model the longest as the penalty term increased; in (b), data shown are the remaining variables in the model.

Table 3. The estimated coefficients for logistic least absolute shrinkage and selection operator (LASSO) regression between dietary data, and well-established breast cancer risk factors with self-reported breast cancer.

Variables	Coefficients (Bootstrap SE)
Well-established Variables	
Age (years)	0.83 (0.41)
Parity (# live births)	−0.05 (0.03)
Age at first menstrual cycle	0
Alcohol (g)	0.03 (0.02)
Other Variables	
Caffeine (mg)	−0.01 (0.02)
Mexican/Hispanic	0
Non-Hispanic Black	0
Other	0
Dietary Variables	
Energy (Kcal)	0
Carbohydrate, % energy	0
Protein, % energy	0
Fat, % energy	0
Cholesterol (mg)	0
Fiber (g)	0
Folate (μg)	0
Vitamin B12 (μg)	0.07 (0.05)
Vitamin B6 (mg)	0
Thiamin (Vitamin B1) (mg)	0
Riboflavin (Vitamin B2) (mg)	0
Calcium (mg)	0
Phosphorous (mg)	0
Magnesium (mg)	0
Iron (mg)	0
Vitamin A (RE)	0
Vitamin C (mg)	0
Vitamin E (mg)	0
Zinc (mg)	0
Sodium (mg)	0
Potassium (mg)	0

4. Discussion

Using a large, cross-sectional, nationally representative sample, in conjunction with modern robust statistical techniques, we applied logistic LASSO regression, which minimizes multicollinearity between dietary variables, to assess the relationship between dietary intakes and breast cancer diagnoses. Via LASSO, we also accounted for well-established breast cancer risk factors, while simultaneously selecting for relevant coefficients from a multitude of variables, ultimately removing all other unrelated variables. Our initial univariate analyses showed that age, age at first menarche, ethnicity, and also vitamin B12 from food, was related to breast cancer. In the ultimate logistic LASSO regression, well-established breast cancer risk factors, including older age and lower parity were associated with increased breast cancer, and vitamin B12, and alcohol and caffeine intakes were also related to self-reported breast cancer. Thus, we showed that increased alcohol consumption and reduced caffeine use were associated with an increase in breast cancer, with only vitamin B12 from diet remaining in the model. To the knowledge of the authors, this is the first study to utilize the powerful LASSO shrinkage technique to assess the relationship between the multitude of dietary variables and other risk factors with breast cancer diagnoses.

Our results of the association of well-established breast cancer risk factors, including age, race/ethnicity, age at menarche, and parity have been observed previously [22–27]. Age continued to remain in the model and was strongly related to breast cancer. In previous studies, aging has been clearly related to breast cancer diagnoses with a majority occurring in postmenopausal women, and more than 77% occurring for women 50 years of age or older [23,28]. Similar to our results, well-established reproductive factors, due to endogenous estrogen exposure, including early age at first menarche (<11 years), pregnancy (ever pregnant), and number of children have been shown to be linked to breast cancer risk [25–28].

Alcohol use has been consistently shown to increase breast cancer risk [3,29–32]. The International Agency for Research on Cancer (IARC) has designated alcohol as a carcinogenic risk factor [29], with the association being observed in both pre- and postmenopausal women. Earlier case/control studies showed increased risk with one reporting a 90% increase in breast cancer risk (OR: 1.9; 95% confidence interval, CI, 1.5–2.4) in ever drinkers compared with never drinkers [31,32], with subsequent epidemiologic studies establishing a positive association between increased quantity of alcohol consumption, showing a dose-response and causal relationship [29,33–36]. Our results are consistent with these previous findings, showing that even after accounting for a multitude of dietary factors, alcohol remained positively associated with breast cancer diagnoses. However, no previous studies have utilized statistical shrinkage techniques to assess the relationship between alcohol use and breast cancer.

Previous studies have shown that caffeine and/or coffee consumption may be associated with reduced breast cancer risk, but data remain equivocal [37–42]. There are several plausible mechanisms by which caffeine and/or coffee consumption may influence breast cancer risk, including the role of caffeine in estrogen metabolism, antioxidant actions of coffee/tea, and tumor differentiation and DNA methylation [43–47]. A recent study of 335,060 women participating in the European Prospective Investigation into Nutrition and Cancer (EPIC) Study reported that caffeinated coffee intake was associated with lower risk of postmenopausal breast cancer: adjusted HR = 0.90, 95% confidence interval (CI): 0.82 to 0.98, for high versus low consumption; $p_{\text{trend}} = 0.029$ [40]. Another large, longitudinal cohort study, the Nurses' Health Study, also showed a significant, yet weak, inverse association of caffeine intake with postmenopausal breast cancer for the highest quintile of intake compared to the lowest, RR: 0.88 (95% CI = 0.79 to 0.97, $p_{\text{trend}} = 0.03$) [43]. However, other prospective cohort studies have shown little to no association between caffeinated coffee intake and risk of breast cancer [41,42,48,49]. It is also plausible that other constituents in coffee and/or tea may either interact with caffeine and/or serve as a proxy in conferring protection against breast cancer [50–53], however our findings are consistent with the larger cohort studies in suggesting an inverse relationship between caffeine intake and breast cancer.

Our findings of a positive relationship between vitamin B12 and breast cancer diagnoses have been reported in previous studies [54–57]. However, other studies have found an inverse and/or no association [58,59]. In a prospective study of 936 incident breast cancer cases, dietary vitamin B12 was associated with increased risk of breast cancer (HR: Quartile 4 vs. Quartile 1 = 1.21 (1.00, 1.46); $p_{\text{trend}} = 0.06$) [55]. Recent EPIC cohort study investigations suggest a weak positive association between plasma vitamin B12 and breast cancer risk, however the association was attenuated by alcohol and/or folate status [54]. As a plausible mechanism, several water-soluble vitamins, including folate, vitamin B6, and vitamin B12 play a critical role in one-carbon metabolism, generating substrates for DNA methylation and DNA syntheses, and therefore modulate cancer risk [60–63]. Vitamin B12, primarily found in meat and dairy products, is involved in DNA methylation and may interfere with gene expression and function, whereby potentially conferring neoplastic cell growth [64,65]. Comparison of our findings with previous studies confirms a positive association of vitamin B12 with breast cancer, however data are inconclusive due to alcohol, folate, and/or epigenetic interactions and should be studied further.

Strengths of our study include the large sample size available via NHANES, which provides sufficient power to detect clinically relevant differences, and the generalizability of results due to

the nationally-representativeness of the NHANES survey data. Additionally, another strength is representation via oversampling of diverse racial/ethnic subgroups, such as Hispanics and African Americans, inherent in the NHANES survey methodology, enabling inclusion of groups that are often underrepresented in the scientific literature on diet and breast cancer. Limitations include the retrospective, cross-sectional design, which does not allow for causal inference, and self-reported data on diet and breast cancer. Further, it is feasible that women may have changed their diet post-diagnoses, which may influence findings, nonetheless, our correlation analyses of time since diagnoses found only one association, possibly due to multiple comparisons, between dietary intakes and time since diagnoses in women with self-reported breast cancer, potentially suggesting little to no change in diet from time since diagnoses. We also reported on dietary intakes from food only, and not on supplemental intake. Supplement data availability during the study time period were limited, and therefore only dietary data were examined in order to increase sample size and preserve power to detect associations.

5. Conclusions

In conclusion, to the knowledge of the authors, we showed for the first time, via a powerful shrinkage technique, that LASSO regression can be a viable option to narrow and decipher the role of a multitude of dietary factors and their relation to breast cancer diagnoses. We showed that established breast cancer factors, including age and parity continue to be associated with breast cancer diagnoses, and that alcohol use was positively associated, while caffeine intake was inversely related to breast cancer diagnoses. Our shrinkage analyses findings also suggest a potential role of dietary vitamin B12 intake and breast cancer diagnoses, however LASSO applications and use in assessing dietary intakes and breast cancer need to be confirmed in other prospective studies and warrant further investigation.

Author Contributions: Conceptualization, A.J.M.; methodology, A.J.M., V.P., R.S. and A.P.; formal Analysis, V.P., R.S.; data curation, A.J.M., R.S.; writing—original draft preparation, A.J.M.; writing—review and editing, A.J.M., V.P., R.S. and A.P.; supervision, A.J.M., A.P.; project administration, A.J.M.; funding acquisition, A.J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was part of the Big Data Discovery and Diversity through Research Education Advancement and Partnerships (BD3-REAP) Project funded by National Institutes of Health (NIH), NIMHHD-R25; # 1R25MD010397-01.

Conflicts of Interest: The authors declare no conflict of interest. The sponsors had no role in the design, execution, interpretation, or writing of the study.

References

1. American Cancer Society. *Breast Cancer Facts & Figures 2019–2020*; American Cancer Society, Inc.: Atlanta, GA, USA, 2019.
2. Patterson, R.E.; Rock, C.L.; Kerr, J.; Natarajan, L.; Marshall, S.J.; Pakiz, B.; Cadmus-Bertram, L.A. Metabolism and breast cancer risk: Frontiers in research and practice. *J. Acad. Nutr. Diet.* **2013**, *113*, 288–296. [[CrossRef](#)]
3. World Cancer Research Fund and American Institute for Cancer Research. Continuous Update Project Report Expert Report 2018. Diet, Nutrition, Physical Activity, and Breast Cancer. Available online: dietandcancerreport.org (accessed on 24 July 2020).
4. Farvid, M.S.; Chen, W.Y.; Rosner, B.A.; Tamimi, R.M.; Willett, W.C.; Eliassen, A.H. Fruit and vegetable consumption and breast cancer incidence: Repeated measures over 30 years of follow-up. *Int J. Cancer.* **2019**, *144*, 1496–1510. [[CrossRef](#)] [[PubMed](#)]
5. Kushi, L.H.; Doyle, C.; McCullough, M.; Rock, C.L.; Demark-Wahnefried, W.; Bandera, E.V.; Gapstur, S.; Patel, A.V.; Andrews, K.; Gansler, T. American Cancer Society 2010 Nutrition and Physical Activity Guidelines Advisory Committee. American Cancer Society Guidelines on nutrition and physical activity for cancer prevention: Reducing the risk of cancer with healthy food choices and physical activity. *CA Cancer J. Clin.* **2012**, *62*, 30–67. [[CrossRef](#)] [[PubMed](#)]
6. Chlebowski, R.T.; Luo, J.; Anderson, G.L.; Barrington, W.; Reding, K.; Simon, M.S.; Manson, J.E.; Rohan, T.E.; Wactawski-Wende, J.; Lane, D.; et al. Weight loss and breast cancer incidence in postmenopausal women. *Cancer* **2019**, *125*, 205–212. [[CrossRef](#)]

7. Hruby, A.; Manson, J.E.; Qi, L.; Malik, V.S.; Rimm, E.B.; Sun, Q.; Willett, W.C.; Hu, F.B. Determinants and Consequences of Obesity. *Am. J. Public Health*. **2016**, *106*, 1656–1662. [CrossRef]
8. Willett, W.C. Diet and breast cancer. *J. Intern. Med.* **2001**, *249*, 395–411. [CrossRef] [PubMed]
9. Hunter, D.J.; Spiegelman, D.; Adami, H.O.; Beeson, L.; van den Brandt, P.A.; Folsom, A.R.; Fraser, G.E.; Goldbohm, R.A.; Graham, S.; Howe, G.R. Cohort studies of fat intake and the risk of breast cancer—a pooled analysis. *N. Engl. J. Med.* **1996**, *334*, 356–361. [CrossRef]
10. Dolley, S. Big Data’s Role in Precision Public Health. *Front. Public Health* **2018**, *6*, 68. [CrossRef]
11. Zipf, G.; Chiappa, M.; Porter, K.S.; Ostchega, Y.; Lewis, B.G.; Dostal, J. National health and nutrition examination survey: Plan and operations, 1999–2010. *Vital Health Stat.* **2013**, *56*, 1–37.
12. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
13. Bickel, P.J.; Ritov, Y.; Tsybakov, A.B. Simultaneous analysis of Lasso and dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732. [CrossRef]
14. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–451.
15. Lee, S.; Kwon, S.; Kim, Y. A modified local quadratic approximation algorithm for penalized optimization problems. *Comput Stat. Data Anal.* **2016**, *94*, 275–286. [CrossRef]
16. Hennessy, S.; Bilker, W.B.; Berlin, J.A.; Strom, B.L. Factors influencing the optimal control-to-case ratio in matched case-control studies. *Am. J. Epidemiol.* **1999**, *149*, 195–197. [CrossRef] [PubMed]
17. United States Department of Agriculture. Agriculture Research Service. Available online: <http://www.ars.usda.gov/ba/bhnrc/fsrg> (accessed on 7 February 2017).
18. United States Department of Agriculture. Available online: <https://fdc.nal.usda.gov/> (accessed on 7 February 2017).
19. McConville, K.S. Improved Estimation for Complex Surveys Using Modern Regression Techniques. Ph.D. Thesis, Colorado State University, Fort Collins, CO, USA, 2011.
20. McConville, K.S.; Breidt, F.J.; Lee, T.C.; Moisen, G.G. Model-assisted survey regression estimation with the lasso. *J. Surv Stat. Method.* **2017**, *5*, 131–158. [CrossRef]
21. Efron, B.; Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*; Cambridge University Press: Cambridge, UK, 2016.
22. Coughlin, S.S. Epidemiology of Breast Cancer in Women. *Adv. Exp. Med. Biol.* **2019**, *1152*, 9–29. [CrossRef]
23. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef]
24. Tamimi, R.M.; Spiegelman, D.; Smith-Warner, S.A.; Wang, M.; Pazaris, M.; Willett, W.C.; Eliassen, A.H.; Hunter, D.J. Population Attributable Risk of Modifiable and Nonmodifiable Breast Cancer Risk Factors in Postmenopausal Breast Cancer. *Am. J. Epidemiol.* **2016**, *184*, 884–893. [CrossRef]
25. Dall, G.V.; Britt, K.L. Estrogen Effects on the Mammary Gland in Early and Late Life and Breast Cancer Risk. *Front. Oncol.* **2017**, *7*, 110. [CrossRef]
26. Gaudet, M.M.; Gierach, G.L.; Carter, B.D.; Luo, J.; Milne, R.L.; Weiderpass, E.; Giles, G.G.; Tamimi, R.M.; Eliassen, A.H.; Rosner, B.; et al. Pooled Analysis of Nine Cohorts Reveals Breast Cancer Risk Factors by Tumor Molecular Subtype. *Cancer Res.* **2018**, *78*, 6011–6021. [CrossRef]
27. Kelsey, J.L.; Gammon, M.D.; John, E.M. Reproductive factors and breast cancer. *Epidemiol. Rev.* **1993**, *15*, 36–47. [CrossRef] [PubMed]
28. Sineshaw, H.M.; Gaudet, M.; Ward, E.M.; Flanders, W.D.; Desantis, C.; Lin, C.C.; Jemal, A. Association of race/ethnicity, socioeconomic status, and breast cancer subtypes in the National Cancer Data Base (2010–2011). *Breast Cancer Res. Treat.* **2014**, *145*, 753–763. [CrossRef] [PubMed]
29. IARC working group on the evaluation of carcinogenic risks to humans. Alcohol Consumption and Ethyl Carbamate. *IARC Monogr. Eval. Carcinog. Risks Hum.* **2010**, *96*, 3.
30. Seitz, H.K.; Maurer, B. The relationship between alcohol metabolism, estrogen levels, and breast cancer risk. *Alcohol Res.* **2007**, *30*, 42–43.
31. Rosenberg, L.; Shapiro, S.; Slone, D.; Kaufman, D.; Helmrich, S.; Miettinen, O.; Engle JR, R. Breast cancer and alcoholic-beverage consumption. *Lancet* **1982**, *319*, 267–271. [CrossRef]
32. Talamini, R.; La Vecchia, C.; Decarli, A.; Franceschi, S.; Grattoni, E.; Grigoletto, E.; Tognoni, G. Social factors, diet and breast cancer in a northern Italian population. *Br. J. Cancer.* **1984**, *49*, 723–729. [CrossRef]
33. Coronado, G.D.; Beasley, J.; Livaudais, J. Alcohol consumption and the risk of breast cancer. *Salud Publica Mex.* **2011**, *53*, 440–447.

34. Pelucchi, C.; Tramacere, I.; Boffetta, P.; Negri, E.; La Vecchia, C. Alcohol consumption and cancer risk. *Nutr. Cancer* **2011**, *63*, 983–990. [[CrossRef](#)]
35. Chen, W.Y.; Rosner, B.; Hankinson, S.E.; Colditz, G.A.; Willett, W.C. Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk. *JAMA*. **2011**, *306*, 1884–1890. [[CrossRef](#)]
36. Secretan, B.; Straif, K.; Baan, R.; Grosse, Y.; El Ghissassi, F.; Bouvard, V.; Benbrahim-Tallaa, L.; Guha, N.; Freeman, C.; Galichet, L.; et al. A review of human carcinogens—Part E: Tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncol.* **2009**, *10*, 1033–1034. [[CrossRef](#)]
37. Baker, J.A.; Beehler, G.P.; Sawant, A.C.; Jayaprakash, V.; McCann, S.E.; Moysich, K.B. Consumption of coffee, but not black tea, is associated with decreased risk of premenopausal breast cancer. *J. Nutr.* **2006**, *136*, 166–171. [[CrossRef](#)]
38. Lowcock, E.C.; Cotterchio, M.; Anderson, L.N.; Boucher, B.A.; El-Sohehy, A. High coffee intake, but not caffeine, is associated with reduced estrogen receptor negative and postmenopausal breast cancer risk with no effect modification by CYP1A2 genotype. *Nutr. Cancer* **2013**, *65*, 398–409. [[CrossRef](#)] [[PubMed](#)]
39. Oh, J.K.; Sandin, S.; Ström, P.; Löf, M.; Adami, H.O.; Weiderpass, E. Prospective study of breast cancer in relation to coffee, tea and caffeine in Sweden. *Int. J. Cancer* **2015**, *137*, 1979–1989. [[CrossRef](#)] [[PubMed](#)]
40. Bhoo-Pathy, N.; Peeters, P.H.M.; Uiterwaal, C.S.P.M.; Bueno-de-Mesquita, H.B.; Bulgiba, A.M.; Bech, B.H.; Overvad, K.; Tjønneland, A.; Olsen, A.; Clavel-Chapelon, F.; et al. Coffee and tea consumption and risk of pre- and postmenopausal breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort study. *Breast Cancer Res.* **2015**, *17*, 15. [[CrossRef](#)] [[PubMed](#)]
41. Fagherazzi, G.; Touillaud, M.S.; Boutron-Ruault, M.C.; Clavel-Chapelon, F.; Romieu, I. No association between coffee, tea or caffeine consumption and breast cancer risk in a prospective cohort study. *Public Health Nutr.* **2011**, *14*, 1315–1320. [[CrossRef](#)]
42. Ishitani, K.; Lin, J.; Manson, J.E.; Buring, J.E.; Zhang, S.M. Caffeine consumption and the risk of breast cancer in a large prospective cohort of women. *Arch. Intern. Med.* **2008**, *168*, 2022–2031. [[CrossRef](#)]
43. Ganmaa, D.; Willett, W.C.; Li, T.Y.; Feskanich, D.; van Dam, R.M.; Lopez-Garcia, E.; Hunter, D.J.; Holmes, M.D. Coffee, tea, caffeine and risk of breast cancer: A 22-year follow-up. *Int. J. Cancer* **2008**, *122*, 2071–2076. [[CrossRef](#)]
44. Sisti, J.S.; Hankinson, S.E.; Caporaso, N.E.; Gu, F.; Tamimi, R.M.; Rosner, B.; Xu, X.; Ziegler, R.; Eliassen, A.H. Caffeine, coffee, and tea intake and urinary estrogens and estrogen metabolites in premenopausal women. *Cancer Epidemiol. Biomark. Prev.* **2015**, *24*, 1174–1183. [[CrossRef](#)]
45. Kotsopoulos, J.; Eliassen, A.H.; Missmer, S.A.; Hankinson, S.E.; Tworoger, S.S. Relationship between caffeine intake and plasma sex hormone concentrations in premenopausal and postmenopausal women. *Cancer* **2009**, *115*, 2765–2774. [[CrossRef](#)]
46. Halvorsen, B.L.; Carlsen, M.H.; Phillips, K.M.; Bøhn, S.K.; Holte, K.; Jacobs, D.R., Jr.; Blomhoff, R. Content of redox-active compounds (ie, antioxidants) in foods consumed in the United States. *Am. J. Clin. Nutr.* **2006**, *84*, 95–135. [[CrossRef](#)]
47. Rahmioglu, N.; Heaton, J.; Clement, G.; Gill, R.; Surdulescu, G.; Zlobecka, K.; Hodgkiss, D.; Ma, Y.; Hider, R.C.; Smith, N.W.; et al. Genetic epidemiology of induced CYP3A4 activity. *Pharmacogenet. Genom.* **2011**, *21*, 642–651. [[CrossRef](#)] [[PubMed](#)]
48. Gierach, G.L.; Freedman, N.D.; Andaya, A.; Hollenbeck, A.R.; Park, Y.; Schatzkin, A.; Brinton, L.A. Coffee intake and breast cancer risk in the NIH-AARP diet and health study cohort. *Int. J. Cancer* **2011**, *131*, 452–460. [[CrossRef](#)] [[PubMed](#)]
49. Boggs, D.A.; Palmer, J.R.; Stampfer, M.J.; Spiegelman, D.; Adams-Campbell, L.L.; Rosenberg, L. Tea and coffee intake in relation to risk of breast cancer in the Black Women’s Health Study. *Cancer Causes Control* **2010**, *21*, 1941–1948. [[CrossRef](#)]
50. Kotsopoulos, J.; Ghadirian, P.; El-Sohehy, A.; Lynch, H.T.; Snyder, C.; Daly, M.; Domchek, S.; Randall, S.; Karlan, B.; Zhang, P.; et al. The CYP1A2 genotype modifies the association between coffee consumption and breast cancer risk among BRCA1 mutation carriers. *Cancer Epidemiol. Biomark. Prev.* **2007**, *16*, 912–916. [[CrossRef](#)]
51. Hunter, D.J.; Manson, J.E.; Stampfer, M.J.; Colditz, G.A.; Rosner, B.; Hennekens, C.H.; Speizer, F.E.; Willett, W.C. A prospective study of caffeine, coffee, tea, and breast cancer. *Am. J. Epidemiol.* **1992**, *136*, 1000–1001.

52. Franceschi, S.; Favero, A.; La Vecchia, C.; Negri, E.; Dal Maso, L.; Salvini, S.; Decarli, A.; Giacosa, A. Influence of food groups and food diversity on breast cancer risk in Italy. *Int. J. Cancer* **1995**, *63*, 785–789. [[CrossRef](#)] [[PubMed](#)]
53. Lubin, F.; Ron, E.; Wax, Y.; Modan, B. Coffee and methylxanthines and breast cancer: A case-control study. *J. Natl. Cancer Inst.* **1985**, *74*, 569–573.
54. Matejicic, M.; de Batlle, J.; Ricci, C. Biomarkers of folate and vitamin B12 and breast cancer risk: Report from the EPIC cohort. *Int. J. Cancer* **2017**, *140*, 1246–1259. [[CrossRef](#)]
55. Bassett, J.K.; Baglietto, L.; Hodge, A.M. Dietary intake of B vitamins and methionine and breast cancer risk. *Cancer Causes Control* **2013**, *24*, 1555–1563. [[CrossRef](#)]
56. Lin, J.; Lee, I.M.; Cook, N.R. Plasma folate, vitamin B-6, vitamin B-12, and risk of breast cancer in women. *Am. J. Clin. Nutr.* **2008**, *87*, 734–743. [[CrossRef](#)]
57. Wu, K.; Helzlsouer, K.J.; Comstock, G.W.; Hoffman, S.C.; Nadeau, M.R.; Selhub, J. A prospective study on folate, B12, and pyridoxal 5'-phosphate (B6) and breast cancer. *Cancer Epidemiol. Biomark. Prev.* **1999**, *8*, 209–217.
58. Essén, A.; Santaolalla, A.; Garmo, H. Baseline serum folate, vitamin B12 and the risk of prostate and breast cancer using data from the Swedish AMORIS cohort. *Cancer Causes Control* **2019**, *30*, 603–615.
59. Zhang, S.M.; Willett, W.C.; Selhub, J. Plasma folate, vitamin B6, vitamin B12, homocysteine, and risk of breast cancer. *J. Natl. Cancer Inst.* **2003**, *95*, 373–380. [[CrossRef](#)] [[PubMed](#)]
60. Xu, X.; Chen, J. One-carbon metabolism and breast cancer: An epidemiological perspective. *J. Genet. Genom.* **2009**, *36*, 203–214. [[CrossRef](#)]
61. Choi, S.W.; Mason, J.B. Folate status: Effects on pathways of colorectal carcinogenesis. *J. Nutr.* **2002**, *132*, 2413S–2418S. [[CrossRef](#)] [[PubMed](#)]
62. Mason, J.B. Biomarkers of nutrient exposure and status in one-carbon (methyl) metabolism. *J. Nutr.* **2003**, *133*, 941S–947S. [[CrossRef](#)]
63. Davis, C.D.; Uthus, E.O. DNA methylation, cancer susceptibility, and nutrient interactions. *Exp. Biol. Med.* **2004**, *229*, 988–995. [[CrossRef](#)]
64. Widschwendter, M.; Jones, P.A. DNA methylation and breast carcinogenesis. *Oncogene*. **2002**, *21*, 5462–5482. [[CrossRef](#)]
65. Gonzalgo, M.L.; Jones, P.A. Mutagenic and epigenetic effects of DNA methylation. *Mutat. Res.* **1997**, *386*, 107–118. [[CrossRef](#)]



© 2020 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/>).