cancers

MDPI

Article

Geographic Variation and Risk Factor Association of Early Versus Late Onset Colorectal Cancer

Weichuan Dong [1,*], Uriel Kim [1,2,3], Johnie Rose [1,3,4], Richard S. Hoehn [5], Matthew Kucmanic [6], Kirsten Eom [7], Shu Li [8], Nathan A. Berger [4,9] and Siran M. Koroukian [1,3,4]

1 Population Cancer Analytics Shared Resource and Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA
2 Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA
3 Center for Community Health Integration, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA
4 Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA
5 Department of Surgery, Division of Surgical Oncology, University Hospitals Cleveland Medical Center, Cleveland, OH 44106, USA
6 Department of Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA 52242, USA
7 MetroHealth Cancer Center, Cleveland, OH 44109, USA
8 School of Digital Sciences, Kent State University, Kent, OH 44240, USA
9 Center for Science, Health and Society, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA
* Correspondence: weichuan.dong@case.edu; Tel.: +1-216-368-3445

Simple Summary: While the incidence of late-onset colorectal cancer (LOCRC) has steadily decreased, the incidence of early-onset colorectal cancer (EOCRC) has continued to increase in the US. This study aims to uncover geographic disparities in EOCRC and understand how risk factors between EOCRC and LOCRC differ. The geographic analysis revealed regions with relatively low LOCRC rates and high EOCRC rates, identifying regions with a disproportionate burden of EOCRC. We then evaluated and compared community-level risk factors associated with incidence rates of EOCRC and LOCRC using the random forest machine learning method. The analysis identified a set of risk factors most predictive of EOCRC and LOCRC, such as diabetes prevalence and physical inactivity, but the importance of these risk factors varied between EOCRC and LORC. Collectively, these findings can help facilitate future studies that further uncover actionable interventions to reduce EOCRC and guide where targeted interventions to reduce EOCRC burden should be deployed.

Abstract: The proportion of patients diagnosed with colorectal cancer (CRC) at age < 50 (early-onset CRC, or EOCRC) has steadily increased over the past three decades relative to the proportion of patients diagnosed at age ≥ 50 (late-onset CRC, or LOCRC), despite the reduction in CRC incidence overall. An important gap in the literature is whether EOCRC shares the same community-level risk factors as LOCRC. Thus, we sought to (1) identify disparities in the incidence rates of EOCRC and LOCRC using geospatial analysis and (2) compare the importance of community-level risk factors (racial/ethnic, health status, behavioral, clinical care, physical environmental, and socioeconomic status risk factors) in the prediction of EOCRC and LOCRC incidence rates using a random forest machine learning approach. The incidence data came from the Surveillance, Epidemiology, and End Results program (years 2000–2019). The geospatial analysis revealed large geographic variations in EOCRC and LOCRC incidence rates. For example, some regions had relatively low LOCRC and high EOCRC rates (e.g., Georgia and eastern Texas) while others had relatively high LOCRC and low EOCRC rates (e.g., Iowa and New Jersey). The random forest analysis revealed that the importance of community-level risk factors most predictive of EOCRC versus LOCRC incidence rates differed meaningfully. For example, diabetes prevalence was the most important risk factor in predicting EOCRC incidence rate, but it was a less important risk factor of LOCRC incidence rate; physical inactivity was the most important risk factor in predicting LOCRC incidence rate, but it was the fourth most important predictor for EOCRC incidence rate. Thus, our community-level analysis

check for updates

Citation: Dong, W.; Kim, U.; Rose, J.; Hoehn, R.S.; Kucmanic, M.; Eom, K.; Li, S.; Berger, N.A.; Koroukian, S.M. Geographic Variation and Risk Factor Association of Early Versus Late Onset Colorectal Cancer. *Cancers* **2023**, *15*, 1006. https://doi.org/10.3390/cancers15041006

Academic Editor: Yutaka Midorikawa

Received: 7 January 2023
Revised: 31 January 2023
Accepted: 2 February 2023
Published: 4 February 2023

demonstrates the geographic variation in EOCRC burden and the distinctive set of risk factors most predictive of EOCRC.

Keywords: colorectal cancer; risk factor; early-onset; geographic information system; machine learning; random forest; regionalization

1. Introduction

Colorectal cancer (CRC) incidence in the US has declined steadily since the mid-1990s, owing to substantial drops in new cases among individuals aged 50 years and older (late-onset CRC, or LOCRC) [1,2]. This trend is likely attributable to improved screening rates in this age group. However, diagnosis of CRC in individuals younger than 50 years old (early-onset CRC, or EOCRC) has nearly doubled over the same period and continues to increase by approximately 2% annually [1]. Thus, uncovering risk factors for EOCRC, especially those that may be unique to early-onset disease versus late-onset disease, can inform efforts that facilitate the prevention and detection of this growing cancer subtype.

Researchers have identified multiple shared risk factors for EOCRC and LOCRC, including sedentary behaviors, Western-style diets, smoking, and diabetes [3–5]. Studies suggest that EOCRC may have distinct clinical and molecular phenotypes compared with LOCRC. This may be attributable to differences in demographic and genetic factors, or potentially modifiable behavioral and environmental factors [6]. Studies also have linked CRC risk with social determinants of health [7], which are often measured by area-level characteristics of individuals' residences as a surrogate. However, the extent to which these factors are associated with EOCRC, and whether they differ from those associated with LOCRC, remains unclear.

Since incidence rates in many sparsely populated counties are often unreliable due to the small number of cases in those areas, it is challenging to conduct population and geographic studies on EOCRC. To our knowledge, no study has conducted a comprehensive area-level analysis comparing demographic, health status, behavioral, clinical care, environmental, and socioeconomic risk factors between EOCRC and LOCRC. This cross-sectional study applied a geographic method (Max-p-regions) to address the data scarcity problem at the county level, followed by comparing the differences in the geographic distributions of the incidence rates of EOCRC and LOCRC. Finally, the study evaluated and compared risk factor importance in the prediction of EOCRC and LOCRC incidence rates.

2. Materials and Methods

2.1. Study Population and Data Sources

The study included individuals diagnosed with CRC before 50 years of age (EOCRC) and at 50 years or older (LOCRC) between 2000 and 2019 from the Surveillance, Epidemiology, and End Results (SEER) database [8]. Data in the SEER program cover 48% of the U.S. population with 97% completeness within the SEER regions [9]. Geographic regions included in this study cover the U.S. states of California, Connecticut, Georgia, Hawaii, Idaho, Illinois, Iowa, Kentucky, Louisiana, Massachusetts, New Jersey, New Mexico, New York, Texas, Utah, and Washington (Seattle–Puget Sound), with a total of 1085 counties. All data were obtained at the county level and were aggregated into a larger geographic unit for analysis as necessary using the max-p-regions method described in further detail below.

2.2. Variables of Interest

The outcomes of interest were the age-adjusted EOCRC incidence rate and age-adjusted LOCRC incidence rate per 100,000 persons. The SEER*STAT software was used to query and extract cancer incidence data and patient characteristics. Predictor variables for the random forest models were obtained from the County Health Rankings and Roadmaps (CHR) [10] and the Health Resources and Services Administration—Area Health Resources

Files (AHRF) [11] to include the relevant domains of population race/ethnicity, health status, health behaviors, clinical care, physical environment, and socioeconomic status (Table 1).

Table 1. County-level explanatory variables used in random forest analyses.

Variable	Description	Year	Original Source
	Ethnicity/Race		
Non-Hispanic Black	Percentage of Non-Hispanic African American population	2011	Census—PE
Non-Hispanic White	Percentage of Non-Hispanic White population		
Hispanic	Percentage of population identifying as Hispanics		
	Health Status		
Diabetes	Percentage of adults (age 20+) with diagnosed diabetes (age-adjusted)	2009	CDC—DIA
Low birthweight	Percentage of live births with low birthweight (<2500 g)	2004–2010	CDC—NCHS
Poor or fair health	Percentage of adults reporting fair or poor health (age-adjusted)	2014	CDC—BRFSS
	Health Behaviors		
Adult obesity	Percentage of the people (age 18+) that report a body mass index $\geq$ 30 (age-adjusted)	2009	CDC—DIA
Physical inactivity	Percentage of adults (age 18+) reporting no leisure-time physical activity (age-adjusted)	2009	
Excessive drinking	Percentage of adults reporting binge or heavy drinking (age-adjusted)	2014	CDC—BRFSS
Insufficient sleep	Percentage of adults who report fewer than 7 h of sleep on average (age-adjusted)	2014	
Smoking	Percentage of adults who are current smokers (age-adjusted)	2014	
Access to recreational facilities	Number of recreational facilities (engaged in fitness and recreational sports) per 100,000 persons	2010	USDA FEA and Census CBP
Food insecurity	Percentage of population who lack adequate access to food	2013	MMP
Teen birth	Number of births per 1000 females ages 15–19	2004–2010	CDC—NCHS
	Clinical Care		
Uninsured adults	Percentage of adults under age 65 without health insurance	2010	Census—SAHIE
Health care costs	Per capita spending of Medicare enrollees	2009	DAHC
Preventable hospital stays	Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees	2010	
Primary care physicians	Primary care physicians in patient care per 100,000 persons	2010	HRSA
Community health centers	Community health centers per 100,000 persons	2013	
	Physical Environment		
Air pollution	Total days when the fine particulate matter (PM2.5) exceeded the 24 h primary standard established by the US EPA.	2008	CDC WONDER
Driving alone to work	Percentage of the workforce that drives alone to work (indicators of the transit system and physical inactivity)	2007–2011	Census—ACS
Rural population	Percentage of people living in rural areas	2010	Census—PE

Table 1. *Cont.*

Socioeconomic Status			
Area deprivation index	A composite measure of 17 census variables designed to describe social deprivation	2008–2012	Census—ACS
Median household income	The income where half of households in a county earn more and half of households earn less	2011	Census—SAIPE
Poverty	Percentage of people in poverty	2010	
Receipt of SNAP benefits	Percentage of people who were food stamp recipients	2010	Census—SNAP
Some college	Percentage of people (age 25–44) with some post-secondary education	2007–2011	Census—ACS
Female-headed households	Percentage of female-headed households	2010	Census (Decennial)
Unemployment	Percentage of population (age 16+) unemployed but seeking work	2011	BLS
Violent crime	Number of reported violent crime offenses per 100,000 persons	2008–2010	FBI—UCR

Abbreviations: ACS: American Community Survey; BLS: Bureau of Labor Statistics; BRFSS: Behavioral Risk Factor Surveillance System; CBP: County Business Patterns; CDC: Centers for Disease Control and Prevention; DAHC: Dartmouth Atlas of Health Care; DIA: Diabetes Interactive Atlas; FEA: Food Environment Atlas; HRSA: Health Resources and Services Administration; NCHS: National Center for Health Statistics; PE: Population Estimates; SAHIE: Small Area Health Insurance Estimates; SAIPE: Small Area Income and Poverty Estimates; SEER: Surveillance, Epidemiology, and End Results; SNAP: Supplemental Nutrition Assistance Program; UCR: Uniform Crime Reporting; USDA: U.S. Department of Agriculture. Note: Area Deprivation Index was obtained using the 'sociome' R package; Community Health Centers, Poverty, Receipt of SNAP benefits, and Female-headed households were obtained from HRSA—Area Health Resources Files; All other variables were obtained from County Health Rankings and Roadmaps.

2.3. Analytic Approach

2.3.1. Composite Counties

To protect patient confidentiality and mitigate unstable cancer incidence rates, counties with fewer than 16 EOCRC or LOCRC cases are typically suppressed. However, since over one-third of counties had fewer than 16 EOCRC or LOCRC cases over the study period, excluding these counties would likely introduce bias by excluding patients from less populated areas. To address this issue, we used a regionalization method called Max-p-Regions [12] to aggregate adjacent, socio-demographically similar counties into the smallest geographic area based on the threshold constraint of at least 16 EOCRC or LOCRC cases and called the newly created area composite county (CC). The county-level area deprivation index (ADI), an index of social deprivation calculated from 17 census variables [13], was used to determine which counties should be grouped together considering their socio-demographical similarity. Next, the age-adjusted incidence rates of EOCRC and LOCRC were calculated at the CC-level. To account for differences in population size of CCs, county-level predictor variables were weighted at the CC-level based on the relative population of the grouped counties.

2.3.2. Random Forest and Variable Importance

Random forest analysis was used to evaluate variable importance (VI) of risk factors in predicting EOCRC incidence rate and LOCRC incidence rate, respectively. Random forest is a tree-based, nonlinear, nonparametric machine learning method that creates and aggregates an ensemble of trees for prediction using random variable selection and bootstrap sampling [14]. In each regression tree of the random forest, every predictor is "competing" with all other predictors during the selection of a split node. The results of the random forest analysis are highly stable and replicable because they are based on the average output of all its regression trees [14]. Compared with traditional linear regression models, random forest can evaluate a large number of predictors without concerns about correlation among them, as demonstrated in a previous study [15].

VI is a measure of node impurity, which reflects the extent to which stratification by a given variable minimizes the variance of responses within resultant subgroups in the regression trees of the random forest. Generally, the frequency of a predictor serving as the splits of trees in the random forest determines its VI. The VI of each predictor was then ordered from highest to lowest, and was also classified into high, median high, median low, and low using the Jenks natural breaks classification method, a data classification method that reduces the variance within classes and maximizes the variance between classes [16].

We created 20,000 trees with all variables from Table 1 included as predictors in the random forest analyses. The number of variables randomly sampled as candidates at each tree split was set to 5. R statistical software (version 4.2.1) and the package "randomForest" were used for the random forest analyses.

Because the random forest is a nonlinear tree-based method and cannot determine the direction of association between the outcome and predictors, we used the correlation coefficient to represent the direction of association as a surrogate (i.e., positive/negative, or +/−) in the VI plots.

3. Results

3.1. Patient Characteristics of EOCRC and LOCRC

Table 2 shows the patient characteristics of EOCRC and LOCRC in the study area. Our study included 136,065 and 1,141,775 individuals who were newly diagnosed with invasive EOCRC and LOCRC, respectively, in the 1085 SEER counties. The overall EOCRC incidence rate and LOCRC incidence rate were 6.9 and 134.7 per 100,000 persons, respectively. Of those, 47.7% of EOCRC and 48.4% of LOCRC patients were women. Non-Hispanic White was the most common race/ethnicity group for both EOCRC and LOCRC patients. Compared to LOCRC, there were more EOCRC patients who were Non-Hispanic Black (13.9% vs. 11.2%), Non-Hispanic Asian or Pacific Islander (7.5% vs. 5.5%), and Hispanic (18.5% vs. 10.4%), while there were fewer EOCRC patients who were Non-Hispanic White (58.7% vs. 72.2%). The number of EOCRC cases increased substantially every 5 years, with a total increase of 31.6% from 2000–2004 to 2015–2019, while the number of LOCRC cases decreased by 10.1% over the same period. Regarding tumor site, there were more EOCRC patients who had tumors in the rectum or rectosigmoid junction compared to LOCRC patients (38.4% vs. 27.9%).

Table 2. Patient characteristics of EOCRC and LOCRC in the study area.

	EOCRC (Age < 50)	**LOCRC (Age 50+)**	***p*-Value**
N	136,065	1,141,775	
Incidence rate *	6.9	134.7	
Sex (%)			<0.0001
Male	71,127 (52.3)	588,623 (51.6)	
Female	64,938 (47.7)	553,152 (48.4)	
Race/Ethnicity (%)			<0.0001
Non-Hispanic White	79,807 (58.7)	824,210 (72.2)	
Non-Hispanic Black	18,953 (13.9)	127,910 (11.2)	
Non-Hispanic American Indian/Alaska Native	711 (0.5)	3632 (0.3)	
Non-Hispanic Asian or Pacific Islander	10,252 (7.5)	62,769 (5.5)	
Hispanic (All Races)	25,240 (18.5)	118,234 (10.4)	
Non-Hispanic Unknown Race	1102 (0.8)	5020 (0.4)	

Table 2. *Cont.*

	EOCRC (Age < 50)	LOCRC (Age 50+)	*p*-Value
Year of Diagnosis (%)			<0.0001
2000–2004	29,908 (22.0)	306,441 (26.8)	
2005–2009	32,582 (23.9)	287,722 (25.2)	
2010–2014	34,203 (25.1)	271,980 (23.8)	
2015–2019	39,372 (28.9)	275,632 (24.1)	
Tumor Site (%)			<0.0001
Colon excluding Sigmoid Colon	55,063 (40.6)	606,845 (53.1)	
Sigmoid Colon	28,722 (21.1)	216,986 (19.0)	
Rectum and Rectosigmoid Junction	52,280 (38.4)	317,944 (27.9)	

* Age-adjusted incidence rate per 100,000 persons. Abbreviations. EOCRC: early-onset colorectal cancer; LOCRC: late-onset colorectal cancer; NOS: not otherwise specified.

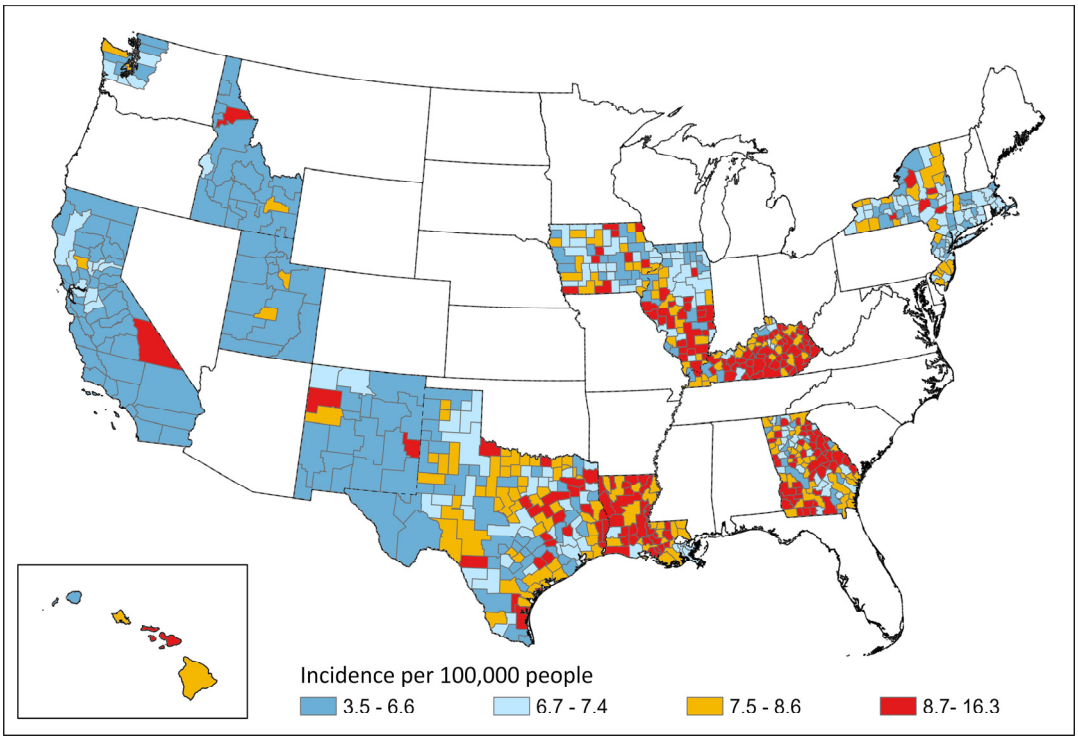
3.2. Geographic Distribution of EOCRC and LOCRC

The max-p-regions method created 709 CCs from the 1085 counties. The medians (IQR) of CC-level EOCRC and LOCRC incidence rates were 7.4 (6.6–8.5) and 142.0 (129.4–156.9) per 100,000 persons, respectively. Figure 1 shows the geographic distributions of the CC-level incidence rates of EOCRC and LOCRC by quartile in the SEER regions. Generally, states in the West (California, Washington (Seattle–Puget Sound), Idaho, Utah, and New Mexico) had lower EOCRC and LOCRC incidence rates than those in the Midwest (Iowa and Illinois) or the South (Kentucky, Louisiana, and Georgia). Kentucky, Louisiana, and Southern Illinois had the highest incidence rates of both EOCRC and LOCRC. Notably, Georgia and eastern Texas had relatively higher EOCRC incidence rates and lower LOCRC incidence rates compared to other areas. In contrast, Iowa, New Jersey, northern Illinois, and upstate New York had relatively lower EOCRC incidence rates and higher LOCRC incidence rates compared to other areas.

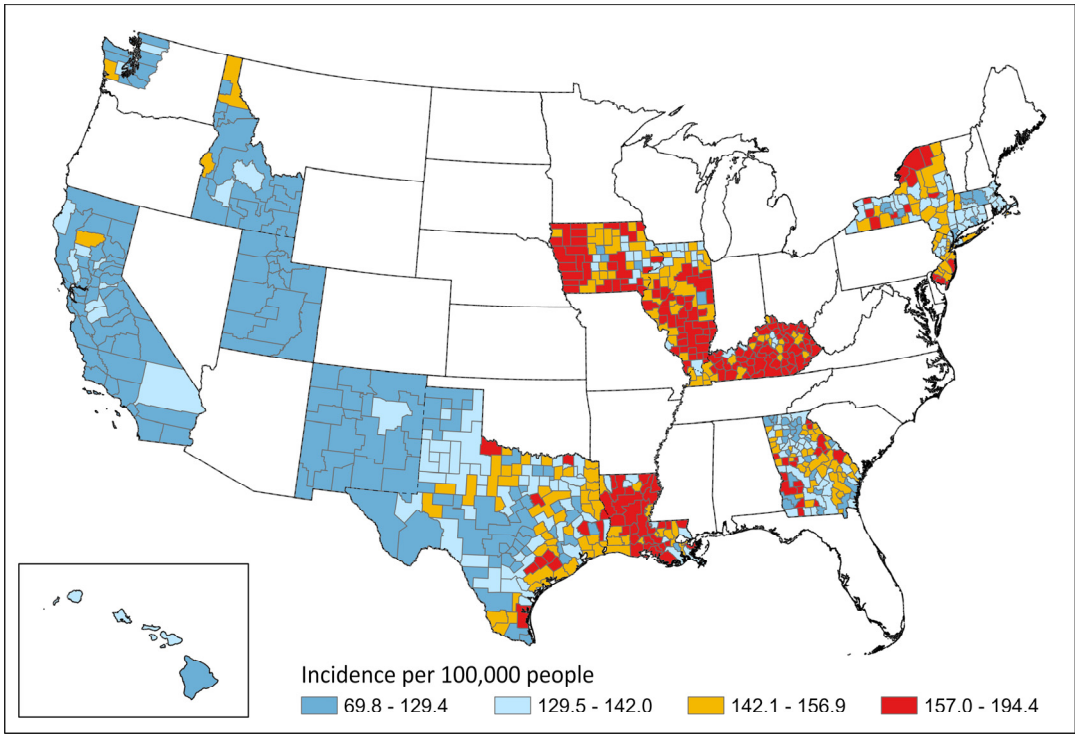
3.3. Importance of Risk Factors in Predicting EOCRC and LOCRC

Figure 2 shows the VI of risk factors from the random forest analyses in predicting EOCRC incidence rate and LOCRC incidence rate. The most important variable is set to 100%, and the VI of the rest of the variables is scaled relative to the most important variable. The position of the horizontal lines (+/−) indicates the direction of association determined by a linear correlation coefficient between the predictor and the outcome. We observed that the direction of association was consistent between the EOCRC and the LOCRC models for all risk factors. Figure 3 presents the differences in the rankings of the variables with the ten highest VI between the EOCRC and the LOCRC models.

For race/ethnicity, the Hispanic population predictor had a high VI (ranked second) in the EOCRC model and medium-high VI (ranked second) in the LOCRC model, with a negative association with both outcomes, meaning that both of these diseases were less likely to be in the Hispanic population. The Non-Hispanic Black predictor was positively associated with both EOCRC and LOCR, but their VI was low for both outcomes (20th and 10th, respectively). The Non-Hispanic White predictor was also positively associated with both EOCRC and LOCRC, and their VI was medium-low for both outcomes (11th for EOCRC and sixth for LOCRC).



(**A**) EOCRC incidence rate



(**B**) LOCRC incidence rate

Figure 1. Geographic distribution of (**A**) EOCRC incidence rate, and (**B**) LOCRC incidence rate. Notes: Incidence rates are age-adjusted and classified by quartile. Data for states in white are not available in SEER.

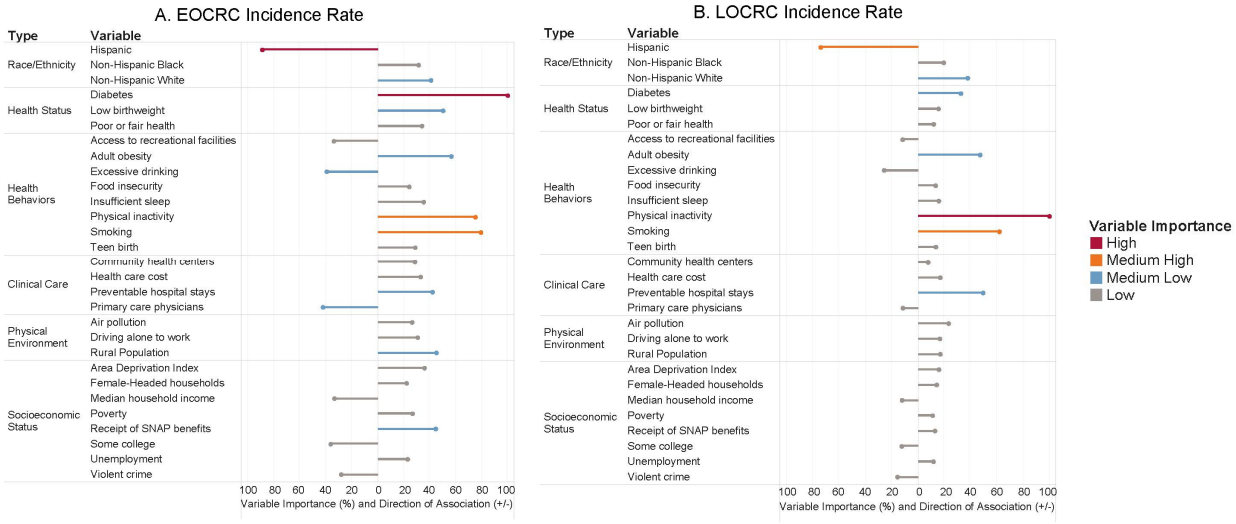


Figure 2. Variable importance and direction of association of risk factors for (**A**) EOCRC incidence rate, and (**B**) LOCRC incidence rate. Notes: (1) The most important variable is set to 100%. The importance of the rest of the variables is scaled relative to the most important variable. (2) Direction of association was determined by a linear correlation coefficient. (3) Categories of variable importance (i.e., high to low) was classified by the Jenks natural breaks method.

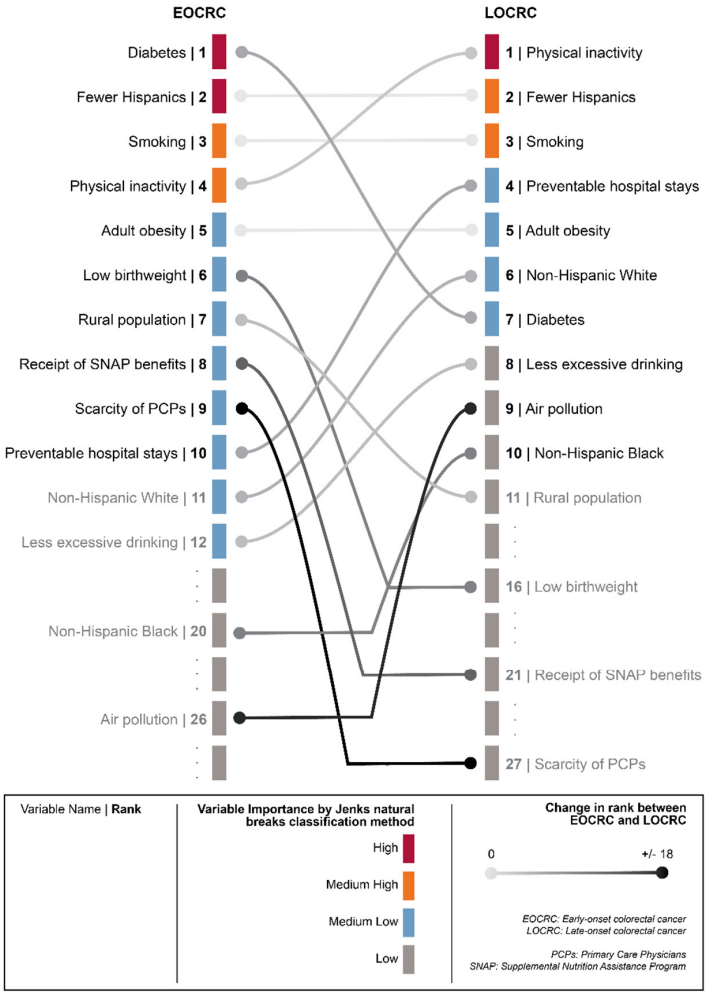


Figure 3. Comparisons of the top ten risk factors based on variable importance between EOCRC and LOCRC models.

For health status risk factors, diabetes prevalence had the highest VI (first) for EOCRC, and it had a medium-low VI (seventh) for LOCRC. Low birthweight was relatively more important in the EOCRC compared to that in the LOCRC model (sixth vs. 16th).

For health behavioral risk factors, both physical inactivity and smoking were positively associated with EOCRC and LOCRC, with physical inactivity ranked fourth for EOCRC and first for LOCRC, and smoking ranked third for both EOCRC and LOCRC. Adult obesity was ranked fifth (medium-low VI) in both models. Excessive drinking was negatively associated with EOCRC and LOCRC and had a medium-low VI (12th) for EOCRC and a low VI for LOCRC (eighth). Other health behavioral risk factors, including access to recreational facilities, food insecurity, insufficient sleep, and teen birth, had a low VI for both EOCRC and LOCRC outcomes.

For clinical care risk factors, the preventable hospital stays predictor was positively associated with both outcomes with a medium-low VI, but its rank of VI increased from 10th for EOCRC to fourth for LOCRC. The predictor of primary care physicians was negatively associated with both outcomes with its VI ranked 10th for EOCRC and 27th for LOCRC. Community health centers and health care cost had a low VI for both EOCRC and LOCRC, despite their positive associations with both outcomes.

Regarding physical environment of the composite counties, air pollution, driving alone to work (indicators of the transit system and physical inactivity), and rural population all had a positive association with both EOCRC and LOCRC outcomes. However, air pollution and driving alone to work had a low VI for both outcomes, and rural population had a medium-low VI (seventh) for EOCRC and a low VI (11th) for LOCRC.

4. Discussion

Using geographic and random forest methods, this cross-sectional study explored the geographic distribution and risk factor association of EOCRC incidence rate and LOCRC incidence rate across the US at the community level. We found substantial geographic disparities in the risk of EOCRC and LOCRC. Certain regions also had a relatively low LOCRC risk and high EOCRC risk (e.g., Georgia and eastern Texas) while other areas presented relatively high LOCRC risk and low EOCRC risk (e.g., Iowa and New Jersey). The random forest analyses revealed that certain risk factors, especially diabetes prevalence and physical inactivity, had different degrees of importance in predicting the incidence rate of EOCRC and the incidence rate of LOCRC.

Diabetes is a known risk factor for CRC [17]. For example, studies from the Netherlands and Sweden found that diabetes was associated with an increased risk of CRC for patients at a younger age compared to their older counterparts [18,19]. Reinforcing these previous studies, our study found that diabetes prevalence was the top important variable in predicting the incidence rate of EOCRC, but not a high-importance variable in predicting the incidence rate of LOCRC. An interesting geographic region of investigation for diabetes and its relationship with EOCRC is the "Diabetes Belt", defined by researchers as a region comprising 644 counties in the American South [20] (and includes the states of Georgia, Kentucky, and Louisiana in our study area). Previous research has also noted that people in the Diabetes Belt were more likely to be non-Hispanic Black, have a sedentary lifestyle, be obese, and have a smaller fitness/recreation facility density than in the rest of the US [20,21]. In our geographic analysis, we found that in Georgia, a state in the Diabetes Belt and whose diabetes prevalence was among the highest, most composite counties were in the top quartile of EOCRC incidence rates but were in lower quartiles of LOCRC incident rates (Figure 1), suggesting that diabetes has a stronger effect on EOCRC versus LOCRC. In recent years, the prevalence of type 2 diabetes in adolescents and young adults has dramatically increased [22,23], and increases have been disproportionally observed among African Americans [23,24]. Thus, future studies should continue to investigate the role of place, race, ethnicity, diabetes, and EOCRC.

Our models identified physical inactivity as having the highest VI for LOCRC and a medium-high VI for EOCRC, consistent with numerous studies suggesting that physical in-

activity or sedentary lifestyle is a substantial risk factor for CRC [3,5,25]. Physical inactivity is also associated with other chronic conditions, such as obesity and diabetes [26,27], which may also increase the risk of CRC. In recent years, sedentary-based work practices and passive behaviors such as smartphone use have increased among young adults, especially in the COVID-19 era [28,29]. A growing body of literature has also linked sedentary lifestyle to an increased risk of EOCRC [30–32]. Our study provides further evidence suggesting that increasing physical activities may be an effective and actionable risk reduction strategy for CRC for people of all ages.

As a risk factor strongly associated with diabetes and physical inactivity, adult obesity had a relatively low VI in predicting either EOCRC incidence rate or LOCRC incidence rate. Despite CRC being one of the obesity-associated cancers [33,34], associations between obesity and EOCRC have not been consistently shown in earlier studies [3,4,35].

Regarding other health behavioral risk factors, smoking was the third important predictor in both the EOCRC and LOCRC models in our study, which supports earlier findings on smoking being a risk factor for CRC [3–5]. Interestingly, excessive drinking was negatively associated with EOCRC and LOCRC in our study, though it should be noted that the risk factor had a relatively low VI. Alcohol consumption, especially heavy drinking, has been associated with increased CRC risk in previous studies [36–38]. This inconsistency may be because excessive drinking is confounded by factors such as the rurality of residence and diet. In another study, moderate alcohol consumption was associated with a reduced CRC risk in populations with greater adherence to a Mediterranean diet [39]. The lack of consistency warrants further investigation into the effect of alcohol consumption on the risk of EOCRC and LOCRC.

There were substantial differences in the risk of EOCRC and LOCRC among racial/ethnic groups. Recent epidemiological data show that the incidence of CRC for Hispanics was 25% and 12% lower than that for Non-Hispanic Blacks and Non-Hispanic Whites in the US, respectively [1]. However, this difference narrowed for patients older than 50, as suggested by a population-based study of patients from California [40]. Despite the negative association between Hispanic population and CRC risk from our study, our data in Table 2 show that the proportion of Hispanic patients with EOCRC was much higher than those with LOCRC (18.5% vs. 10.4%). This might be attributable to changes in lifestyle between the US-born Hispanic population versus first-generation Hispanic immigrants; not only are US-born Hispanics younger on average compared to first-generation Hispanic immigrants [41], but they also report lower intake of fruits and vegetables [42], higher rates of smoking [43], higher rates of physical inactivity [44], and higher rates of obesity [45]. Previous studies have also shown that Hispanics born in the US have a higher risk of colorectal cancer death compared to their foreign-born counterparts [46] and even Non-Hispanic Whites [47]. Thus, as the proportion of US-born Hispanics increases and the proportion of foreign-born Hispanics decreases [48], the association between Hispanic ethnicity and CRC burden may evolve [49].

Finally, compared to race/ethnicity, health status, and health behaviors, risk factors related to clinical care, physical environment, and socioeconomic status did not show high VI in predicting either EOCRC or LOCRC incidence rate. However, these risk factors might still influence the risk of EOCRC and LOCRC through their impact on other risk factors such as smoking, physical inactivity, obesity, and diabetes, as suggested by their directions of association in Figure 2. Future studies should investigate the causal pathways between these risk factors and CRC.

Our findings should be interpreted in light of the following limitations. First, we were not able to consider temporal variations in either the outcomes or the predictors during the 20-year study period. Second, this community-level study did not consider individual-level characteristics in the analyses. However, many of the associations identified from our study are consistent with previous individual-level studies on risk factors of EOCRC and LOCRC. Finally, we could not consider spatially varying associations [50] between the risk factors and EOCRC and LOCRC due to the relatively small number of geographic units (composite counties) and non-contiguous regions in the study area.

We note that this random forest approach is data-driven and results are stable and reproducible based on the data inputs. Future studies using new additional data may impact the results of EOCRC and LOCRC. Also, the focus here is not to discover causal pathways for EOCRC or LOCRC, noting that causality cannot be established from this study, given its cross-sectional nature. In addition, the results obtained from data-driven approaches should be interpreted with caution, and the plausibility of the emerging associations should be determined in light of the relevant body of knowledge. Nevertheless, evaluating the association between community-level risk factors and outcomes is an essential first step toward establishing causality.

5. Conclusions

This community-level study evaluated and compared associations between various risk factors and EOCRC and LOCRC using novel geographic and machine learning approaches. The geographic analysis revealed regions with an elevated risk of EOCRC but a lower risk of LOCRC, identifying regions with a disproportionate burden of EOCRC. The machine learning analysis identified risk factor profiles specific to EOCRC and LOCRC. Collectively, these findings can help facilitate future studies that further uncover actionable interventions to reduce EOCRC and guide where targeted interventions to reduce EOCRC burden should be deployed.

Author Contributions: Conceptualization, W.D., U.K., J.R., N.A.B. and S.M.K.; methodology, W.D., U.K., J.R., M.K., S.L., N.A.B. and S.M.K.; validation, W.D.; investigation, W.D.; resources, N.A.B. and S.M.K.; data curation, W.D. and U.K.; writing—original draft preparation, W.D.; writing—review and editing, All Authors; visualization, W.D., U.K., J.R. and S.L.; supervision, N.A.B. and S.M.K.; funding acquisition, J.R. and S.M.K. We confirm the originality of content. W.D. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Cancer Institute Case Comprehensive Cancer Center (P30 CA043703).

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the fact that this study did not constitute human participant research.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be shared upon request.

Conflicts of Interest: Weichuan Dong and Siran M. Koroukian reported receiving grants from American Cancer Society (RWIA-20-111-02 RWIA) and by contracts from Cleveland Clinic Foundation, including a subcontract from Celgene Corporation. Siran M. Koroukian was also supported by grants from the Centers for Disease Control and Prevention, U48 DP005030-05S1 and U48 DP006404-03S7; National Institutes of Health (R15 NR017792, UH3-DE025487, and R01 AG074946-01) and American Cancer Society (132678-RSGI-19-213-01-CPHPS). Uriel Kim is supported by grants from the National Institute of General Medical Sciences (5T32GM007250), National Center for Advancing Translational Sciences (5TL1TR002549), and the PhRMA Foundation (PDHO18). Johnie Rose reported receiving grants from NIH/National Cancer Institute during the conduct of the study; holding stock in Vinya Intelligence Inc outside the submitted work; and having a patent issued for US 270,799 B2 "In-home remote monitoring systems and methods for predicting health status decline". No other disclosures were reported.

References

1. Siegel, R.L.; Miller, K.D.; Goding Sauer, A.; Fedewa, S.A.; Butterly, L.F.; Anderson, J.C.; Cercek, A.; Smith, R.A.; Jemal, A. Colorectal cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 145–164. [CrossRef]
2. Wang, W.; Chen, W.; Lin, J.; Shen, Q.; Zhou, X.; Lin, C. Incidence and characteristics of young-onset colorectal cancer in the United States: An analysis of SEER data collected from 1988 to 2013. *Clin. Res. Hepatol. Gastroenterol.* **2019**, *43*, 208–215. [CrossRef]
3. Archambault, A.N.; Lin, Y.; Jeon, J.; Harrison, T.A.; Bishop, D.T.; Brenner, H.; Casey, G.; Chan, A.T.; Chang-Claude, J.; Figueiredo, J.C.; et al. Nongenetic Determinants of Risk for Early-Onset Colorectal Cancer. *JNCI Cancer Spectr.* **2021**, *5*, pkab029. [CrossRef]

4. Hayes, R.B. Advances in Understanding Early-Onset Colorectal Cancer. *Cancer Epidemiol. Biomark. Prev.* **2021**, *30*, 1775–1777. [CrossRef]
5. Sinicrope, F.A. Increasing Incidence of Early-Onset Colorectal Cancer. *N. Engl. J. Med.* **2022**, *386*, 1547–1558. [CrossRef] [PubMed]
6. Mauri, G.; Sartore-Bianchi, A.; Russo, A.G.; Marsoni, S.; Bardelli, A.; Siena, S. Early-onset colorectal cancer in young individuals. *Mol. Oncol.* **2019**, *13*, 109–131. [CrossRef] [PubMed]
7. Doubeni, C.A.; Laiyemo, A.O.; Major, J.M.; Schootman, M.; Lian, M.; Park, Y.; Graubard, B.I.; Hollenbeck, A.R.; Sinha, R. Socioeconomic status and the risk of colorectal cancer: An analysis of more than a half million adults in the National Institutes of Health-AARP Diet and Health Study. *Cancer* **2012**, *118*, 3636–3644. [CrossRef] [PubMed]
8. Surveillance, Epidemiology, and End Results (SEER) Program, SEER*Stat Database: Incidence—SEER Research Plus Limited-Field Data, 22 Registries, November 2021 Sub (2000–2019)—Linked To County Attributes—Total U.S. 1969-2020 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2022, based on the November 2021 Submission. Available online: www.seer.cancer.gov (accessed on 28 November 2022).
9. Surveillance, Epidemiology, and End Results (SEER) Program. Overview of the SEER Program. Available online: https://seer.cancer.gov/about/overview.html (accessed on 27 January 2023).
10. County Health Rankings & Roadmaps. Available online: https://www.countyhealthrankings.org (accessed on 27 January 2023).
11. Health Resources and Services Administration—Area Health Resources Files. Available online: https://data.hrsa.gov/topics/health-workforce/ahrf (accessed on 27 January 2023).
12. Duque, J.C.; Anselin, L.; Rey, S.J. The Max-P-Regions Problem*. *J. Reg. Sci.* **2012**, *52*, 397–419. [CrossRef]
13. Singh, G.K. Area deprivation and widening inequalities in US mortality, 1969–1998. *Am. J. Public Health* **2003**, *93*, 1137–1143. [CrossRef]
14. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
15. Dong, W.; Bensken, W.P.; Kim, U.; Rose, J.; Berger, N.A.; Koroukian, S.M. Phenotype Discovery and Geographic Disparities of Late-Stage Breast Cancer Diagnosis across U.S. Counties: A Machine Learning Approach. *Cancer Epidemiol. Biomark. Prev.* **2022**, *31*, 66–76. [CrossRef]
16. Jenks, G.F. The data model concept in statistical mapping. *Int. Yearb. Cartogr.* **1967**, *7*, 186–190.
17. Ma, Y.; Yang, W.; Song, M.; Smith-Warner, S.A.; Yang, J.; Li, Y.; Ma, W.; Hu, Y.; Ogino, S.; Hu, F.B.; et al. Type 2 diabetes and risk of colorectal cancer in two large U.S. prospective cohorts. *Br. J. Cancer* **2018**, *119*, 1436–1442. [CrossRef] [PubMed]
18. de Kort, S.; Masclee, A.A.; Sanduleanu, S.; Weijenberg, M.P.; van Herk-Sukel, M.P.; Oldenhof, N.J.; van den Bergh, J.P.; Haak, H.R.; Janssen-Heijnen, M.L. Higher risk of colorectal cancer in patients with newly diagnosed diabetes mellitus before the age of colorectal cancer screening initiation. *Sci. Rep.* **2017**, *7*, 46527. [CrossRef]
19. Ali Khan, U.; Fallah, M.; Tian, Y.; Sundquist, K.; Sundquist, J.; Brenner, H.; Kharazmi, E. Personal History of Diabetes as Important as Family History of Colorectal Cancer for Risk of Colorectal Cancer: A Nationwide Cohort Study. *Am. J. Gastroenterol.* **2020**, *115*, 1103–1109. [CrossRef] [PubMed]
20. Barker, L.E.; Kirtland, K.A.; Gregg, E.W.; Geiss, L.S.; Thompson, T.J. Geographic distribution of diagnosed diabetes in the U.S.: A diabetes belt. *Am. J. Prev. Med.* **2011**, *40*, 434–439. [CrossRef]
21. Myers, C.A.; Slack, T.; Broyles, S.T.; Heymsfield, S.B.; Church, T.S.; Martin, C.K. Diabetes prevalence is associated with different community factors in the diabetes belt versus the rest of the United States. *Obesity* **2017**, *25*, 452–459. [CrossRef]
22. Lascar, N.; Brown, J.; Pattison, H.; Barnett, A.H.; Bailey, C.J.; Bellary, S. Type 2 diabetes in adolescents and young adults. *Lancet Diabetes Endocrinol.* **2018**, *6*, 69–80. [CrossRef] [PubMed]
23. Virani, S.S.; Alonso, A.; Aparicio, H.J.; Benjamin, E.J.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Cheng, S.; Delling, F.N.; et al. Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. *Circulation* **2021**, *143*, e254–e743. [CrossRef]
24. Wang, M.C.; Shah, N.S.; Carnethon, M.R.; O'Brien, M.J.; Khan, S.S. Age at Diagnosis of Diabetes by Race and Ethnicity in the United States From 2011 to 2018. *JAMA Intern. Med.* **2021**, *181*, 1537–1539. [CrossRef] [PubMed]
25. Slattery, M.L.; Potter, J.; Caan, B.; Edwards, S.; Coates, A.; Ma, K.N.; Berry, T.D. Energy balance and colon cancer—beyond physical activity. *Cancer Res.* **1997**, *57*, 75–80. Available online: https://aacrjournals.org/cancerres/article/57/1/75/503178/Energy-Balance-and-Colon-Cancer-beyond-Physical (accessed on 27 January 2023). [PubMed]
26. Hu, F.B.; Li, T.Y.; Colditz, G.A.; Willett, W.C.; Manson, J.E. Television watching and other sedentary behaviors in relation to risk of obesity and type 2 diabetes mellitus in women. *JAMA* **2003**, *289*, 1785–1791. [CrossRef]
27. Qi, Q.; Li, Y.; Chomistek, A.K.; Kang, J.H.; Curhan, G.C.; Pasquale, L.R.; Willett, W.C.; Rimm, E.B.; Hu, F.B.; Qi, L. Television watching, leisure time physical activity, and the genetic predisposition in relation to body mass index in women and men. *Circulation* **2012**, *126*, 1821–1827. [CrossRef] [PubMed]
28. Ricci, F.; Izzicupo, P.; Moscucci, F.; Sciomer, S.; Maffei, S.; Di Baldassarre, A.; Mattioli, A.V.; Gallina, S. Recommendations for Physical Inactivity and Sedentary Behavior During the Coronavirus Disease (COVID-19) Pandemic. *Front. Public Health* **2020**, *8*, 199. [CrossRef] [PubMed]
29. Wilms, P.; Schröder, J.; Reer, R.; Scheit, L. The Impact of "Home Office" Work on Physical Activity and Sedentary Behavior during the COVID-19 Pandemic: A Systematic Review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12344. [CrossRef]
30. Nguyen, L.H.; Liu, P.H.; Zheng, X.; Keum, N.; Zong, X.; Li, X.; Wu, K.; Fuchs, C.S.; Ogino, S.; Ng, K.; et al. Sedentary Behaviors, TV Viewing Time, and Risk of Young-Onset Colorectal Cancer. *JNCI Cancer Spectr.* **2018**, *2*, pky073. [CrossRef] [PubMed]

31. Rogers, C.R.; Moore, J.X.; Qeadan, F.; Gu, L.Y.; Huntington, M.S.; Holowatyj, A.N. Examining factors underlying geographic disparities in early-onset colorectal cancer survival among men in the United States. *Am. J. Cancer Res.* **2020**, *10*, 1592–1607. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7269786 (accessed on 27 January 2023).
32. Schmid, D.; Leitzmann, M.F. Television viewing and time spent sedentary in relation to cancer risk: A meta-analysis. *J. Natl. Cancer Inst.* **2014**, *106*, dju098. [CrossRef]
33. Steele, C.B.; Thomas, C.C.; Henley, S.J.; Massetti, G.M.; Galuska, D.A.; Agurs-Collins, T.; Puckett, M.; Richardson, L.C. Vital Signs: Trends in Incidence of Cancers Associated with Overweight and Obesity—United States, 2005–2014. *MMWR Morb. Mortal. Wkly. Rep.* **2017**, *66*, 1052–1058. [CrossRef]
34. Koroukian, S.M.; Dong, W.; Berger, N.A. Changes in Age Distribution of Obesity-Associated Cancers. *JAMA Netw. Open* **2019**, *2*, e199261. [CrossRef]
35. Liu, P.H.; Wu, K.; Ng, K.; Zauber, A.G.; Nguyen, L.H.; Song, M.; He, X.; Fuchs, C.S.; Ogino, S.; Willett, W.C.; et al. Association of Obesity With Risk of Early-Onset Colorectal Cancer Among Women. *JAMA Oncol.* **2019**, *5*, 37–44. [CrossRef] [PubMed]
36. Fedirko, V.; Tramacere, I.; Bagnardi, V.; Rota, M.; Scotti, L.; Islami, F.; Negri, E.; Straif, K.; Romieu, I.; La Vecchia, C.; et al. Alcohol drinking and colorectal cancer risk: An overall and dose-response meta-analysis of published studies. *Ann. Oncol.* **2011**, *22*, 1958–1972. [CrossRef]
37. Moskal, A.; Norat, T.; Ferrari, P.; Riboli, E. Alcohol intake and colorectal cancer risk: A dose-response meta-analysis of published cohort studies. *Int. J. Cancer* **2007**, *120*, 664–671. [CrossRef]
38. Pedersen, A.; Johansen, C.; Grønbaek, M. Relations between amount and type of alcohol and colon and rectal cancer in a Danish population based cohort study. *Gut* **2003**, *52*, 861–867. [CrossRef] [PubMed]
39. Klarich, D.S.; Brasser, S.M.; Hong, M.Y. Moderate Alcohol Consumption and Colorectal Cancer Risk. *Alcohol Clin. Exp. Res.* **2015**, *39*, 1280–1291. [CrossRef]
40. Ellis, L.; Abrahão, R.; McKinley, M.; Yang, J.; Somsouk, M.; Marchand, L.L.; Cheng, I.; Gomez, S.L.; Shariff-Marco, S. Colorectal Cancer Incidence Trends by Age, Stage, and Racial/Ethnic Group in California, 1990–2014. *Cancer Epidemiol. Biomark. Prev.* **2018**, *27*, 1011–1018. [CrossRef] [PubMed]
41. Patten, E. The Nation's Latino Population Is Defined by Its Youth. Pew Research Center 2016. Available online: https://www.pewresearch.org/hispanic/2016/04/20/the-nations-latino-population-is-defined-by-its-youth/ (accessed on 27 January 2023).
42. Bermudez, O.I.; Falcon, L.M.; Tucker, K.L. Intake and food sources of macronutrients among older Hispanic adults: Association with ethnicity, acculturation, and length of residence in the United States. *J. Am. Diet. Assoc.* **2000**, *100*, 665–673. [CrossRef] [PubMed]
43. Pulvers, K.; Cupertino, A.P.; Scheuermann, T.S.; Cox, L.S.; Ho, Y.Y.; Nollen, N.L.; Cuellar, R.; Ahluwalia, J.S. Daily and Nondaily Smoking Varies by Acculturation among English-Speaking, US Latino Men and Women. *Ethn. Dis.* **2018**, *28*, 105–114. [CrossRef]
44. Joseph, R.P.; Benitez, T.J.; Ainsworth, B.E.; Todd, M.; Keller, C. Acculturation and Physical Activity among Latinas Enrolled in a 12-Month Walking Intervention. *West. J. Nurs. Res.* **2018**, *40*, 942–960. [CrossRef]
45. Lara, M.; Gamboa, C.; Kahramanian, M.I.; Morales, L.S.; Hayes Bautista, D.E. Acculturation and Latino health in the United States: A review of the literature and its sociopolitical context. *Annu. Rev. Public Health* **2005**, *26*, 367–397. [CrossRef]
46. Chen, H.; Wu, A.H.; Wang, S.; Bookstein, A.; Le Marchand, L.; Wilkens, L.R.; Haiman, C.A.; Cheng, I.; Monroe, K.R.; Setiawan, V.W. Cancer Mortality Patterns by Birthplace and Generation Status of Mexican Latinos: The Multiethnic Cohort. *J. Natl. Cancer Inst.* **2022**, *114*, 959–968. [CrossRef] [PubMed]
47. Pinheiro, P.S.; Callahan, K.E.; Gomez, S.L.; Marcos-Gragera, R.; Cobb, T.R.; Roca-Barcelo, A.; Ramirez, A.G. High cancer mortality for US-born Latinos: Evidence from California and Texas. *BMC Cancer* **2017**, *17*, 478. [CrossRef] [PubMed]
48. Gonzalez-Barrera, A. More Mexicans Leaving Than Coming to the US. Pew Research Center 2015. Available online: https://www.pewresearch.org/hispanic/2015/11/19/more-mexicans-leaving-than-coming-to-the-u-s/ (accessed on 27 January 2023).
49. Pinheiro, P.S.; Callahan, K.E.; Stern, M.C.; de Vries, E. Migration from Mexico to the United States: A high-speed cancer transition. *Int. J. Cancer* **2018**, *142*, 477–488. [CrossRef]
50. Dong, W.; Bensken, W.P.; Kim, U.; Rose, J.; Fan, Q.; Schiltz, N.K.; Berger, N.A.; Koroukian, S.M. Variation in and Factors Associated With US County-Level Cancer Mortality, 2008–2019. *JAMA Netw. Open* **2022**, *5*, e2230925. [CrossRef] [PubMed]