HHS Public Access

Author manuscript

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2023 August 06.

Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2023 February 06; 32(2): 217–225.

doi:10.1158/1055-9965.EPI-22-0442.

Dietary factors and early-onset colorectal cancer in the United States—an ecologic analysis

Jianjiu Chen[1,*,†], **Isabella L Zhang**[1,†], **Mary Beth Terry**[1,2], **Wan Yang**[1,2,*]

[1]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York, United States

[2]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, New York, United States

Abstract

Background: Incidence of early-onset colorectal cancer (EOCRC; e.g., diagnosed before age 50) in the US has increased substantially since the 1990s but the underlying reasons remain unclear.

Methods: We examined the ecologic associations between dietary factors and EOCRC incidence in adults aged 25-49 during 1977-2016 in the US, using negative binomial regression models, accounting for age, period, and race. The models also incorporated an age-mean centering (AMC) approach to address potential confounding by age. We stratified the analysis by sex and computed incidence rate ratio (IRR) for each study factor. Study factor data (for 18 variables) came from repeated national surveys; EOCRC incidence data came from the Surveillance Epidemiology, and End Results Program.

Results: Results suggest that confounding by age on the association with EOCRC likely existed for certain study factors (e.g., calcium intake), and that AMC can alleviate the confounding. EOCRC incidence was positively associated with smoking (IRR [95% CI]: 1.17 [1.10-1.24] for men; 1.15 [1.09-1.21] for women) and alcohol consumption (IRR [95% CI]: 1.08 [1.04-1.12] for men; 1.08 [1.04-1.11] for women). No strong associations were found for most other study factors (e.g., fiber and calcium).

Conclusions: Alcohol consumption was positively associated with EOCRC and has increased among young adults since the 1980s, which may have contributed to the EOCRC incidence increases since the 1990s. The AMC approach may help alleviate age confounding in similar ecologic analyses.

Impact: Increases in alcohol consumption may have contributed to the recent increases in colorectal cancer incidence among young adults.

Keywords

Early-onset colorectal cancer; dietary factors; alcohol; ecologic analysis; confounding bias

[*]**Corresponding Authors:** Jianjiu Chen; Address: 722 West 168th Street, Room 517, New York, New York, United States 10032; jc5586@cumc.columbia.edu; Wan Yang; Address: 722 West 168th Street, Room 514, New York, New York, United States 10032; Tel: (212) 305-0421; wy2202@cumc.columbia.edu.

[†]These authors contributed equally to this work.

1 Introduction

Recent studies have showed increases in early-onset colorectal cancer (EOCRC, e.g., those diagnosed before age 50) incidence in the United States (US) since roughly the 1990s.[1-3] Studies have also projected a further increase in EOCRC incidence (e.g., >90% higher by 2030 compared to 2010[4]), if this trend continued. Thus, accurate identification of modifiable risk factors of EOCRC is urgently needed to inform effective prevention in younger adults.

While there is a body of risk factor research on CRC primarily based on cases 50 years and older,[5,6] important research gaps on EOCRC exist. Exposure during early life and critical development period are widely believed to be important in EOCRC development,[7,8] yet studies of such exposure are largely absent. In typical cohorts, exposure measurements start in the 40s, the age of cohort recruitment. Moreover, risk factors of EOCRC and older cases may differ. Compared with older CRC cases, EOCRC is associated with more aggressive pathology and late diagnosis.[7,8] As such, current risk-classification tools based on family history and inflammatory bowel disease could wrongly classify many EOCRC cases as average risk, resulting in late diagnosis.[7] It is also challenging to study risk factors of EOCRC using traditional cohort and/or nested case-control designs. Because the absolute EOCRC risk is relatively low, prohibitively large sample sizes would be needed to provide sufficient statistical power. For example, assuming an incidence rate as that among US women aged 25-49 during 2011-2016 (i.e., 12.9/100,000),[9] to observe 500 cases over five years, a cohort of 0.78 million would be needed.

Given the above research gap and challenges, we conducted an ecologic analysis to examine the association of EOCRC incidence with a range of dietary factors, which are of major interest in EOCRC etiology and amenable to public health interventions. We focused on the US population aged 25-49 (i.e., age groups shown to experience substantial EOCRC incidence increases[2,3,10,11]) during 1977-2016. We also proposed a set of regression models to address two common challenges in similar ecologic analyses (i.e., time lag from exposure to disease and confounding by age). The proposed ecologic approach allows efficient and low-cost investigations of various exposures at different life stages and could be used to study other early-onset cancers with similar rapid increases in recent decades.[10]

2 Methods

2.1 Study design

In a previous study, Pfeiffer et al. examined ecologic associations between concurrent exposures and breast cancer incidence in the US across population groups defined by age, period, race, and sex.[12] Here, to further account for the potential latent period from exposure to cancer diagnosis, we propose two strategies: regress the outcome on i) the exposure 10 years ago (equivalent to lagging the outcome) or ii) the cumulative exposure over the 10 years before the outcome.

Another challenge in our study is potential confounding by age, when both the outcome (here, cancer incidence) and exposure can be associated with age and sometimes in opposite directions (see, e.g., fat intake in Figure 1A). For such exposures, including age as a

covariate in the model may not be able to handle the discordant association of age with the outcome versus exposure. Moreover, for exposures with similar positive association with age as for CRC, residual age confounding is also possible. To address this, we propose an age-mean centering (AMC) approach. Briefly, we remove the association between an exposure and age, by subtracting the age-specific mean exposure for each age, and use these age-removed exposure data in the models (see details below). In so doing, we decouple the age association with the exposure and allow the covariate age to account for its association with the outcome alone (Figure 1). This approach is similar to a strategy in behavioral sciences that disaggregates between-person and within-person effect.[13] Here, we tested five regression models combining the above strategies.

2.2 Study Factors

2.2.1 Data Source—We obtained study factor data from the National Health and Nutrition Examination Surveys (NHANES)[14], the National Health Interview Surveys (NHIS)[15], and the Behavioral Risk Factor Surveillance System (BRFSS)[16]. These three programs conduct repeated national cross-sectional studies in the US over several decades (see Table 1 for survey designs and included survey cycles).[17-19] We included the following dietary factors: smoking, the intake of alcohol, tea, coffee, caffeine, whole fruit, fruit juice, total fruit (whole fruit and fruit juice combined), cholesterol, protein, fiber, calcium, magnesium, fat, saturated fat, total energy, and carbohydrate, and serum folate (see Table S1 for the availability of the study factors in the surveys and sample sizes; see Table S2 for the measurements). Further details on compiling study factor data and handling of periods with no data are described in the Supplementary Methods and Figures S1-S2.

2.2.2 Computing study factor levels—We harmonized study factor data from the different surveys and computed the weighted prevalence for each study factor for each population group defined by age, period, race, and sex.[20,21] The study population (whites and blacks aged 25-49 during 1977-2016) was divided into 160 subgroups: Five 5-year age groups (25-29, 30-34, …, 45-49) × eight 5-year periods (1977-1981, 1982-1986, …, 2012-2016) × two race groups (whites and blacks) × two sexes (men and women). In addition, we computed weighted prevalence for the population groups aged 20-24 and during 1972-1976 for use in the lagged or cumulative models. See Table S3 for specific age and period groups used in each model. Due to small sample sizes, we did not include races other than whites and blacks; we also did not stratify by ethnicity (Hispanic/non-Hispanic), as such information was unavailable from the surveys (e.g., NHANES I and II) or cancer surveillance programs (see below) for earlier periods.

For the no-lag and lagged models (see below), all exposures were categorized by quintiles, as done in Pfeiffer et al.[12] The quintiles were determined based on all population groups (i.e., 80 subgroups for men/women when data were complete). For the other three models (AMC no-lag, AMC lagged, and AMC cumulative; see below), exposures were analyzed as continuous variables, because the AMC-processed exposures no longer spanned a wide range of quintile categories for different age groups and could lead to unstable model estimates using quintiles.

2.3 EOCRC incidence

We obtained EOCRC incidence data from the Surveillance Epidemiology, and End Results (SEER) Program using SEER*STAT.[9,22,23] To match with the exposures, the EOCRC incidence data were aggregated to the same 160 groups specified above. As the coverage of SEER expanded over time, we used SEER data in two ways. In the main analysis, we used SEER 9, which included nine registries, covered 9.4% of the US population, and provided EOCRC incidence throughout our study period (1977-2016). As a sensitivity analysis, we combined SEER 9 with SEER 13 (13 registries; 13.5% coverage; 1992-2016) and SEER 18 (18 registries; 27.8% coverage; 2000-2016). The SEER program, albeit covering a subset of the US population, is representative of the general US population;[24] in addition, SEER started in 1973, earlier than many other national cancer surveillance programs (vs. e.g., the National Program of Cancer Registries starting in 1992).

2.4 Statistical Analysis

Using the population groups defined above, we applied five negative binomial regression models to examine the association between EOCRC incidence and each study factor for men and women, separately.

2.4.1 No-lag model—The no-lag model is similar to the model of Pfeiffer et al.[12] but differs in two ways: i) we included race as a covariate in order to include subgroups from both races (whites and blacks) in the same model to increase sample sizes; ii) we considered age-period interaction. The no-lag model equation is

$$log\left(\lambda_{a,p,r}\right) = \mu + \beta_a + \gamma_p + \delta_{a,p} + \pi_r + (\theta_0 Z_{a,p,r,0} + \theta_1 Z_{a,p,r,1} + \cdots + \theta_4 Z_{a,p,r,4}) \quad (1)$$

where $\lambda_{a,p,r}$ is the expected EOCRC incidence rate for age a, period p, and race r. β_a, γ_p, π_r, and $\delta_{a,p}$ represent the effects of age, period, race, and age-period interaction, respectively; μ is the intercept. $\delta_{a,p}$ was included only when the term was significant (P<0.05) in a model including age, period, race, and interaction terms for all age-period combinations. $Z_{a,p,r,q}$ is an indicator variable that is 1 for age a, period p, race r, and exposure quintile q, and 0 otherwise; a (a = 0,1, … , A) represents the five 5-year age groups: 25-29, …, 45-49; p (p = 0,1, … , P) represents the eight 5-year periods: 1977-1981, …, 2012-2016; r (r = 0,1) represents white and black; and q (q = 0,1,2,3,4) represents the quintile of exposure.

2.4.2 Lagged model—The lagged model used the same structure as Eq (1), except that $Z_{a,p,r,q}$ was replaced by $Z_{a-2,p-2,r,q}$. That is, the exposure occurred 10 years before EOCRC diagnosis (i.e., two 5-year periods ago, hence $p-2$ in the subscript) when the EOCRC cases were 10 years younger (i.e., two 5-year age intervals ago, hence $a-2$). The 10-year lag was chosen, given the likely induction time[5] and data availability (note the youngest age group, i.e., 25-29, can no longer be included due to a lack of earlier measurements; details in Table S3). In addition, we also tested models with a 5-year or 15-year lag to explore pattern across different lags.

2.4.3 AMC no-lag model—We used AMC to address potential confounding by age. For each study factor, we calculated the age-mean centered exposure per

$$R_{a,\,p,\,r} = Z_{a,\,p,\,r} - \overline{Z_{a,\,r}} \quad (2)$$

where $Z_{a,p,r}$ represents the exposure (on continuous scales) for age a, period p, and race r; $\overline{Z_{a,\,r}}$ is the mean of $Z_{a,p,r}$ for age a and race r across all study periods. By subtracting the age-, race-specific mean, the residuals, $R_{a,p,r}$, would still retain the time trend, which is of interest here, but remove the association with age (Figure 1). The AMC no-lag model equation is

$$log\left(\lambda_{a,\,p,\,r}\right) = \mu + \beta_a + \gamma_p + \delta_{a,\,p} + \pi_r + \theta R_{a,\,p,\,r} \quad (3)$$

using the same notations as Eqs (1)-(2). Of note, for all three AMC models (the AMC no-lag model here and the others below) where continuous exposures were used, we standardized the age-removed exposure (mean=0; SD=1) before regression, which allows comparison of the estimates across different study factors and models.

2.4.4 AMC lagged model—The AMC lagged model extends the AMC no-lag model to include the time-lag from exposure to cancer diagnosis. The AMC lagged model equation is the same as Eq (3) except that $R_{a,p,r}$ is replaced by $R_{a-2,p-2,r}$.

2.4.5 AMC cumulative model—The AMC cumulative model uses exposures summed over the 10 years before cancer diagnosis. The model equation is the same as Eq (3) except that $R_{a,p,r}$ is replaced by $R_{a-1,p-1,r} + R_{a-2,p-2,r}$.

2.4.6 Examine the association between age and each study factor—For men and women, separately, we regressed each study factor upon age using 12 groups defined by age and race: six 5-year age groups (20-24, 25-29, …45-49) × two race groups (whites and blacks).

2.4.7 Assess the association between each study factor and EOCRC—All models estimated the incidence rate ratio (IRR) of EOCRC in relation to each study factor, including the mean, 95% confidence interval, and P-value (see Tables 2 and S4). In addition, we used the Bayesian information criterion (BIC) to assess the strength of estimated associations.[25] Specifically, for each study factor and model (one of the five described above), we also tested a corresponding null model with all covariates but the study factor. We calculated the BIC for both models and the diffidence $\Delta BIC = BIC_0 - BIC_f$ (BIC_f for the full model including the study factor and BIC_0 for the null model). $\Delta BIC > 0$ indicates the EOCRC data are better explained when the study factor is included, thus supporting the association between the study factor and EOCRC. The evidence was deemed weak, positive, strong, and very strong for ΔBICs in the ranges of 0-2, 2-6, 6-10, and >10, respectively.[25] $\Delta BIC < 0$ implies an absence of such evidence. All data processing and analyses were conducted using R (https://www.r-project.org).

2.4.8 Method validation—To test the models, we performed two sets of model validation. First, we tested the models on model-generated synthetic data, for which the underlying associations are known and thus can be compared to model estimates. Second,

we applied the models to older age groups (i.e., 35-59-year-olds) and a subset of well-studied exposures (smoking, alcohol consumption, and calcium intake).[6,26] See details in the Supplementary Methods, Tables S5-S7, and Figures S3-S5.

2.5 Data availability statement

The study factor data are publicly available at the websites of NHANES[14], NHIS[15], and BRFSS[16]. The EOCRC incidence data are available at the SEER website[9,22,23].

3 Results

3.1 Method validation

As detailed in the Supplementary Methods, synthetic testing showed that both the lagged and AMC-lagged models were able to accurately identify the true direction of association in most tests (overall accuracy: 79% and 80% by the lagged and AMC-lagged models, respectively; Figures S4-S5). When the association between EOCRC and exposure was close to the null (i.e., IRRs close to 1), the AMC-lagged model was more accurate than the lagged model (71% vs. 65% accuracy; Figure S5), suggesting the AMC approach may alleviate potential biases to more accurately estimate the true association. Furthermore, model results for those aged 35-59 were generally consistent with findings in the literature (i.e., positive associations of CRC with smoking and alcohol consumption and a negative association with calcium intake, primarily based on cases 50 years and older[6,26]); see the red cells (representing positive association) for smoking and alcohol and blue cells (negative association) for calcium in Figure S6 and Table S7 for specific estimates.

3.2 Effect of AMC on estimated associations

We designed AMC to address potential age confounding between study factors and EOCRC. In multiple instances, changes of estimated associations after AMC were consistent with the expected. For instance, as calcium intake decreased with age (Figure S7) while CRC increased with age, age confounding could bias the estimated association between calcium intake and EOCRC towards the negative (Table S5). Indeed, without removing the negative association between calcium intake and age, the no-lag and lagged models estimated negative associations (see blue cells in Figure 2) with larger ΔBICs (ΔBIC>6 except for men using the lagged model; Table 2), indicating stronger evidence for this association. In comparison, the AMC models, designed to remove the age association with calcium intake, generally estimated negative associations with lower ΔBICs, indicating weaker evidence for this association.

For tea, coffee, and caffeine, intake generally increased with age (Figure S7), which could nudge the estimated association with EOCRC towards the positive (Table S5). Indeed, without removing the age association with these exposures, in multiple instances, the no-lag and lagged models estimated positive associations for these exposures (see red cells in Figure 2). In contrast, with AMC, the models in general estimated negative or no association (see light blue or white cells in Figure 2).

3.3 Association between study factors and EOCRC

For smoking, the models found a positive association with EOCRC for both men and women (Figure 2). For 25-49-year-old men, the no-lag model estimated that IRRs were 1.11 (95% CI: 0.95-1.29) and 1.26 (95% CI: 1.04-1.53) for the top two quintiles (Table 2). When smoking prevalence 10 years before EOCRC diagnosis was used, the lagged model estimated that IRRs increased from 1.12 (95% CI: 1.02-1.24) for the second quintile to 1.33 (95% CI: 1.14-1.55) for the fifth. Consistently, estimated IRRs were 1.14 (95% CI: 1.05-1.23) per the AMC no-lag, 1.17 (95% CI: 1.10-1.24) per the AMC lagged, and 1.20 (1.13-1.29) per the AMC cumulative models. For the three AMC models, comparison with the corresponding null models showed strong to very strong support for this association (ΔBIC ranged from 6.1 to 24.9; Table 2).

For alcohol consumption, the models also generally found a positive association with EOCRC for both young men and women (Figure 2). For 25-49-year-old men, the lagged model estimated the IRRs increased from 1.10 (95% CI: 0.97-1.24) for the second quintile to 1.28 (95% CI: 1.13-1.46) for the fifth; for the AMC lagged and AMC cumulative models, estimated IRRs were 1.08 (95% CI: 1.04-1.12) and 1.06 (95% CI: 1.03-1.09), respectively (Table 2). The three models incorporating the time-lag also outperformed their corresponding null models (ΔBICs ranged from 8.1 to 14.5; Table 2), further supporting the association. Models without the time-lag generally found no association for alcohol consumption (except the AMC no-lag model for men).

For the intake of whole fruit, fruit juice, and total fruit, the estimated associations with EOCRC tend to be negative, but the overall evidence was not strong (Figure 2). For the intake of cholesterol, protein, fiber, and magnesium, the estimated associations with EOCRC were either nonsignificant or inconsistent across different models for meaningful interpretation (Figure 2).

The estimated associations between a few study factors and EOCRC were unexpected: negative associations for fat, total energy, and carbohydrate intake, and a positive association for serum folate (Figure 2).

Model results using EOCRC data combining SEER 9, 13, and 18 were similar to those above using SEER 9 data alone (Figure S8 and Table S8). Results from models using different lags are also similar to the main analyses using a 10-year lag; we did not find any clear pattern (Figure S9) except for alcohol, for which the IRRs were the largest with a 10-year lag.

4 Discussion

To explore reasons underlying the recent increases in EOCRC incidence, we have examined the ecologic association between EOCRC and 18 dietary factors. Given the ecological nature of the study, model results represent a first assessment to generate hypotheses regarding potential risk factors to inform more in-depth investigation. Overall, we found that smoking and alcohol consumption starting in young adulthood were positively associated with EOCRC. While these exposures are long-established carcinogens for many cancers

including CRC,[26,27] most studies are based on older populations and mid to late life exposure.[26,28] Given the likely long induction time,[5] our findings suggest that primary prevention strategies for EOCRC, which are urgently needed, should incorporate tobacco and alcohol control measures targeting younger populations. The findings also suggest smoking and alcohol consumption may be important risk factors for identifying young adults for early screening and detection of EOCRC in clinical settings.

The contributions of smoking and alcohol consumption to the recent increases in EOCRC, however, likely differ. As shown in Figure 3, smoking prevalence has been decreasing significantly in recent decades (see details of the break-point trend analysis in the Supplementary Methods), suggesting changes in smoking are likely not the reason behind the recent EOCRC increases. In contrast, alcohol consumption decreased significantly from 1971 to around 1980, consistent with the decrease of EOCRC incidence from 1973 to the early 1990s; alcohol consumption then increased since the 1980s, albeit not statistically significant, followed by the increases in EOECR incidence since the 1990s (Figure 3). These lagged, concordant trends of alcohol consumption and EOCRC incidence resemble the parallel trends in smoking and lung cancer that have strongly supported smoking as a main cause of lung cancer.[26] Consistently, using the approach in Figure 3 of Pfeiffer et al.[12], we showed that, compared to the adjusted EOCRC incidence setting alcohol consumption at the lowest quintile, for both men and women, the observed EOCRC incidence was higher from 1992 onwards and the gap reached the maximum during recent periods (e.g., 2012-2016), when alcohol consumption levels were the highest (see Figure S10 and details in Supplementary Methods). Given these analyses, we hypothesize that increase in alcohol consumption is a key contributor to the recent EOCRC incidence increases. Further investigation is warranted while teasing out the effect of other potential risk factors.

We found some, albeit weak evidence for negative associations of caffeine, whole fruit, fruit juice, and total fruit intake with EOCRC (18/24 of the IRRs in the range of 0.95-0.99 after AMC). The literature on biological effects of these dietary factors also suggests negative associations.[29-31] More in-depth investigation into the potential role of fruit and caffeine using stronger epidemiological designs may thus prove fruitful for EOCRC prevention.

For fiber, calcium, and magnesium intake, we found either no or weak negative association with EOCRC. In contrast, epidemiological studies among older adults suggest these nutrients are protective against CRC.[6,32] For instance, an umbrella review of meta-analyses of cohort studies found convincing evidence for a negative association of CRC with fiber and calcium intake, separately, and some evidence of a negative association with magnesium.[6] Unlike previous studies using cohorts, we used aggregated population-level data, due to the challenges studying EOCRC as noted in the Introduction. This ecologic design may be less powered to identify milder risk factors, especially for younger population (e.g., aged 25-49 here). Moreover, unlike other study factors (e.g., smoking), fiber and magnesium data were unavailable during 1972-1987, further reducing the sample sizes and statistical power. Nonetheless, the direction of our estimates for calcium and magnesium (see the blue cells indicating negative associations in Figure 2) is consistent with previous findings.

Importantly, we note that fiber, calcium, and magnesium intake among blacks were significantly lower than those among whites (P<0.001, paired t-test; Figures S11-S13), and also considerably lower than the recommended levels per current dietary guildlines.[33] Supporting the disparities in intake of these nutrients and potential impact on EOCRC, models including these nutrients partly explained the higher incidence for blacks than whites (e.g., estimated IRRs for black compared to white men: 1.03-1.16 vs. 1.19-1.25 using the lagged model with vs without one of these nutrients; Table S9). Given the higher EOCRC among blacks and multiple health benefits of these nutrients, these findings suggest increasing the intake of these nutrients may help mitigate EOCRC risk among blacks.

Some of our findings were at odds with the literature. In particular, for CRC, past studies found positive associations with high fat diet and total energy intake,[32,34,35] no association with carbohydrate intake,[36] and negative associations with folate intake.[37,38] Model estimates here were inconsistent with these previous findings, particularly for young men, which highlights limitations in this ecologic analysis. Nonetheless, we note that while EOCRC increased during the later part of our study period (from 1990s onwards), fat intake had been decreasing among young men (Figure S14). Similar time-trends were observed for total energy and carbohydrate (Figures S15-16). These trends suggest that, at the population level, the changes in fat, total energy, and carbohydrate intake are likely not associated with the recent increases in EOCRC. For folate intake, serum folate concentration increased during 1987-2016 likely due to the folic acid fortification program implemented in 1998[39] (Figure S17; see Table S2 for reasons for excluding earlier serum folate data); this coincided with the increases in EOCRC during the time period. The positive association between serum folate and EOCRC may have been an artefact of such concurrent changes. We thus caution the above limitations, even though ecologic studies could be invaluable in examining potential risk factors taking advantage of long-term population data. Further, we advocate for comprehensive result interpretation combining ecologic modeling results, findings from the literature, and careful inspection of underlying data, as demonstrated here.

We note several study limitations, apart from the ecological design. First, while our analysis included cigarette smoking, other forms of tobacco consumption were not included due to a lack of long-term data. For example, e-cigarettes have gained popularity among youth and young adults in the US in the 2010s. The potential impacts of such exposure, particularly during critical development periods, warrant future investigations. Second, due to challenges in converting and harmonizing intake of various vegetable items (e.g., inconsistent classification/inclusion schemes and definitions of serving size[40,41]), we were unable to analyze the association of EOCRC with total vegetable intake. Third, this study focused on testing the proposed methods and dietary factors. Future work will extend to non-dietary factors, including those that have been found to affect CRC risks among older adults (e.g., body weight and physical exercise[5]). Fourth, this study estimated the marginal effect of each study factor, as done in Pfeiffer et al.[12] Future work considering potential interactions among various study factors is under way. Fifth, while our models accounted for and estimated the age and period effect, to incorporate the risk factor data and estimate their associations with EOCRC, the models were not formulated as conventional age-period-cohort models[42] to enable estimation of birth cohort effect.

In sum, we found that alcohol consumption was strongly associated with EOCRC incidence and has increased since the 1980s, which may have contributed to recent EOCRC increases among US adults aged 25-49. We have also proposed an AMC approach, which may be applied in ecologic studies of risk factors and other diseases where large-cohort data are unavailable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Declaration of interest:

W. Yang and M.B. Terry report funding from the Data Science Institute and Irving Institute for Cancer Dynamics Seed Funds Program at Columbia University and the National Cancer Institute (grant number: R01CA257971) for the conduct of this study. All other authors declare no potential conflicts of interest.

References

1. Siegel RL, Fedewa SA, Anderson WF, Miller KD, Ma J, Rosenberg PS, et al. Colorectal Cancer Incidence Patterns in the United States, 1974-2013. J Natl Cancer Inst. 2017;109(8).
2. Kehm RD, Lima SM, Swett K, Mueller L, Yang W, Gonsalves L, et al. Age-specific Trends in Colorectal Cancer Incidence for Women and Men, 1935-2017. Gastroenterology. 2021;161(3):1060–1062 e1063. [PubMed: 34058214]
3. Yang W, Kehm RD, Terry MB. Survival model methods for analyses of cancer incidence trends in young adults. Stat Med. 2020;39(7):1011–1024. [PubMed: 32022306]
4. Bailey CE, Hu CY, You YN, Bednarski BK, Rodriguez-Bigas MA, Skibber JM, et al. Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975-2010. JAMA Surg. 2015;150(1):17–22. [PubMed: 25372703]
5. Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet. 2014;383(9927):1490–1502. [PubMed: 24225001]
6. Veettil SK, Wong TY, Loo YS, Playdon MC, Lai NM, Giovannucci EL, et al. Role of Diet in Colorectal Cancer Incidence: Umbrella Review of Meta-analyses of Prospective Observational Studies. JAMA Netw Open. 2021;4(2):e2037341. [PubMed: 33591366]
7. Stoffel EM, Murphy CC. Epidemiology and Mechanisms of the Increasing Incidence of Colon and Rectal Cancers in Young Adults. Gastroenterology. 2020;158(2):341–353. [PubMed: 31394082]
8. Hofseth LJ, Hebert JR, Chanda A, Chen H, Love BL, Pena MM, et al. Early-onset colorectal cancer: initial clues and current views. Nat Rev Gastroenterol Hepatol. 2020;17(6):352–364. [PubMed: 32086499]
9. seer.cancer.gov [Internet]. Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission [cited 2022 Sep 14]. Available from: https://seer.cancer.gov/data-software/.
10. Yang W, Terry MB. Do Temporal Trends in Cancer Incidence Reveal Organ System Connections for Cancer Etiology? Epidemiology. 2020;31(4):595–598. [PubMed: 32221269]
11. Kehm RD, Yang W, Tehranifar P, Terry MB. 40 Years of Change in Age- and Stage-Specific Cancer Incidence Rates in US Women and Men. JNCI Cancer Spectr. 2019;3(3):pkz038. [PubMed: 31414075]
12. Pfeiffer RM, Webb-Vargas Y, Wheeler W, Gail MH. Proportion of U.S. Trends in Breast Cancer Incidence Attributable to Long-term Changes in Risk Factor Distributions. Cancer Epidemiol Biomarkers Prev. 2018;27(10):1214–1222. [PubMed: 30068516]

13. Curran PJ, Bauer DJ. The disaggregation of within-person and between-person effects in longitudinal models of change. Annu Rev Psychol. 2011;62:583–619. [PubMed: 19575624]
14. cdc.gov [Internet]. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. In. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention [cited 2022 Sep 14]. Available from: https://wwwn.cdc.gov/nchs/nhanes/default.aspx.
15. nhis.ipums.org [Internet]. Blewett LA, Drew JAR, King ML, Williams KCW. IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset]. In. Minneapolis, MN: IPUMS; 2019 [cite 2022 Sep 14]. Available from: https://nhis.ipums.org/nhis-action/variables/group.
16. cdc.gov [Internet]. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. In. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention [cited 2022 Sep 14]. Available from: https://www.cdc.gov/brfss/.
17. cdc.gov [Internet]. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Examination Protocol. In. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention [cited 2022 Sep 14]. Available from: https://www.cdc.gov/nchs/nhanes/index.htm.
18. nhis.ipums.org [Internet]. NHIS sample design. IPUMS [cited 2022 Sep 14]. Available from: https://nhis.ipums.org/nhis/userNotes_sampledesign.shtml.
19. cdc.gov [Internet]. Behavioral Risk Factor Surveillance System. Survey Data & Documentation. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention [cited 2022 Sep 14]. Available from: https://www.cdc.gov/brfss/data_documentation/index.htm.
20. cdc.gov [Internet]. National Health and Nutrition Examination Survey. Tutorial Module 3: Weighting. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention [cited 2022 Sep 14]. Available from: https://wwwn.cdc.gov/nchs/nhanes/tutorials/Module3.aspx.
21. Korn EL, Graubard BI. Chapter 8: Analyses using multiple surveys. In: Analysis of health surveys. John Wiley & Sons, Inc; 1999. p. 278–84.
22. seer.cancer.gov [Internet]. Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER Research Data, 13 Registries, Nov 2020 Sub (1992-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission [cited 2022 Sep 14]. Available from: https://seer.cancer.gov/data-software/.
23. seer.cancer.gov [Internet]. Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER Research Data, 18 Registries, Nov 2020 Sub (2000-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission [cited 2022 Sep 14]. Available from: https://seer.cancer.gov/data-software/.
24. Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology: Toward Strengthening the Critical Relationship. Am J Surg Pathol. 2016;40(12):e94–e102. [PubMed: 27740970]
25. Raftery AE. Bayesian Model Selection in Social Research. Sociological Methodology. 1995;25(1995):111–163.
26. The Health Consequences of Smoking-50 Years of Progress: A Report of the Surgeon General. Atlanta (GA): Centers for Disease Control and Prevention (US); 2014. p. 197–203.
27. Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, Bouvard V, et al. Carcinogenicity of alcoholic beverages. Lancet Oncol. 2007;8(4):292–293. [PubMed: 17431955]
28. Cho E, Smith-Warner SA, Ritz J, van den Brandt PA, Colditz GA, Folsom AR, et al. Alcohol intake and colorectal cancer: a pooled analysis of 8 cohort studies. Ann Intern Med. 2004;140(8):603–613. [PubMed: 15096331]
29. Cui WQ, Wang ST, Pan D, Chang B, Sang LX. Caffeine and its main targets of colorectal cancer. World J Gastrointest Oncol. 2020;12(2):149–172. [PubMed: 32104547]

30. Steinmetz KA, Potter JD. Vegetables, fruit, and cancer. II. Mechanisms. Cancer Causes Control. 1991;2(6):427–442. [PubMed: 1764568]

31. Aune D, Lau R, Chan DS, Vieira R, Greenwood DC, Kampman E, et al. Nonlinear reduction in risk for colorectal cancer by fruit and vegetable intake based on meta-analysis of prospective studies. Gastroenterology. 2011;141(1):106–118. [PubMed: 21600207]

32. Song M, Garrett WS, Chan AT. Nutrients, foods, and colorectal cancer prevention. Gastroenterology. 2015;148(6):1244–1260 e1216. [PubMed: 25575572]

33. DietaryGuidelines.gov [Internet]. U.S. Department of Agriculture and U.S. Department of Health and Human Services. Dietary Guidelines for Americans, 2020-2025. 9th Edition. December 2020 [cited 2022 Sep 14]. Available from: https://www.dietaryguidelines.gov/sites/default/files/2020-12/Dietary_Guidelines_for_Americans_2020-2025.pdf.

34. Beyaz S, Mana MD, Roper J, Kedrin D, Saadatpour A, Hong SJ, et al. High-fat diet enhances stemness and tumorigenicity of intestinal progenitors. Nature. 2016;531(7592):53–58. [PubMed: 26935695]

35. Sun Z, Liu L, Wang PP, Roebothan B, Zhao J, Dicks E, et al. Association of total energy intake and macronutrient consumption with colorectal cancer risk: results from a large population-based case-control study in Newfoundland and Labrador and Ontario, Canada. Nutr J. 2012;11:18. [PubMed: 22449145]

36. Aune D, Chan DS, Lau R, Vieira R, Greenwood DC, Kampman E, et al. Carbohydrates, glycemic index, glycemic load, and colorectal cancer risk: a systematic review and meta-analysis of cohort studies. Cancer Causes Control. 2012;23(4):521–535. [PubMed: 22418776]

37. Liu Y, Yu Q, Zhu Z, Zhang J, Chen M, Tang P, et al. Vitamin and multiple-vitamin supplement intake and incidence of colorectal cancer: a meta-analysis of cohort studies. Med Oncol. 2015;32(1):434. [PubMed: 25491145]

38. Bollheimer LC, Buettner R, Kullmann A, Kullmann F. Folate and its preventive potential in colorectal carcinogenesis. How strong is the biological and epidemiological evidence? Crit Rev Oncol Hematol. 2005;55(1):13–36. [PubMed: 15927841]

39. Yetley EA, Johnson CL. Folate and vitamin B-12 biomarkers in NHANES: history of their measurement and use. Am J Clin Nutr. 2011;94(1):322S–331S. [PubMed: 21593508]

40. Patterson BH, Block G, Rosenberger WF, Pee D, Kahle LL. Fruit and vegetables in the American diet: data from the NHANES II survey. Am J Public Health. 1990;80(12):1443–1449. [PubMed: 2240327]

41. Branum AM, Rossen LM. The contribution of mixed dishes to vegetable intake among US children and adolescents. Public Health Nutr. 2014;17(9):2053–2060. [PubMed: 23962488]

42. Holford TR. Understanding the effects of age, period, and cohort on incidence and mortality rates. Annu Rev Public Health. 1991;12:425–457. [PubMed: 2049144]

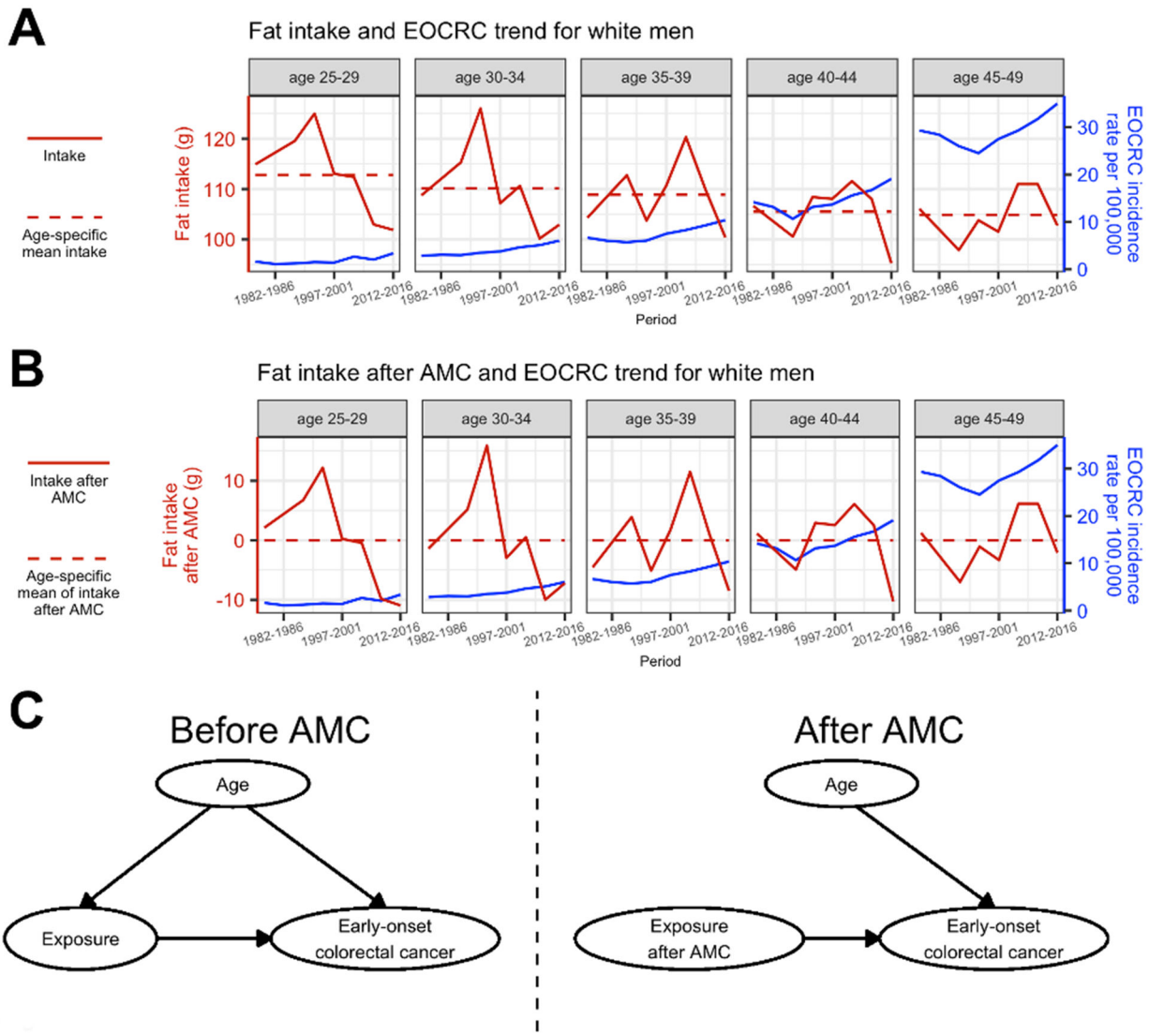

Figure 1. Schematic of the age-mean centering (AMC) approach to alleviate age confounding.

(A) Trends in fat intake and early onset colorectal cancer (EOCRC) among white men aged 25-49 show an example where both the exposure (here, fat intake) and outcome (here, EOCRC) are associated with age and could act in opposite directions, i.e., fat intake tends to decrease with age whereas EOCRC incidence tends to increase with age. (B) After removing the age-specific mean (i.e., subtracting the average of all fat intake values for that age and race group across all periods, shown by the horizontal dashed lines in A), the age-mean centered fat intake values are now on similar scales for all age groups (0 means for all age groups as shown by the horizontal dashed lines; estimated association with age: 0.00 (95% CI: −0.17, 0.17) using a regression model), whereas the trends over time remain the same as the raw data shown in A. (C) Diagrams of causal relationships before and after applying AMC to the exposure data. Without AMC, age is associated with both the exposure and outcome and could bias the estimate of exposure-outcome association (left panel). After AMC, age is no longer associated with the exposure (right panel); in addition, because changes in exposure over time (i.e., calendar years) needed to examine the changes in

disease outcome over time are still retained as shown in B, the age-mean centered exposure can be used to examine the association between the exposure and disease outcome.

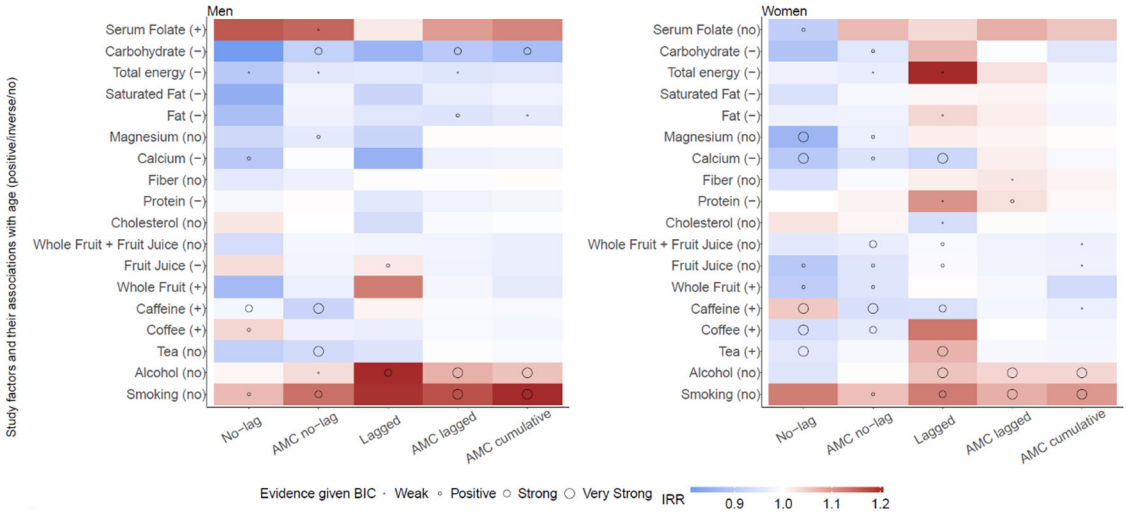

Figure 2.

Estimated incidence rate ratios (IRRs) of early-onset colorectal cancer (EOCRC) incidence in relation to study factors, for 25-49-year-old men and women. EOCRC incidence data were obtained from SEER 9. All IRR estimates were adjusted for age, period, and race. For the no-lag and lagged models, the mean of four IRRs (for the quintiles) is shown here. Of note, the estimated IRRs before and after AMC are not comparable because the exposure was on different scales (quintiles vs continuous). The evidence for an association was deemed weak, positive, strong, and very strong for ΔBICs in the ranges of 0-2, 2-6, 6-10, and >10, respectively. Numerical values of the IRRs and ΔBICs are shown in Tables 2 and S4.

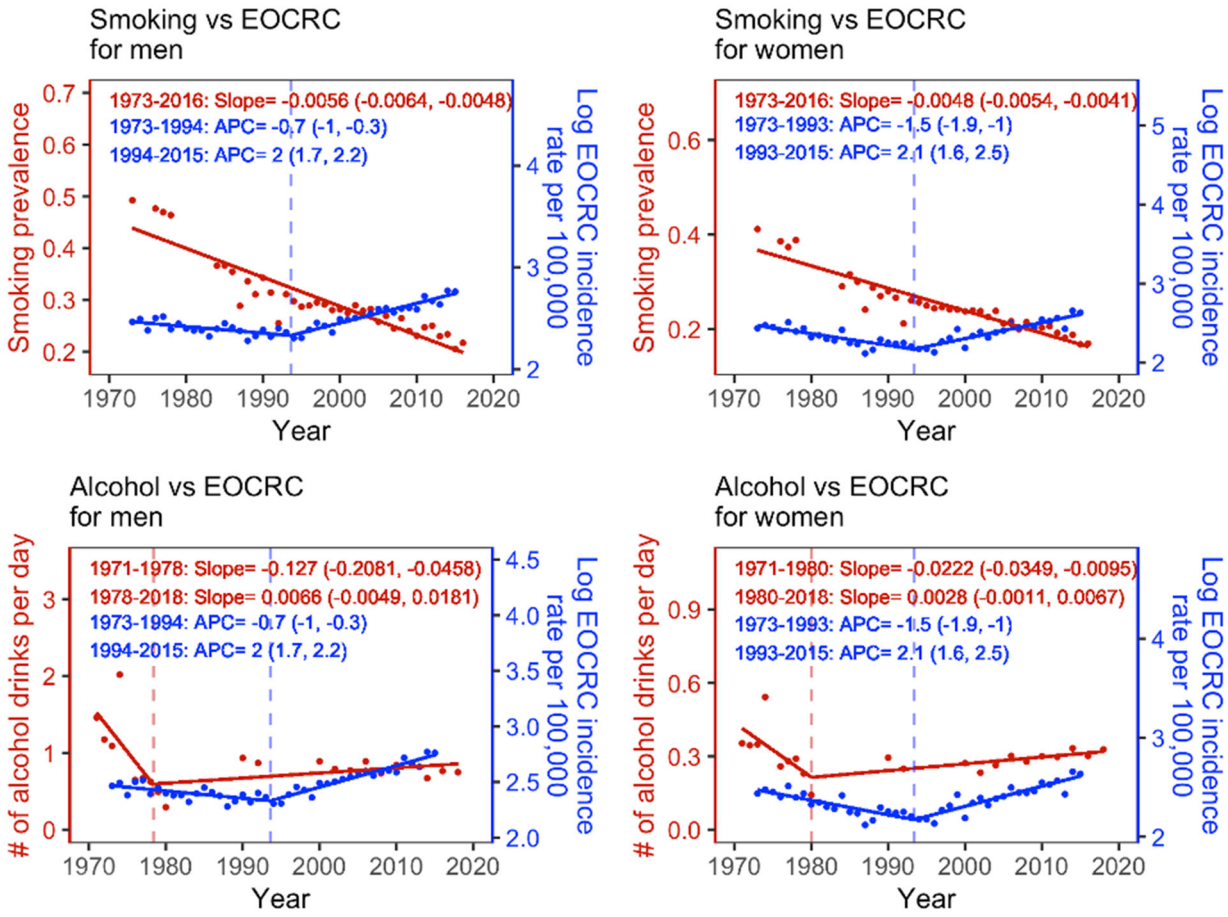

Figure 3.

Trends of smoking, alcohol consumption, and early-onset colorectal cancer (EOCRC) incidence for 25-49-year-old men and women. All estimates of smoking prevalence, numbers of alcohol drinks per day, and EOCRC incidence rates were age-standardized using the age structure of US population in 2000. Red dots show available measurements for smoking (top row) and alcohol consumption (bottom row) from NHANES; blue dots show annual EOCRC incidence (log transformed) from SEER 9. The estimated annual percentage changes [APCs; means (95% confidence intervals)] for EOCRC incidence are shown in blue; the corresponding estimates for the slopes of smoking prevalence and alcohol consumption are shown in red. Note that alcohol consumption data for Year 1981-1989 were not available, which may affect the accuracy of estimated trend (e.g., timing of break-point). See details of the trend analysis in the Supplementary Methods.

Table 1.

Overview of the sources of study factor data.

	NHANES	NHIS	BRFSS
Target population	Civilian, noninstitutionalized US population	Civilian, noninstitutionalized US population	US adult population
Sampling approach	Multistage, probability sampling	Multistage, probability sampling	Sampling based on landline (and also cellular telephone numbers since 2011)
Data collection	Household interviews and physical examinations	Household interviews	Telephone interviews
Included surveys	NHANES I (1971-1974), NHANES II (1976 to 1980), NHANES III (1988 to 1994), continuous NHANES (1999-2016)	NHIS (1976, 1977, 1985, 1987, 1988, 1990-1995, and 1997-2016)	BRFSS (1984-2016)

NHANES=National Health and Nutrition Examination Surveys; NHIS=National Health Interview Surveys; BRFSS=Behavioral Risk Factor Surveillance System;

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2023 August 06.

Table 2.

Estimated associations of early-onset colorectal cancer (EOCRC) incidence with smoking, alcohol consumption, and calcium intake (using SEER 9 data; age groups: 25-49). Five models (2nd column) were used to estimate incidence rate ratio (IRR) of EOCRC for each unit/quintile change in each study factor (1st column). Of the five models, the no-lag and lagged models used study factor levels defined by quintile (3rd column) as input and estimated the IRR for each quintile, relative to the first quintile (i.e., the reference); the age-mean centering (AMC) models used continuous values as input and estimated the IRR for each unit change of the study factor. Of note, the estimated IRRs before and after AMC are not comparable because the exposure was on different scales. p-values, and ΔBIC and assessment of the strength of evidence (see details in main text) are also shown for each model.

			Men					Women				
Study Factor	**Method**	**Quintile**	**IRR (95% CI)**	**p-value (quintiles)**	**p-value (cont.)[a]**	**ΔBIC**	**Evidence**	**IRR (95% CI)**	**p-value (quintiles)**	**p-value (cont.)[a]**	**ΔBIC**	**Evidence**
Smoking	No-lag	1	1		<0.001	2	Positive	1		<0.001	−5.4	No Improvement
		2	0.92 (0.82-1.04)	0.21				1.13 (0.99-1.28)	0.07			
		3	0.98 (0.87-1.10)	0.71				1.20 (1.06-1.37) **	0.005			
		4	1.11 (0.95-1.29)	0.19				1.28 (1.11-1.48) ***	<0.001			
		5	1.26 (1.04-1.53) *	0.02				1.37 (1.13-1.66) **	0.002			
	Lagged	1	1		<0.001	−0.9	No Improvement	1		<0.001	7.7	Strong
		2	1.12 (1.02-1.24) *	0.02				1.15 (1.07-1.24) ***	<0.001			
		3	1.13 (1.01-1.27) *	0.03				1.19 (1.07-1.32) **	0.002			
		4	1.21 (1.06-1.37) **	0.004				1.26 (1.10-1.43) ***	<0.001			
		5	1.33 (1.14-1.55) ***	<0.001				1.39 (1.18-1.62) ***	<0.001			
	AMC no-lag		1.14 (1.05-1.23) ***		<0.001	6.1	Strong	1.12 (1.03-1.21) **		0.005	3.3	Positive
	AMC lagged		1.17 (1.10-1.24) ***		<0.001	21.4	Very Strong	1.15 (1.09-1.21) ***		<0.001	21	Very Strong
	AMC cumulative		1.20 (1.13-1.29) ***		<0.001	24.9	Very Strong	1.19 (1.12-1.27) ***		<0.001	27.8	Very Strong
Alcohol	No-lag	1	1		0.5	−15.5	No Improvement	1		0.81	−4.9	No Improvement
		2	0.98 (0.88-1.09)	0.72				0.87 (0.78-0.96) **	0.009			
		3	1.02 (0.91-1.15)	0.69				0.87 (0.77-0.97) *	0.01			
		4	0.99 (0.89-1.11)	0.92				0.98 (0.89-1.09)	0.76			
		5	1.03 (0.92-1.16)	0.57				0.92 (0.83-1.03)	0.14			
Alcohol	Lagged	1	1		<0.001	8.1	Strong	1		<0.001	16.3	Very Strong

Study Factor	Method	Quintile	Men					Women				
			IRR (95% CI)	p-value (quintiles)	p-value (cont.)[a]	ΔBIC	Evidence	IRR (95% CI)	p-value (quintiles)	p-value (cont.)[a]	ΔBIC	Evidence
		2	1.10 (0.97-1.24)	0.13				1.04 (0.94-1.15)	0.45			
		3	1.17 (1.05-1.31) **	0.005				1.02 (0.91-1.13)	0.77			
		4	1.27 (1.13-1.43) ***	<0.001				1.15 (1.05-1.26) **	0.003			
		5	1.28 (1.13-1.46) ***	<0.001				1.23 (1.11-1.37) ***	<0.001			
	AMC no-lag		1.03 (1-1.06) *		0.03	0.3	Weak	1 (0.97-1.04)		0.81	−4.2	No Improvement
	AMC lagged		1.08 (1.04-1.12) ***		<0.001	11.3	Very Strong	1.08 (1.04-1.11) ***		<0.001	15.6	Very Strong
	AMC cumulative		1.06 (1.03-1.09) ***		<0.001	14.5	Very Strong	1.07 (1.04-1.11) ***		<0.001	17.7	Very Strong
Calcium	No-lag	1	1		0.44	6	Positive	1		0.008	16.2	Very Strong
		2	0.91 (0.81-1.01)	0.08				0.87 (0.78-0.98) *	0.02			
		3	0.85 (0.74-0.97) *	0.01				0.81 (0.71-0.92) **	0.001			
		4	0.88 (0.76-1.03)	0.1				0.74 (0.62-0.87) ***	<0.001			
		5	0.96 (0.77-1.19)	0.69				0.80 (0.63-1.01)	0.06			
	Lagged	1	1		0.29	−3.7	No Improvement	1		0.04	13.5	Very Strong
		2	0.86 (0.77-0.96) **	0.006				0.92 (0.83-1.02)	0.12			
		3	0.84 (0.74-0.97) *	0.02				0.80 (0.70-0.91) ***	<0.001			
		4	0.85 (0.73-0.99) *	0.04				0.81 (0.69-0.95) **	0.01			
		5	0.85 (0.70-1.04)	0.12				0.90 (0.73-1.10)	0.31			
	AMC no-lag		1 (0.96-1.04)		0.86	−4.2	No Improvement	0.90 (0.84-0.97) **		0.008	2.2	Positive
	AMC lagged		0.98 (0.95-1.01)		0.12	−1.4	No Improvement	1.03 (0.95-1.11)		0.46	−3.3	No Improvement
	AMC cumulative		0.98 (0.94-1.02)		0.33	−3.1	No Improvement	0.98 (0.90-1.06)		0.6	−3.8	No Improvement

*P<0.05

**P<0.01

***P<0.001; IRR=Incidence rate ratio; BIC=Bayesian information criterion; AMC=Age-mean centering.

IRRs were adjusted for age, period, and race.

[a]For the no-lag and lagged models, we generated p-value (cont.) by treating the median of the quintiles as a continuous variable in regression.

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2023 August 06.