Journal Pre-proof

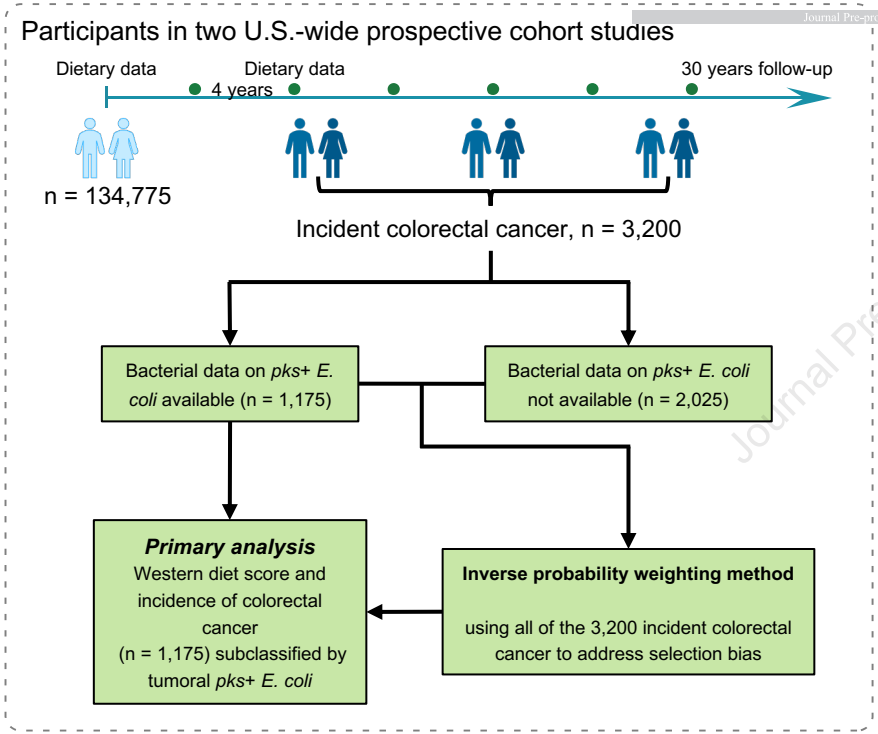
Western-style Diet, *pks* Island-Carrying *Escherichia coli*, and Colorectal Cancer: Analyses from Two Large Prospective Cohort Studies

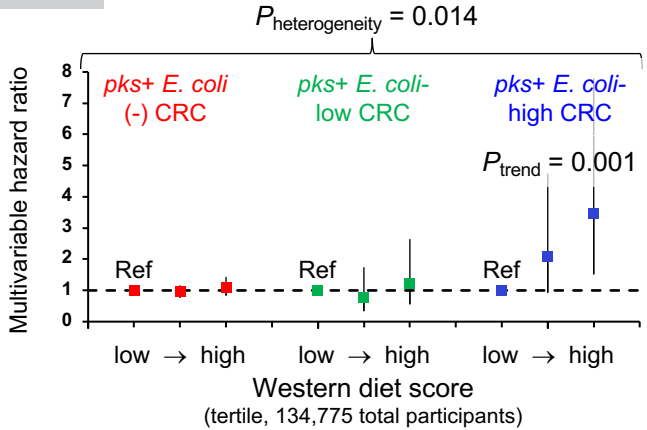
Kota Arima, Rong Zhong, Tomotaka Ugai, Melissa Zhao, Koichiro Haruki, Naohiko Akimoto, Mai Chan Lau, Kazuo Okadome, Raaj S. Mehta, Juha P. Väyrynen, Junko Kishikawa, Tyler S. Twombly, Shanshan Shi, Kenji Fujiyoshi, Keisuke Kosumi, Yoko Ogata, Hideo Baba, Fenglei Wang, Kana Wu, Mingyang Song, Xuehong Zhang, Charles S. Fuchs, Cynthia L. Sears, Walter C. Willett, Edward L. Giovannucci, Jeffrey A. Meyerhardt, Wendy S. Garrett, Curtis Huttenhower, Andrew T. Chan, Jonathan A. Nowak, Marios Giannakis, Shuji Ogino

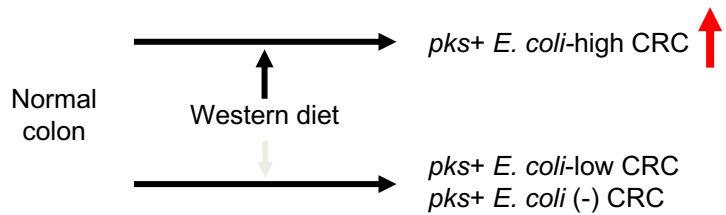
PII: S0016-5085(22)00672-2
DOI: https://doi.org/10.1053/j.gastro.2022.06.054
Reference: YGAST 65158

To appear in: *Gastroenterology*
Accepted Date: 15 June 2022

Please cite this article as: Arima K, Zhong R, Ugai T, Zhao M, Haruki K, Akimoto N, Lau MC, Okadome K, Mehta RS, Väyrynen JP, Kishikawa J, Twombly TS, Shi S, Fujiyoshi K, Kosumi K, Ogata Y, Baba H, Wang F, Wu K, Song M, Zhang X, Fuchs CS, Sears CL, Willett WC, Giovannucci EL, Meyerhardt JA, Garrett WS, Huttenhower C, Chan AT, Nowak JA, Giannakis M, Ogino S, Western-style Diet, *pks* Island-Carrying *Escherichia coli*, and Colorectal Cancer: Analyses from Two Large Prospective Cohort Studies, *Gastroenterology* (2022), doi: https://doi.org/10.1053/j.gastro.2022.06.054.







Gastroenterology

Original Research Article

Western-style Diet, *pks* Island-Carrying *Escherichia coli*, and Colorectal Cancer: Analyses from Two Large Prospective Cohort Studies

Authors
Kota Arima[1,2,3], Rong Zhong[1,4,5], Tomotaka Ugai[1,4], Melissa Zhao[1], Koichiro Haruki[1,3], Naohiko Akimoto[1], Mai Chan Lau[1], Kazuo Okadome[1], Raaj S. Mehta[6,7], Juha P. Väyrynen[1,3,8], Junko Kishikawa[1], Tyler S. Twombly[1], Shanshan Shi[1], Kenji Fujiyoshi[1], Keisuke Kosumi[1], Yoko Ogata[2], Hideo Baba[2], Fenglei Wang[9], Kana Wu[9], Mingyang Song[6,7,9], Xuehong Zhang[9,10], Charles S. Fuchs[11,12,13,14], Cynthia L. Sears[15], Walter C. Willett[9,10], Edward L. Giovannucci[4,9,10], Jeffrey A. Meyerhardt[3], Wendy S. Garrett[3,16,17], Curtis Huttenhower[18], Andrew T. Chan[6,7,10,16,19], Jonathan A. Nowak[1], Marios Giannakis[3,17,19], and Shuji Ogino[1,4,17,20].

The first four authors contributed equally as co-first authors.
The last four authors contributed equally as co-last authors.

Author Affiliations
[1] Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; [2] Department of Gastroenterological Surgery, Graduate School of Medical Sciences, Kumamoto University, Kumamoto, Japan; [3] Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA; [4] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; [5] Department of Epidemiology and Biostatistics and Ministry of Education Key Lab of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China; [6] Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA; [7] Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA; [8] Cancer and Translational Medicine Research Unit, Medical Research Center Oulu, Oulu University Hospital, and University of Oulu, Oulu, Finland; [9] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA; [10] Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; [11] Yale Cancer Center, New Haven, CT; [12] Department of Medicine, Yale School of Medicine, New Haven, CT; [13] Smilow Cancer Hospital, New Haven, CT; [14] Genentech, South San Francisco, CA; [15] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; [16] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA; [17] Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA; [18] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; [19] Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; [20] Cancer Immunology and Cancer Epidemiology Programs, Dana-Farber Harvard Cancer Center, Boston, MA.

Word Count (Text): 6,432 words (inclusive of main text, references, and figure/table legends)
Figures and Tables: 5 tables, 2 figures, 8 supplementary tables, and 1 supplementary figure.

Short Running Head: Diet, Colorectal Cancer Incidence, *pks*+ *Escherichia coli*

Funding: This work was supported by U.S. National Institutes of Health (NIH) grants (P01 CA87969 to M.J. Stampfer; UM1 CA186107 to M.J. Stampfer; P01 CA55075 to W.C. Willett; UM1 CA167552 to W.C. Willett; U01 CA167552 to W.C. Willett and L.A. Mucci; P50 CA127003 to C.S.F.; R01 CA118553 to C.S.F.; R01 CA169141 to C.S.F.; R01 CA137178 to A.T.C.; K24 DK098311 to A.T.C.; R35 CA197735 to S.O.; R01 CA151993 to S.O.; K07 CA190673 to R.N.; and K07 CA188126 to X.Z.); by Cancer Research UK Grand Challenge Award (UK C10674/A27140 to C.L.S., W.S.G., C.H., M.G., and S.O.); by the Stand Up to Cancer Colorectal Cancer Dream Team Translational Research Grant (SU2C-AACR-DT22-17 to C.S.F. and M.G.), administered by the American Association for Cancer Research, a scientific partner of SU2C; and by grants from the Project P Fund, The Friends of the Dana-Farber Cancer Institute, Bennett Family Fund, and the Entertainment Industry Foundation through National Colorectal Cancer Research Alliance. K.A. and T.U. were supported by a grant from Overseas Research Fellowship (201860083 to K.A., 201960541 to T.U.) from Japan Society for the Promotion of Science. R.Z. was supported by a fellowship grant from Huazhong University of Science and Technology. T.U., K.H., and K.F. were supported by fellowship grants from the Uehara Memorial Foundation. T.U. was supported by a grant from Yasuda Medical Foundation. K.H. was supported by the Mitsukoshi Health and Welfare Foundation. A.T.C. is a Stuart and Suzanne Steele MGH Research Scholar. J.A.M. research is supported by the Douglas Gray Woodruff Chair fund, the Guo Shu Shi Fund, Anonymous Family Fund for Innovations in Colorectal Cancer, Project P fund, and the George Stone Family Foundation. M.G. was supported by a Conquer Cancer Foundation of ASCO Career Development Award.
The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Corresponding Author
Shuji Ogino, MD, PhD, MS
Program in MPE Molecular Pathological Epidemiology,
Department of Pathology,
Brigham and Women's Hospital
221 Longwood Avenue, EBRC Room 404A, Boston, MA, 02115
Tel: 617-732-5161; Fax: 617-264-5149
Email: sogino@bwh.harvard.edu

Disclosure of Potential Conflicts of Interest: A.T.C. previously served as a consultant for Bayer Healthcare and Pfizer Inc. J.A.M. has also served as an advisor/consultant to Ignyta, Array Pharmaceutical, and Cota Healthcare. C.S.F. is

currently employed by Genentech, a subsidiary of Roche, and previously served as a consultant for Agios, Bain Capital, Bayer, Celgene, Dicerna, Five Prime Therapeutics, Gilead Sciences, Eli Lilly, Entrinsic Health, Genentech, KEW, Merck, Merrimack Pharmaceuticals, Pfizer Inc, Sanofi, Taiho, and Unum Therapeutics; C.S.F. also serves as a Director for CytomX Therapeutics and owns unexercised stock options for CytomX and Entrinsic Health. M.G. receives research funding from Bristol-Myers Squibb, Merck, Servier, and Janssen. This study was not funded by any of these commercial entities. No other conflicts of interest exist. The other authors declare that they have no conflicts of interest.

Abbreviations: AJCC, American Joint Committee on Cancer; CI, confidence interval; CIMP, CpG island methylator phenotype; Ct, cycle threshold; FFPE, formalin-fixed paraffin-embedded; HPFS, Health Professionals Follow-up Study; HR, hazard ratio; IPW, inverse probability weighting; LINE-1, long-interspersed nucleotide element-1; MSI, microsatellite instability; NHS, Nurses' Health Study; PCR, polymerase chain reaction; pks, polyketide synthase; SD, standard deviation.

Authors' Contributions: Study concept and design: S.O. Acquisition, analysis, or interpretation of data: K.A., R.Z., T.U., M.Z., K.H., N.A., M.C.L., K.O., R.S.M., J.P.V., J.K., T.S.T., S.S., K.F., K.K., Y.O., M.S., X.Z., C.L.S., W.S.G., C.H., J.A.N., M.G., and S.O. Drafting of the manuscript: K.A., R.Z., T.U., M.Z., K.H., N.A., M.C.L., J.P.V., J.K., T.S.T., K.K., H.B., F.W., K.W., M.S., X.Z., E.L.G., J.A.M., W.S.G., C.H., M.G., and S.O. Critical revision of the manuscript for important intellectual content: K.A., R.Z., T.U., M.S., X.Z., C.S.F., C.L.S., W.C.W., E.L.G., J.A.M., W.S.G., C.H., A.T.C., J.A.N., M.G., and S.O. Statistical analysis: K.A., R.Z., T.U., and S.O. Obtained funding: X.Z., C.S.F., C.L.S., W.C.W., W.S.G., C.H., A.T.C., M.G., and S.O. Study supervision: C.S.F., C.L.S., W.C.W., E.L.G., J.A.N., M.G., and S.O. All authors read and approved the final manuscript.

Use of Standardized Official Symbols: We use HUGO (Human Genome Organisation) Gene Nomenclature Committee-approved official symbols (or root symbols) for genes and gene products, including BRAF, CACNA1G, CD14, CDKN2A, CRABP1, IGF2, KRAS, MLH1, NEUROG1, PIK3CA, RUNX3, SLCO2A1, SOCS1, TP53, and WNT; all of which are described at www.genenames.org. Gene symbols are italicized whereas symbols for gene products are not italicized.

ABSTRACT

BACKGROUND AND AIMS: Evidence supports a carcinogenic role of *Escherichia coli* carrying the polyketide synthase (*pks*) island that encodes enzymes for colibactin biosynthesis. We hypothesized that the association of western-style diet (rich in red and processed meat) with colorectal cancer incidence might be stronger for tumors containing higher amounts of *pks*+ *E. coli*.

METHODS: Western diet score was calculated using food frequency questionnaire data obtained every four years during follow-up of 134,775 participants in two U.S.-wide prospective cohort studies. Using quantitative polymerase chain reaction, we measured *pks*+ *E. coli* DNA in 1,175 tumors among 3,200 incident colorectal cancer cases that had occurred during the follow-up. We utilized the 3,200 cases and inverse probability weighting (to adjust for selection bias due to tissue availability), integrated in multivariable-adjusted duplication-method Cox proportional hazards regression analyses.

RESULTS: The association of the western diet score with colorectal cancer incidence was stronger for tumors containing higher levels of *pks*+ *E. coli* ($P_{heterogeneity}$ = 0.014). Multivariable-adjusted hazard ratios (with 95% confidence interval) for the highest (vs. lowest) tertile of the western diet score were 3.45 (1.53-7.78) (P_{trend} = 0.001) for *pks*+ *E. coli*-high tumors, 1.22 (0.57-2.63) for *pks*+ *E. coli*-low tumors, and 1.10 (0.85-1.42) for *pks*+ *E. coli*-negative tumors. The *pks*+ *E. coli* level was associated with lower disease

stage but not with tumor location, microsatellite instability, or *BRAF*, *KRAS*, or *PIK3CA* mutations.

CONCLUSIONS: Western-style diet is associated with higher incidence of colorectal cancer containing abundant *pks*+ *E. coli*, supporting a potential link between diet, the intestinal microbiota, and colorectal carcinogenesis.

KEYWORDS: immunology; microbiome; molecular pathological epidemiology.

INTRODUCTION

Accumulating evidence indicates that certain intestinal microorganisms influence colorectal tumor development through DNA damage, inflammation, and other mechanisms[1-5]. Among intestinal bacteria, *Escherichia coli* (*E. coli*) strains of the B2 phylotype commonly harbor the 54-kb *polyketide synthase* (*pks*) pathogenicity island that encodes enzymes for colibactin biosynthesis[1,3,6]. Experimental studies have shown that colibactin can alkylate DNA, induce DNA double-strand breaks, and cause a specific somatic mutational pattern in human cells[7-9]. However, human population studies are needed to better understand the role of *pks* island-carrying *E. coli* (hereafter referred to as *pks*+ *E. coli*) in colorectal cancer.

Diet and nutrition are considered crucial factors for colorectal cancer development. A meta-analysis has shown a weak-to-modest relationship between western dietary patterns and colorectal cancer risk[10]. An experimental study indicates that western-style diet (characterized by high intake of red and processed meat, sugar, and refined grains and low intake of in vegetables and legumes) can induce systemic and intestinal inflammation[11]. Considering the possible interplay between diet and pathogenic bacteria, it is of particular interest to study western-style diet in relation to *pks*+ *E. coli* within colorectal tumor tissue. Such analyses may contribute to the development of cancer prevention strategies targeting diet and microbiota.

In this study, we tested the hypothesis that the association of western-style diet with colorectal cancer incidence might be stronger for tumors containing higher amounts of *pks*+ *E. coli*. We utilized a molecular pathological epidemiology database of two

U.S.-wide longitudinal prospective cohort studies with incident colorectal cancer cases. This comprehensive dataset offered a unique opportunity to examine long-term dietary patterns of individuals (who had not known whether they would develop cancers or not) in relation to colorectal cancer incidence subclassified by *pks*+ *E. coli* levels, while adjusting for potential confounders and selection bias due to tissue availability. In addition, we comprehensively assessed clinical, pathological, molecular, and prognostic features according to the amount of *pks*+ *E. coli* in colorectal carcinoma tissue.

MATERIALS AND METHODS

Study Population and Dietary Assessment

We used two prospective cohort studies in the U.S., namely the Nurses' Health Study (NHS, 121,700 women aged 30-55 years at enrollment in 1976)[12,13] and the Health Professionals Follow-up Study (HPFS, 51,529 men aged 40-75 years at enrollment in 1986)[13,14] (**Figure 1**). Study participants had been followed by use of questionnaires every two years on lifestyle and diagnoses of major diseases. The response rate has exceeded 90% for each follow-up questionnaire cycle in both cohorts. Dietary data were collected using self-administrated semi-quantitative food frequency questionnaires in 1984, 1986, and every four years thereafter in the NHS, and every four years since 1986 in the HPFS. Validity of semiquantitative food frequency questionnaires in the assessment of dietary intake was extensively assessed

and documented in studies using diet records and plasma nutrients[15-17]. Total nutrient intakes were calculated by summing intakes from all foods and adjusted for total energy intake by the residual method. In this study, we used data from 134,775 participants who provided sufficient longitudinal dietary information.

The participants had been followed since the baseline questionnaire return until colorectal cancer diagnosis, loss to follow-up, end of follow-up (June 1, 2014 for the NHS, January 1, 2014 for the HPFS), or death, whichever came first. Participants who had major illnesses including colorectal cancer reported those through questionnaires. Unreported lethal colorectal cancer cases were ascertained through use of the National Death Index. Clinical information such as tumor location and disease stage based on the American Joint Committee on Cancer (AJCC) classification was extracted from medical record by a study physician[18]. We included both colon and rectal carcinomas based on the colorectal continuum model[19]. We gathered formalin-fixed paraffin-embedded (FFPE) tissue blocks from pathology files of hospitals throughout the U.S. where the patients' tumors were resected. Histopathological features including tumor differentiation, extracellular mucin, and signet ring cells were evaluated by the study pathologist (S.O.)[20]. In this study, the inverse probability weighting (IPW) method using both cases with available tissue bacterial data (n = 1,175) and those without tissue bacterial data (n = 2,025) was integrated into duplication-method Cox proportional hazards regression analysis to adjust for selection bias due to tissue bacterial data availability (**Figure 1**). Characteristics of the cases with tissue bacterial data were similar to those without tissue bacterial data (**Table S1**). In addition, during an assay validation step, we utilized tissues from 21 anonymized colorectal cancer patients who

had surgical resections performed at the Brigham and Women's Hospital or Kumamoto University.

Informed consent was obtained from all participants at enrollment and consent for tissue specimen use was additionally obtained before tissue collection. This study protocol was approved by the institutional review boards of the Brigham and Women's Hospital (Boston, MA) and Kumamoto University (Kumamoto, Japan), and those of participating registries as required.

Tumor Tissue Analyses

Genomic DNA was extracted from archival FFPE tissue sections of colorectal carcinoma using the QIAamp DNA FFPE Tissue Kit and GeneRead DNA FFPE Kit (Qiagen, Hilden, Germany). We used custom TaqMan primer-probe sets (Applied Biosystems, Foster City, CA) for the *clbB* gene DNA sequence of *pks*+ *E. coli*[4] and for the reference human gene *SLCO2A1* that has been used in other bacterial assays on FFPE tissue-derived DNA[21] (the names used follow the recommendations for standardized nomenclature of genes and their products by an expert panel[22]). Genomic DNA concentration derived from samples was measured by Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA). Each reaction contained 20 ng of genomic DNA and was assayed in 20 µl reactions containing 1x final concentration TaqMan Environmental Master Mix 2.0 (Applied Biosystems), in a 96-well optical polymerase chain reaction (PCR) plate. Amplification and detection of DNA were performed with a QuantStudio 3 Real-time PCR System (Thermo Fisher Scientific) using the following reaction

conditions; 10 minutes at 95°C and 45 cycles of 15 seconds at 95°C, 30 seconds at 57°C, and 30 seconds at 72°C. The primer and probe sequences for each TaqMan Gene Expression Assay were as follows: *pks*+ *E. coli* forward primer, 5'-GCAACATACTCGCCCAGACT-3'; *pks*+ *E. coli* reverse primer, 5'-TCTCAAGGCGTTGTTGTTTG-3'; *pks*+ *E. coli* FAM probe, 5'-CAAGGTGCGCGCTAGGCTGT-3'; *SLCO2A1* forward primer, 5'-ATCCCCAAAGCACCTGGTTT-3'; *SLCO2A1* reverse primer, 5'-AGAGGCCAAGATAGTCCTGGTAA-3'; *SLCO2A1* VIC probe, 5'-CCATCCATGTCCTCATCTC-3'. To validate our PCR assay, Sanger dideoxy sequencing was performed on the PCR product from three anonymized colorectal carcinoma patients in which the PCR assay detected *pks*+ *E. coli* DNA. The PCR product (165 bp) using the forward and reverse primer sets was isolated by agarose gel electrophoresis. The isolated PCR product was amplified by subcloning and sequenced by Sanger dideoxy sequencing using BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific). We used Competent Quick DH5a (Toyobo, Japan) and QIAprep Spin Miniprep Kit (Qiagen) in the transformation and extraction. We confirmed that the PCR product had sequence of the *clbB* gene of *pks*+ *E. coli* in all of the three cases. In two colorectal carcinoma cases with detectable *pks*+ *E. coli*, the cycle threshold (Ct) values for *pks*+ *E. coli* and *SLCO2A1* decreased linearly with the amount of input DNA (in a log scale) from the same specimens ($r^2 > 0.95$) (**Figure 2A**). We also confirmed that the Ct values (for *pks*+ *E. coli*) decreased linearly with the amount of input DNA (in a log scale) from *pks*+ *E. coli* DNA (American Type Culture Collection, Manassas, VA) ($r^2 > 0.99$) (**Figure 2B**) and that there was no amplification of DNA from

E. coli without the *pks* island (DH10B) (Thermo Fisher Scientific) as a negative control. These positive and negative controls were used in each PCR run on the HPFS and the NHS specimens. Furthermore, in six colorectal carcinoma cases (three positive and three negative cases for *pks*+ *E. coli* DNA), the interassay coefficient of variation of Ct values from each specimen was <1% for both targets in repeated assays of five different batches (**Table S2**). In the cohort cases, each specimen was analyzed in duplicate for each target in a single batch, and we used the mean of the two Ct values for each target. The amount of *pks*+ *E. coli* was calculated as a relative unitless value normalized with *SLCO2A1* using the $2^{-\Delta Ct}$ method (where ΔCt = the average Ct value of *pks*+ *E. coli* – the average Ct value of *SLCO2A1*) as previously described[23]. Cases with detectable *pks*+ *E. coli* were dichotomized into high-level vs. low-level based on the median cut-off point. Microsatellite instability (MSI) status was determined based on PCR of 10 microsatellite markers (D2S123, D5S346, D17S250, BAT25, BAT26, BAT40, D18S55, D18S56, D18S67, and D18S487) as previously described[18]. CpG island methylator phenotype (CIMP) was determined using MethyLight assays[24] of the 8 promoter CpG islands (*CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1*, *RUNX3*, and *SOCS1*)[25]. Methylation level of long-interspersed nucleotide element-1 (LINE-1) was measured using bisulfite PCR and Pyrosequencing as previously described[26]. PCR and pyrosequencing were performed for *KRAS* (codons 12, 13, 61, and 146)[27,28], *BRAF* (codon 600), and *PIK3CA* (exons 9 and 20), as previously described[29].

Statistical Analysis

Detailed statistical analysis methods are described in **Supplementary Methods**. All statistical analyses were performed using SAS software (version 9.4, SAS Institute, Cary, NC), and all *P* values were two-sided. We adjusted the two-sided α level to 0.012 (approximately 0.05/4) for multiple hypothesis testing by Bonferroni correction, considering our use of one heterogeneity trend test (for levels of *pks*+ *E. coli*) and three stratum-specific (high, low, and negative *pks*+ *E. coli*) statistical trend tests.

The western-style diet was derived by principal component analyses of the extensive diet data, as previously described and validated[12,14]. Each participant was assigned a factor score, determined by adding the reported frequencies of food item intakes weighted by the factor loadings (**Table 1**). To capture long-term habitual consumption, we calculated the cumulative mean of the western diet scores from all data-available preceding food frequency questionnaires up to each questionnaire cycle. **Table 1** shows the distribution of western diet score in each cohort.

To limit the number of primary hypotheses, our primary hypothesis testing was the assessment of heterogeneity of the association of the western diet score with the incidence of colorectal cancer subclassified by tissue bacterial amount. We examined heterogeneity across the ordinal tumor subtypes (by the one degree-of-freedom statistical trend test for negative vs. low vs. high) in the multivariable-adjusted duplication-method Cox proportional hazards model using the meta-regression method with a subtype-specific random effect term[30]. For statistical trend tests, the diet score was used as a continuous variable with cohort-specific ceilings at the 10th and 90th

percentiles to eliminate outlier effects. We also examined hazard ratios for each cancer subgroup by comparing dietary score tertiles as secondary analyses. To control for selection bias due to tissue bacterial data availability in the 1,175 cases, we used the 3,200 incident colorectal cancer cases and inverse probability weighting (IPW) method[31] combined with Cox proportional hazards regression models. Multivariable Cox regression models were stratified by age, sex (cohort), and questionnaire year and additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), pack-years smoked (continuous with 50 pack-years ceiling), first-degree relative family history of colorectal cancer (yes vs. no), previous colonoscopy/sigmoidoscopy (yes vs. no), physical activity (continuous with 50 metabolic-equivalent-task-score hours/week ceiling), aspirin or nonsteroidal anti-inflammatory drug use (≥2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling). In the NHS (female)-only analyses, we additionally adjusted for postmenopausal hormone use (yes vs. no). For individuals with missing data in one questionnaire, data from preceding questionnaires were used.

In secondary analyses to assess clinical, pathological, and molecular features according to *pks*+ *E. coli* status (negative, low, and high), we used the chi-square test (for categorical variables), an analysis of variance (for continuous variables), or Spearman correlation analysis (for ordinal variables). In secondary analyses to assess patient survival, we used IPW-adjusted Kaplan-Meier analysis and multivariable Cox proportional hazards regression models (see details in **Supplementary Methods**). In secondary analyses of the association of red meat variables (total, unprocessed, and processed red meat intake) with the incidence of colorectal cancer subclassified by

pks+ *E. coli* status, we examined heterogeneity across the ordinal tumor subtypes (by the one degree-of-freedom statistical trend test for *pks*+ *E. coli* negative vs. low vs. high) in the multivariable-adjusted duplication-method Cox proportional hazards model using the meta-regression method with a subtype-specific random effect term.

RESULTS

We utilized data from 134,775 participants of the Health Professionals Follow-up Study (HPFS) and the Nurses' Health Study (NHS) (**Table 2** and **Figure 1**). During 3,766,179 person-years of follow-up, we documented 3,200 incident colorectal cancer cases. In multivariable analyses using each cohort, the western diet score was weakly associated with colorectal cancer incidence (**Table S3**). Because the results were similar in the two cohorts (P for heterogeneity > 0.6), we combined the two cohorts for further analyses to maximize statistical power while adjusting for cohort (i.e., sex).

We developed and validated the assay to quantify *pks*+ *E. coli* DNA in tumor tissue. The assay, which was successfully conducted in duplicate in 1,175 cases among the 3,200 colorectal cancer patients, detected *pks*+ *E. coli* in 111 patients while 1,064 patients were negative for this bacterium. Clinical, pathological, and molecular features according to the amount of *pks*+ *E. coli* in colorectal carcinoma tissue are shown in **Table 3**. The amount of *pks*+ *E. coli* DNA was inversely associated with AJCC stage (P = 0.008) but not with the other features examined.

We examined the association of the western diet score with colorectal cancer incidence, using all 3,200 incident cases, the 1,175 cases with bacterial data, and the remaining 2,025 cases without bacterial data (**Table S4**). There was no substantial difference in the results from these three analyses. To adjust for selection bias due to bacterial data availability, we used the inverse probability weighting (IPW) method[31] on the 3,200 cases for further analyses.

Our analysis showed that the association of western diet scores with colorectal cancer incidence differed by tissue *pks*+ *E. coli* levels (*P* for heterogeneity = 0.014; **Table 4**), and was stronger for tumors containing higher-level *pks*+ *E. coli*. Multivariable hazard ratios in individuals with scores in the highest (vs. the lowest) tertile of western diet scores were 3.45 [95% confidence interval (CI), 1.53-7.78; *P* for trend = 0.001 across the tertiles] for colorectal cancer with high-level *pks*+ *E. coli*, 1.22 (95% CI, 0.57-2.63) for cancer with low-level *pks*+ *E. coli*, and 1.10 (95% CI, 0.85-1.42) for cancer without detectable *pks*+ *E. coli*. In a sensitivity analysis, we confirmed that the analysis without IPW yielded results (**Table S5**) similar to the IPW-adjusted analysis.

In secondary subgroup analyses, we found similar differential associations by *pks*+ *E. coli* status in men and women (**Table 5**). In analyses using patients stratified by tumor microsatellite instability (MSI) status (**Table S6**), the differential association by *pks*+ *E. coli* status was apparent for the non-MSI-high subtype while statistical power was limited for the MSI-high subtype.

In secondary analyses to assess the prognostic association of the amount of *pks*+ *E. coli*, we conducted survival analysis using Kaplan-Meier method and Cox proportional hazard regression (**Figure S1** and **Table S7**). In univariable analyses of

colorectal cancer-specific survival, there was a statistically-insignificant favorable prognostic association of the amount of *pks*+ *E. coli* (P_{trend} = 0.028, with the α level of 0.012), which did not persist in multivariable analyses (P_{trend} > 0.16).

We further examined whether red meat variables (total, unprocessed, and processed red meat intake amounts), which was the largest component of the western diet score, might be differentially associated with colorectal cancer by *pks*+ *E. coli* status (**Table S8**). We found that the differential association by the amount of *pks*+ *E. coli* was not statistically significant for any of these red meat variables ($P_{heterogeneity}$ > 0.05).

DISCUSSION

Colorectal cancer is a heterogeneous group of neoplastic diseases influenced by many factors including diet, lifestyle, and intestinal microbiota[32-37]. Using two prospective cohort studies in the U.S. with three decades of follow-up, we discovered a stronger association of western-style diet with the incidence of colorectal carcinoma containing higher amounts of *pks*+ *E. coli*. Our findings provide evidence for western-style diet characterized by high intake of red and processed meat, sugar, and refined grains as a risk factor for colorectal cancer, especially its subtype containing a high amount of *pks*+ *E. coli*. Our novel data can inform research efforts devoted to developing cancer prevention strategies that modify diet and the intestinal microbiome.

Previous metagenomic studies have cast light on the role of the intestinal microbiome in colorectal carcinogenesis[38-40]. Molecular pathological analyses of

colorectal cancer have also supported the role of specific intestinal microbes such as colibactin-producing *pks*+ *E. coli* in tumor development[41,42]. A recent study has elucidated the structure of colibactin and enabled the synthesis of colibactin[6]. Experimental studies indicate that the genotoxic colibactin can alkylate DNA on adenine residues[7] and induce double-strand breaks, leading to a specific mutational signature[8,9]. Another study has shown that organoids that recovered from short-term infection with *pks*+ *E. coli* reveal characteristics of colorectal carcinoma cells, such as enhanced proliferation, WNT-independence, and impaired differentiation, at least in part through alterations in TP53-signaling[43]. In addition, evidence suggests that *pks*+ *E. coli* suppresses the host immune response in the tumor microenvironment[44]. Taken together, although *pks*+ *E. coli* likely plays a role in colorectal carcinogenesis, it currently remains uncertain when and how *pks*+ *E. coli* exerts an effect on tumor development. Investigating the detailed mechanism and the associations of this bacterium with lifestyle and dietary risk factors is of particular interest.

Dietary influences on the microbiome in stool and colonic tissue have been investigated. Experimental studies have shown that daily microbiome variation is related to food group choices[45], and that a high-fat diet can alter intestinal bacterial composition[11] and lead to the development of systemic inflammation[46,47]. Observational studies have found relationships between low quality diet and inflammatory diet with intestinal dysbiosis[48,49] as well as between western-style diet and high plasma soluble CD14 level, a biomarker of mucosal barrier dysfunction[50]. These lines of evidence suggest that dietary factors can influence intestinal microbial composition and inflammatory status. The prior principal component analysis on diet data in the

population revealed two dominant dietary patterns, namely the western-style pattern and prudent dietary pattern[51]. A meta-analysis indicates a weak-to-moderate association between western-style diet and colorectal cancer risk[10]. In contrast, the prudent dietary pattern characterized by high intake of fruits, vegetables, fish, poultry, and whole-grains have been inversely associated with colorectal cancer risk[10]. Nonetheless, the strength of the association remains uncertain due to residual or unmeasured confounding by other healthy or unhealthy behaviors associated with the dietary patterns.

Utilizing the molecular pathological epidemiology approach[33,52-54], we found a strong association between western diet and the colorectal cancer subgroup containing high levels of *pks*+ *E. coli.* This specific link between western diet and *pks*+ *E. coli* suggests potentially interactive carcinogenic effects. In further analysis of red meat variable, we did not observe a statistically significant association of any red meat variable with the incidence of colorectal cancer by *pks*+ *E. coli* status. Our data suggest that red meat intake by itself is unlikely the sole factor that contributed to the differential association of the western diet with colorectal cancer by *pks*+ *E. coli* status. One possibility is that western diet may promote the proliferation and activity of *pks*+ *E. coli* and/or strengthen the carcinogenic effects of *pks*+ *E. coli* through alteration of the local tissue microenvironment. It is evident that the molecular pathological epidemiology approach allows for the generation of intriguing hypotheses based on human population data. Although our analyses showed the correlation between western diet and the incidence of colorectal cancer containing high abundance of *pks*+ *E. coli*, a replication using additional independent cohorts and experimental research is necessary.

In addition, we found that the association of western diet with colorectal cancer incidence according to *pks*+ *E. coli* might be different by sex (**Table 5**). Although intriguing, those results were obtained by our secondary subgroup analyses, and as such, generalizability needs to be tested in independent datasets. If replicated, our findings may inform differential interactive influences of western diet and *pks*+ *E. coli* in male vs. female. While the mechanisms underlying these sex-specific effects remain to be elucidated, differences in biological features of colorectal cancer between men and women have been demonstrated[55-57]. Additional studies are warranted to investigate how western diet and *pks*+ *E. coli* may exert interactive carcinogenic effects, and which specific food items might contribute to the observed differential associations between western diet and colorectal cancer incidence according to *pks*+ *E. coli* status.

We acknowledge limitations in the current study. First, unmeasured and/or residual confounding might have substantially influenced our findings. We included most established risk factors in our analysis models with little evidence for substantial confounding by the included variables. Second, tissue bacterial data was unavailable for some incident cancer cases within the cohorts, which might have caused selection bias. However, by using all 3,200 incident colorectal cancers and the inverse probability weighting (IPW) method[31], we were able to adjust for selection bias with the available covariates. Analyses with and without the IPW adjustment yielded similar results. Third, measurement errors were inherently present in the assessments of diet and tissue bacterial amounts, particularly with the use of FFPE tissue specimens. We utilized repeated assessments of diet every four years, which allowed us to estimate the effects of long-term dietary patterns. For bacterial analyses, we carefully optimized and

validated our quantitative PCR assay for FFPE tissue specimens, to ensure high analytical sensitivity and specificity. Our validation study also demonstrated a high linearity ($r^2 > 0.95$) and high precision (with <1% interassay coefficient of variation) of the assay. Fourth, our cohort populations mainly consisted of non-Hispanic Whites, and thus our findings need to be replicated in independent populations. Fifth, we utilized information on microbial contents in tumor tissue, which was not prospectively collected unlike dietary data. Therefore, establishing a cause-and-effect relationship between the microbial species and colorectal cancer requires additional studies. Finally, our findings were based on the observational cohort studies which had certain inherent limitations in data collections. Hence, additional epidemiological studies and experimental confirmation are ultimately needed.

There exist notable strengths in the current study. First, our dietary data had been prospectively and repeatedly collected for more than 30 years through validated food frequency questionnaires[58]. Second, our prospective cohort design enabled the collection of diet and other lifestyle data without knowing who would develop colorectal cancer later, thereby eliminating differential recall bias between cancer patients and cancer-free individuals. Third, the prospective study design also enabled us to leverage all 3,200 incident colorectal cancer cases with the IPW method to adjust for selection bias caused by tissue bacterial data availability. Fourth, we utilized molecular pathological epidemiology methods which can provide novel etiological insights into diet and bacterial species, thereby augmenting causal inference. Fifth, the cancer patient group was assembled from hundreds of hospitals located throughout the U.S., which increases the generalizability of our findings in contrast to studies based on only one or

a few hospitals. Nonetheless, our findings should be replicated in independent populations.

In conclusion, we have found that the association of western diet with colorectal cancer incidence is stronger for tumors containing higher amounts of *pks+ E. coli*. Our findings provide evidence supporting the role of the gut microbiota in mediating the pathogenic link between diet and colorectal cancer. This study also underscores the importance of diet as a modifiable factor that may contribute to cancer prevention.

Acknowledgments

We would like to thank the participants and staff of the Health Professionals Follow-up Study and the Nurses' Health Study for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data.

References

1. Nougayrede JP, Homburg S, Taieb F, et al. Escherichia coli induces DNA double-strand breaks in eukaryotic cells. Science 2006;313:848-51.

2. Wu S, Rhee KJ, Albesiano E, et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nat Med 2009;15:1016-22.

3. Arthur JC, Perez-Chanona E, Muhlbauer M, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science 2012;338:120-3.

4. Dejea CM, Fathi P, Craig JM, et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science 2018;359:592-597.

5. Barrett M, Hand CK, Shanahan F, et al. Mutagenesis by Microbe: the Role of the Microbiota in Shaping the Cancer Genome. Trends Cancer 2020;6:277-287.

6. Xue M, Kim CS, Healy AR, et al. Structure elucidation of colibactin and its DNA cross-links. Science 2019;365.

7. Wilson MR, Jiang Y, Villalta PW, et al. The human gut bacterial genotoxin colibactin alkylates DNA. Science 2019;363.

8. **Pleguezuelos-Manzano C**, **Puschhof J**, **Rosendahl Huber A**, et al. Mutational signature in colorectal cancer caused by genotoxic pks(+) E. coli. Nature 2020;580:269-273.

9. Dziubanska-Kusibab PJ, Berger H, Battistini F, et al. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. Nat Med 2020;26:1063-1069.

10. Magalhaes B, Peleteiro B, Lunet N. Dietary patterns and colorectal cancer: systematic review and meta-analysis. Eur J Cancer Prev 2012;21:15-23.

11. O'Keefe SJ, Li JV, Lahti L, et al. Fat, fibre and cancer risk in African Americans and rural Africans. Nat Commun 2015;6:6342.

12. Fung T, Hu FB, Fuchs C, et al. Major dietary patterns and the risk of colorectal cancer in women. Arch Intern Med 2003;163:309-14.

13. **Nishihara R**, **Wu K**, **Lochhead P**, et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. N Engl J Med 2013;369:1095-105.

14. Hu FB, Rimm EB, Stampfer MJ, et al. Prospective study of major dietary patterns and risk of coronary heart disease in men. Am J Clin Nutr 2000;72:912-21.

15. Feskanich D, Rimm EB, Giovannucci EL, et al. Reproducibility and validity of food intake measurements from a semiquantitative food frequency questionnaire. J Am Diet Assoc 1993;93:790-6.

16. Yuan C, Spiegelman D, Rimm EB, et al. Relative Validity of Nutrient Intakes Assessed by Questionnaire, 24-Hour Recalls, and Diet Records as Compared With Urinary Recovery and Plasma Concentration Biomarkers: Findings for Women. Am J Epidemiol 2018;187:1051-1063.

17. Al-Shaar L, Yuan C, Rosner B, et al. Reproducibility and Validity of a Semiquantitative Food Frequency Questionnaire in Men Assessed by Multiple Methods. Am J Epidemiol 2021;190:1122-1132.

18. **Yamauchi M**, **Morikawa T**, **Kuchiba A**, et al. Assessment of colorectal cancer molecular features along bowel subsites challenges the conception of distinct dichotomy of proximal versus distal colorectum. Gut 2012;61:847-54.

19. **Yamauchi M**, **Lochhead P**, **Morikawa T**, et al. Colorectal cancer: a tale of two sides or a continuum? Gut 2012;61:794-7.

20. **Inamura K**, **Yamauchi M**, **Nishihara R**, et al. Prognostic significance and molecular features of signet-ring cell and mucinous components in colorectal carcinoma. Ann Surg Oncol 2015;22:1226-1235.

21. **Mima K**, **Sukawa Y**, **Nishihara R**, et al. Fusobacterium nucleatum and T Cells in Colorectal Carcinoma. JAMA Oncol 2015;1:653-61.

22. **Fujiyoshi K, Bruford EA, Mroz P**, et al. Opinion: Standardizing gene product nomenclature-a call to action. Proc Natl Acad Sci U S A 2021;118.

23. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. Nat Protoc 2008;3:1101-8.

24. Ogino S, Kawasaki T, Brahmandam M, et al. Precision and performance characteristics of bisulfite conversion and real-time PCR (MethyLight) for quantitative DNA methylation analysis. J Mol Diagn 2006;8:209-17.

25. **Nosho K**, **Irahara N**, **Shima K**, et al. Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. PLoS One 2008;3:e3698.

26. **Irahara N**, **Nosho K**, **Baba Y**, et al. Precision of pyrosequencing assay to measure LINE-1 methylation in colon cancer, normal colonic mucosa, and peripheral blood cells. J Mol Diagn 2010;12:177-83.

27. Ogino S, Kawasaki T, Brahmandam M, et al. Sensitive sequencing method for KRAS mutation detection by Pyrosequencing. J Mol Diagn 2005;7:413-21.

28. **Imamura Y**, **Lochhead P**, **Yamauchi M**, et al. Analyses of clinicopathological, molecular, and prognostic associations of KRAS codon 61 and codon 146 mutations in colorectal cancer: cohort study and literature review. Mol Cancer 2014;13:135.

29. **Liao X**, **Lochhead P**, **Nishihara R**, et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. N Engl J Med 2012;367:1596-606.

30. Wang M, Spiegelman D, Kuchiba A, et al. Statistical methods for studying disease subtype heterogeneity. Stat Med 2016;35:782-800.

31. **Liu L**, **Nevo D**, **Nishihara R**, et al. Utility of inverse probability weighting in molecular pathological epidemiology. Eur J Epidemiol 2018;33:381-392.

32. **Hamada T**, **Nowak JA**, **Milner DA, Jr.**, et al. Integration of microbiology, molecular pathology, and epidemiology: a new paradigm to explore the pathogenesis of microbiome-driven neoplasms. J Pathol 2019;247:615-628.

33. **Akimoto N**, **Ugai T**, **Zhong R**, et al. Rising incidence of early-onset colorectal cancer - a call to action. Nat Rev Clin Oncol 2021;18:230-243.

34. Gunter MJ, Alhomoud S, Arnold M, et al. Meeting report from the joint IARC-NCI international cancer seminar series: a focus on colorectal cancer. Ann Oncol 2019;30:510-519.

35. Tilg H, Adolph TE, Gerner RR, et al. The Intestinal Microbiota in Colorectal Cancer. Cancer Cell 2018;33:954-964.

36. Rajpoot M, Sharma AK, Sharma A, et al. Understanding the microbiome: Emerging biomarkers for exploiting the microbiota for personalized medicine against cancer. Semin Cancer Biol 2018;52:1-8.

37. Li J, Zhang AH, Wu FF, et al. Alterations in the Gut Microbiota and Their Metabolites in Colorectal Cancer: Recent Progress and Future Prospects. Front Oncol 2022;12:841552.

38. Yachida S, Mizutani S, Shiroma H, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 2019;25:968-976.

39. Thomas AM, Manghi P, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:667-678.

40. Ternes D, Karta J, Tsenkova M, et al. Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? Trends Microbiol 2020;28:401-423.

41. Thakur BK, Malaise Y, Martin A. Unveiling the Mutational Mechanism of the Bacterial Genotoxin Colibactin in Colorectal Cancer. Mol Cell 2019;74:227-229.

42. Arthur JC. Microbiota and colorectal cancer: colibactin makes its mark. Nat Rev Gastroenterol Hepatol 2020;17:317-318.

43. Iftekhar A, Berger H, Bouznad N, et al. Genomic aberrations after short-term exposure to colibactin-producing E. coli transform primary colon epithelial cells. Nat Commun 2021;12:1003.

44. Lopes A, Billard E, Casse AH, et al. Colibactin-positive Escherichia coli induce a procarcinogenic immune environment leading to immunotherapy resistance in colorectal cancer. Int J Cancer 2020;146:3147-3159.

45. Johnson AJ, Vangay P, Al-Ghalith GA, et al. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. Cell Host Microbe 2019;25:789-802 e5.

46. Wan Y, Wang F, Yuan J, et al. Effects of dietary fat on gut microbiota and faecal metabolites, and their relationship with cardiometabolic risk factors: a 6-month randomised controlled-feeding trial. Gut 2019;68:1417-1429.

47. Groschel C, Prinz-Wohlgenannt M, Mesteri I, et al. Switching to a Healthy Diet Prevents the Detrimental Effects of Western Diet in a Colitis-Associated Colorectal Cancer Model. Nutrients 2019;12.

48. Liu Y, Ajami NJ, El-Serag HB, et al. Dietary quality and the colonic mucosa-associated gut microbiome in humans. Am J Clin Nutr 2019;110:701-712.

49. Zheng J, Hoffman KL, Chen JS, et al. Dietary inflammatory potential in relation to the gut microbiome: results from a cross-sectional study. Br J Nutr 2020;124:931-942.

50. Tabung FK, Birmann BM, Epstein MM, et al. Influence of Dietary Patterns on Plasma Soluble CD14, a Surrogate Marker of Gut Barrier Dysfunction. Curr Dev Nutr 2017;1.

51. **Mehta RS**, **Nishihara R**, **Cao Y**, et al. Association of Dietary Patterns With Risk of Colorectal Cancer Subtypes Classified by Fusobacterium nucleatum in Tumor Tissue. JAMA Oncol 2017;3:921-927.

52. Ogino S, Chan AT, Fuchs CS, et al. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. Gut 2011;60:397-411.

53. Ogino S, Nowak JA, Hamada T, et al. Insights into Pathogenic Interactions Among Environment, Host, and Tumor at the Crossroads of Molecular Pathology and Epidemiology. Annu Rev Pathol 2019;14:83-103.

54. **Mima K**, **Kosumi K**, **Baba Y**, et al. The microbiome, genetics, and gastrointestinal neoplasms: the evolving field of molecular pathological epidemiology to analyze the tumor-immune-microbiome interaction. Hum Genet 2021;140:725-746.

55. Lopes-Ramos CM, Kuijjer ML, Ogino S, et al. Gene Regulatory Network Analysis Identifies Sex-Linked Differences in Colon Cancer Drug Metabolism. Cancer Res 2018;78:5538-5547.

56. Berntsson J, Eberhard J, Nodin B, et al. Pre-diagnostic anthropometry, sex, and risk of colorectal cancer according to tumor immune cell composition. Oncoimmunology 2019;8:e1664275.

57. Abancens M, Bustos V, Harvey H, et al. Sexual Dimorphism in Colon Cancer. Front Oncol 2020;10:607909.

58. Rimm EB, Giovannucci EL, Stampfer MJ, et al. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. Am J Epidemiol 1992;135:1114-26; discussion 1127-36.

Author names in bold designate shared co-first authorship.

Figure Legends

Figure 1. Flow diagram of the study population in the Nurses' Health Study and the Health Professionals Follow-up Study.

Abbreviation: HPFS, Health Professionals Follow-up Study; NHS, Nurses' Health Study.

Figure 2. Assessment of linearity in quantitative real-time PCR assay. (A): Quantitative real-time PCR assay for *pks*+ *E. coli* DNA and the human reference gene *SLCO2A1* using 2-fold dilution series (10, 20, 40, and 80 ng) from the same DNA specimen from formalin-fixed paraffin-embedded tissue. (B): Quantitative real-time PCR assay for *pks*+ *E. coli* DNA using 10-fold dilution series (0.0001, 0.001, 0.01, 0.1, and 1 ng) from DNA from cultured *pks*+ *E. coli*. Symbols indicate mean, error bars, standard deviation of cycle threshold values of quadruplicate runs. The coefficient of determination (r^2) in the assays for *pks*+ *E. coli* DNA and *SLCO2A1* is shown.

Abbreviation: PCR, polymerase chain reaction.

Table 1. Distribution of Western Diet Scores and Factor Loading Matrix for Western-style Diet in Each Cohort

	Health Professionals Follow-up Study	Nurses' Health Study
Distribution (percentile)		
Minimum	-3.58	-3.98
1%	-1.51	-1.45
5%	-1.17	-1.06
10%	-0.97	-0.85
25%	-0.60	-0.47
50%	-0.10	0.015
75%	0.49	0.56
90%	1.14	1.12
95%	1.59	1.50
99%	2.52	2.31
Maximum	10.0	9.66
Food item*		
Unprocessed red meat	0.66	0.61
Processed meat	0.61	0.58
High fat dairy food	0.51	0.50
French fries	0.49	0.46
Eggs	0.47	0.41
Desserts§	0.43	0.45
Condiments‖	0.39	0.36
Refined grains	0.38	0.38
Butter	0.38	0.50
Mayonnaise	0.36	0.34
Margarine	0.34	0.32
Snacks**	0.34	
Pizza	0.33	0.36
Creamy soups	0.31	0.32
Sugar-sweetened beverages	0.31	0.33
Potatoes		0.34

* Only items with correlation coefficients >0.30 are presented. With the orthogonal rotation used, correlations are identical to factor loading matrix.

§ Desserts include chocolate, candy bars, cookies, brownies, cake, pie, and pastries.

‖ Conditions include soy sauce, non-dairy creamer, Worcestershire sauce, red chili sauce, and pepper.

** Snacks include chips, popcorn, and crackers.

Table 2. Age-standardized Characteristics According to Western Diet Score Tertiles in the Health Professionals Follow-up Study (Men, 1986-2014) and the Nurses' Health Study (Women, 1980-2014)

	Health Professionals Follow-up Study			Nurses' Health Study		
	Western diet score			Western diet score		
Characteristic*	Tertile 1	Tertile 2	Tertile 3	Tertile 1	Tertile 2	Tertile 3
Participants, No.	17,429	14,217	15,803	27,645	27,702	31,979
Mean age, years	65.1	64.7	63.3	62.8	61.3	60.0
Mean body mass index, kg/m^2	25.3	25.9	26.3	24.9	25.3	25.9
Mean physical activity, METS-hours/week†	29.6	27.2	26.2	12.4	10.5	9.3
Mean pack-years smoked	9.0	11.3	15.2	11.6	12.5	13.8
Family history of colorectal cancer, %	15.1	14.7	14.5	19.1	18.8	18.9
History of previous endoscopy, %	25.1	24.9	22.6	28.6	27.6	26.3
Current multivitamin use, %	48.5	47.8	44.3	55.1	52.3	48.2
Regular aspirin or NSAID use, %‡	43.7	48.4	49.4	57.7	60.6	60.8
Postmenopausal, %				74.8	74.0	72.7
Current hormone use, %§				47.1	44.8	42.4
Dietary intake (means)						
Total calorie intake, kcal/day	1,610	1,900	2,420	1,430	1,620	1,990
Unprocessed red meat, servings/day	0.30	0.56	0.88	0.45	0.63	0.82
Processed red meat, servings/day	0.12	0.28	0.56	0.15	0.28	0.47
Poultry, servings/day	0.40	0.39	0.39	0.34	0.31	0.31
Fruit, servings/day	2.82	2.39	2.25	2.49	2.18	2.07
Vegetable, servings/day	3.31	3.15	3.26	2.88	2.62	2.63
Alcohol, g/day	8.0	11.2	13.9	6.5	6.1	5.5
Folate, µg/day	619	545	481	494	424	371
Calcium, mg/day	1,030	928	860	1,070	927	819
Vitamin D, IU/day	515	428	368	429	355	299
Dietary fiber, g/day	25.7	21.5	19.1	18.9	16.1	14.3

* Updated information throughout follow-up was used to calculate the mean for continuous variables and percentage for categorical variables. All variables are age-standardized except age.

† Physical activity is represented by the product sum of the metabolic equivalent task score (METS) of each specific recreational activity and hours spent on that activity per week.

‡ Regular users are defined as ≥2 standard (325-mg) tablets of aspirin or ≥ 2 tablets of NSAIDs per week.

§ Proportion of current menopausal hormone use is calculated among postmenopausal women only.

Abbreviations: METS, metabolic equivalent task score; NSAIDs, non-steroidal anti-inflammatory drugs.

Table 3. Clinical, Pathological, and Molecular Characteristics of Colorectal Cancer Cases According to the Amount of *pks*+ *E. coli* DNA in Colorectal Cancer Tissue

Characteristic*	All cases (N = 1175)	Amount of *pks*+ *E. coli* DNA in colorectal cancer tissue			*P* value†
		Negative (N = 1064)	Low (N = 55)	High (N = 56)	
Sex					0.28
Female (NHS)	656 (56%)	588 (55%)	31 (56%)	37 (66%)	
Male (HPFS)	519 (44%)	476 (45%)	24 (44%)	19 (34%)	
Mean age ± SD (years)	69.0 ± 8.8	68.9 ± 8.8	69.6 ± 10.0	69.9 ± 8.1	0.61
Year of diagnosis					0.09
1995 or before	399 (34%)	371 (35%)	12 (22%)	16 (29%)	
1996-2000	375 (32%)	333 (31%)	18 (33%)	24 (43%)	
2001-2008	401 (34%)	360 (34%)	25 (45%)	16 (29%)	
Family history of colorectal cancer in first-degree relative(s)					0.17
Absent	935 (80%)	854 (81%)	42 (76%)	39 (71%)	
Present	235 (20%)	206 (19%)	13 (24%)	16 (29%)	
Tumor location					0.70
Cecum	202 (17%)	182 (17%)	11 (20%)	9 (16%)	
Ascending to transverse	364 (31%)	334 (32%)	16 (29%)	14 (25%)	
Descending to sigmoid	355 (30%)	314 (30%)	19 (35%)	22 (39%)	
Rectum	249 (21%)	229 (22%)	9 (16%)	11 (20%)	
AJCC disease stage					0.008
I	262 (24%)	230 (23%)	12 (24%)	20 (42%)	
II	354 (33%)	319 (32%)	19 (37%)	16 (33%)	
III	311 (29%)	288 (29%)	14 (27%)	9 (19%)	
IV	157 (14%)	148 (15%)	6 (12%)	3 (6.3%)	
Tumor size ± SD (cm)	4.4 ± 2.0	4.4 ± 2.0	4.7 ± 2.0	4.5 ± 2.1	0.39
Tumor differentiation					0.42
Well to moderate	1053 (90%)	954 (90%)	47 (85%)	52 (93%)	
Poor	118 (10%)	106 (10%)	8 (15%)	4 (7.1%)	
MSI status					0.49
Non-MSI-high	947 (83%)	860 (83%)	40 (78%)	47 (87%)	
MSI-high	188 (17%)	170 (17%)	11 (22%)	7 (13%)	
CIMP status					0.63
Low/negative	885 (82%)	803 (82%)	36 (82%)	46 (87%)	
High	197 (18%)	182 (18%)	8 (18%)	7 (13%)	
Mean LINE-1 methylation level ± SD	63.0 ± 9.8	63.0 ± 9.8	63.4 ± 11.5	63.5 ± 7.2	0.88

KRAS mutation					0.50
Wild-type	645 (59%)	586 (59%)	30 (64%)	29 (53%)	
Mutant	443 (41%)	400 (41%)	17 (36%)	26 (47%)	
BRAF mutation					0.55
Wild-type	942 (84%)	852 (84%)	41 (82%)	49 (89%)	
Mutant	177 (16%)	162 (16%)	9 (18%)	6 (11%)	
PIK3CA mutation					0.51
Wild-type	878 (84%)	795 (84%)	39 (85%)	44 (90%)	
Mutant	168 (16%)	156 (16%)	7 (15%)	5 (10%)	

* Percentage indicates the proportion of patients with a specific clinical, pathological, or molecular characteristic among all patients or in strata of the amount of *pks*+ *E. coli* DNA in colorectal cancer tissue.
† To assess associations between the categories (negative, low, and high) of *pks*+ *E. coli* DNA in colorectal cancer tissue and categorical data, the chi-square test was performed. To compare age, and LINE-1 methylation level, an analysis of variance was performed. To compare AJCC disease stage, Spearman analysis was performed.
Abbreviations: AJCC, American Joint Committee on Cancer; CIMP, CpG island methylator phenotype; HPFS, Health Professionals Follow-up Study; LINE-1, long-interspersed nucleotide element-1; MSI, microsatellite instability; NHS, Nurses' Health Study; SD, standard deviation.

Table 4. Incidence of Colorectal Cancer by *pks+ E. coli* Status in Relation to Cumulative Average Western Diet Score in the Combined Cohorts of the Health Professionals Follow-up Study (1986-2014) and the Nurses' Health Study (1980-2014)

	Western diet score				
	Tertile 1	Tertile 2	Tertile 3	*P* for trend*	*P* for heterogeneity†
Person-years	1,255,030	1,254,558	1,256,591		
Overall colorectal cancer					
Cases, No. (total n=1,175)	392	391	392		
Age-adjusted HR (95% CI)§	1 (referent)	1.04 (0.93-1.15)	1.28 (1.14-1.44)	<0.001	
Multivariable HR (95% CI)‖	1 (referent)	0.98 (0.88-1.09)	1.14 (1.01-1.29)	0.010	
pks+ E. coli status					0.014
pks+ E. coli negative					
Cases, No. (total n=1,064)	364	354	346		
Age-adjusted HR (95% CI)§	1 (referent)	1.00 (0.82-1.21)	1.23 (0.96-1.58)	0.068	
Multivariable HR (95% CI)‖	1 (referent)	0.95 (0.78-1.15)	1.10 (0.85-1.42)	0.40	
pks+ E. coli low					
Cases, No. (total n=55)	18	15	22		
Age-adjusted HR (95% CI)§	1 (referent)	0.82 (0.36-1.83)	1.37 (0.64-2.97)	0.51	
Multivariable HR (95% CI)‖	1 (referent)	0.77 (0.35-1.73)	1.22 (0.57-2.63)	0.76	
pks+ E. coli high					
Cases, No. (total n=56)	10	22	24		
Age-adjusted HR (95% CI)§	1 (referent)	2.24 (1.00-5.04)	3.83 (1.69-8.66)	<0.001	
Multivariable HR (95% CI)‖	1 (referent)	2.11 (0.94-4.73)	3.45 (1.53-7.78)	0.001	

* The trend test was performed using the western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores above 90th percentile and those below the 10th percentile, respectively, to eliminate outlier effects.

† The meta-regression method with a subtype-specific random effect term was used to test whether the association has a trend across the ordinal subtypes in the multivariable-adjusted model, where the western diet score was used as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles.

§ Duplication-method Cox proportional hazards model weighted by inverse probabilities based on tissue bacterial data availability for competing risks data was used with total caloric intake adjusted and stratification by age (in months), sex (i.e., cohort) and year of questionnaire return.

‖ Additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity (continuous with a ceiling at 50 metabolic equivalent task score-hours per week), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling).

Abbreviations: CI, confidence interval; HR, hazard ratio.

Table 5. Incidence of Colorectal Cancer by *pks*+ *E. coli* Status in Relation to Cumulative Average Western Diet Score in the Health Professionals Follow-up Study (1986-2014) and the Nurses' Health Study (1980-2014)

	Western diet score				
	Tertile 1	Tertile 2	Tertile 3	*P* for trend*	*P* for heterogeneity†
Health Professionals Follow-up Study (men)					
Person-Years	365,506	365,602	365,680		
pks+ *E. coli* status					0.71
pks+ *E. coli* negative					
Cases, No. (total n=476)	156	158	162		
Age-adjusted HR (95% CI)§	1 (referent)	1.19 (0.95-1.50)	1.41 (1.09-1.82)	<0.001	
Multivariable HR (95% CI)‖	1 (referent)	1.12 (0.89-1.42)	1.26 (0.96-1.65)	0.015	
pks+ *E. coli* positive					
Cases, No. (total n=43)	10	15	18		
Age-adjusted HR (95% CI)§	1 (referent)	1.77 (0.80-3.95)	2.33 (1.05-5.14)	0.042	
Multivariable HR (95% CI)‖	1 (referent)	1.62 (0.72-3.61)	2.05 (0.93-4.50)	0.10	
Nurses' Health Study (women)					
Person-years	889,524	888,957	890,911		
pks+ *E. coli* status					0.018
pks+ *E. coli* negative					
Cases, No. (total n=588)	208	196	184		
Age-adjusted HR (95% CI)§	1 (referent)	0.95 (0.75-1.20)	1.18 (0.87-1.61)	0.28	
Multivariable HR (95% CI)‖	1 (referent)	0.89 (0.70-1.13)	1.03 (0.75-1.41)	0.90	
pks+ *E. coli* positive					
Cases, No. (total n=68)	18	22	28		
Age-adjusted HR (95% CI)§	1 (referent)	1.17 (0.59-2.31)	2.09 (1.07-4.09)	0.019	
Multivariable HR (95% CI)‖	1 (referent)	1.10 (0.55-2.18)	1.81 (0.92-3.54)	0.058	

* The trend test was performed using the western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores above 90th percentile and those below the 10th percentile, respectively, to eliminate outlier effects.

† The meta-regression method with a subtype-specific random effect term was used to test whether the association has a trend across the ordinal subtypes (negative vs. low vs. high) in the multivariable-adjusted model, where the western diet score was used as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles.

§ Duplication-method Cox proportional hazards model weighted by inverse probabilities based on tissue bacterial data availability for competing risks data was used with total caloric intake adjusted and stratification by age (in months) and year of questionnaire return.

‖ Additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity (continuous with a ceiling at 50 metabolic equivalent task score-hours per week), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling). We additionally adjusted for postmenopausal hormone use (yes vs. no) for the Nurses' Health Study analysis.

Abbreviations: CI, confidence interval; HR, hazard ratio.

Figure 1

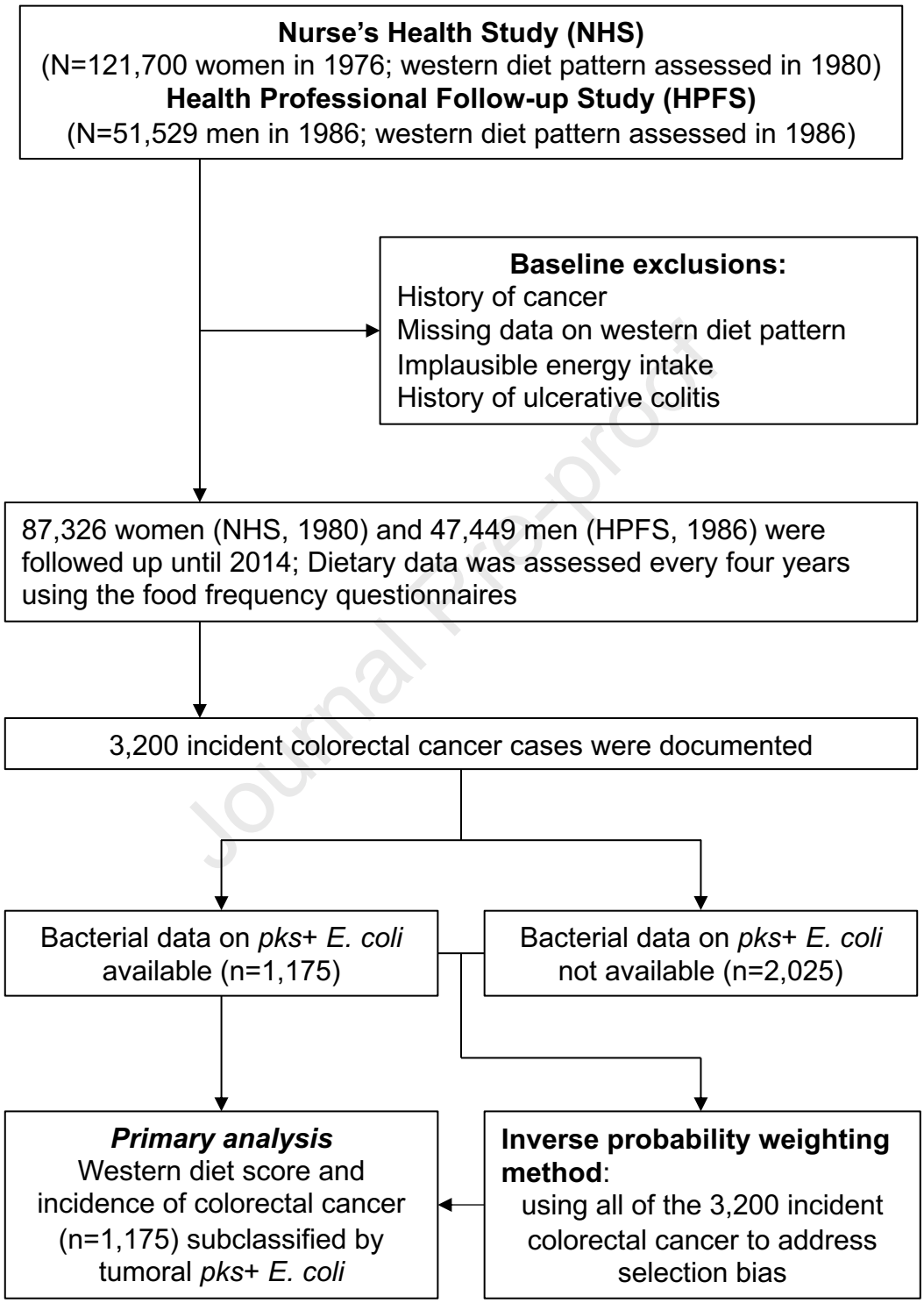
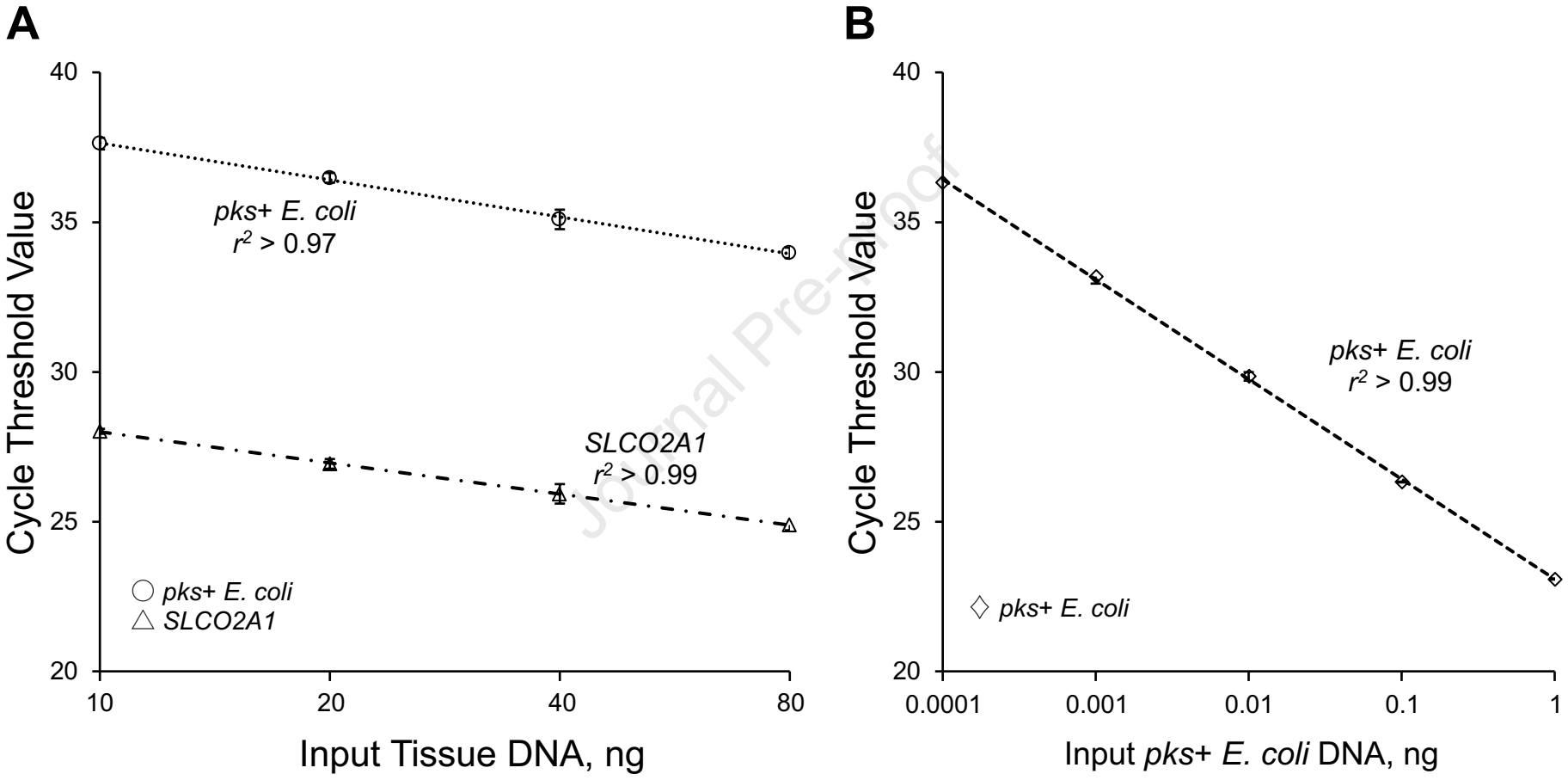


Figure 2.



What You Need to Know and Lay Summary

Western-style Diet, *pks* Island-Carrying *Escherichia coli*, and Colorectal Cancer: Analyses from Two Large Prospective Cohort Studies

Authors
Kota Arima, Rong Zhong, Tomotaka Ugai, Melissa Zhao, Koichiro Haruki, Naohiko Akimoto, Mai Chan Lau, Kazuo Okadome, Raaj S. Mehta, Juha P. Väyrynen, Junko Kishikawa, Tyler S. Twombly, Shanshan Shi, Kenji Fujiyoshi, Keisuke Kosumi, Yoko Ogata, Hideo Baba, Fenglei Wang, Kana Wu, Mingyang Song, Xuehong Zhang, Charles S. Fuchs, Cynthia L. Sears, Walter C. Willett, Edward L. Giovannucci, Jeffrey A. Meyerhardt, Wendy S. Garrett, Curtis Huttenhower, Andrew T. Chan, Jonathan A. Nowak, Marios Giannakis, and Shuji Ogino.

What You Need to Know

BACKGROUND AND CONTEXT

Western-style diet has been weakly associated with colorectal cancer risk; however, it remains unclear whether the association of western-style diet with colorectal cancer incidence varies by gut microbe.

NEW FINDINGS

Utilizing two U.S. longitudinal prospective cohort studies, we found that the association of western-style diet with colorectal cancer incidence was stronger for tumors containing higher amounts of *pks*+ *E. coli*.

LIMITATIONS

Our cohorts consisted predominantly of non-Hispanic Whites. Therefore, further studies using other populations are needed, as well as experimental confirmation to investigate the mechanisms.

IMPACT

Our findings provide evidence supporting the role of the specific bacterium in mediating a pathogenic link between diet and colorectal cancer and the importance of diet for cancer prevention.

Lay Summary

We found that western-style diet (rich in red and processed meat and sugar) increased risk of colorectal cancer containing high amounts of specific toxin-producing *E. coli* bacterium.

Supplementary Appendix

Western-style Diet, *pks* Island-Carrying *Escherichia coli*, and Colorectal Cancer: Analyses from Two Large Prospective Cohort Studies

Authors
Kota Arima, Rong Zhong, Tomotaka Ugai, Melissa Zhao, Koichiro Haruki, Naohiko Akimoto, Mai Chan Lau, Kazuo Okadome, Raaj S. Mehta, Juha P. Väyrynen, Junko Kishikawa, Tyler S. Twombly, Shanshan Shi, Kenji Fujiyoshi, Keisuke Kosumi, Yoko Ogata, Hideo Baba, Fenglei Wang, Kana Wu, Mingyang Song, Xuehong Zhang, Charles S. Fuchs, Cynthia L. Sears, Walter C. Willett, Edward L. Giovannucci, Jeffrey A. Meyerhardt, Wendy S. Garrett, Curtis Huttenhower, Andrew T. Chan, Jonathan A. Nowak, Marios Giannakis, and Shuji Ogino.

Supplementary Methods

Statistical Analysis

All statistical analyses were performed using SAS software (version 9.4, SAS Institute, Cary, NC), and all *P* values were two-sided. We adjusted the two-sided α level to 0.012 (approximately 0.05/4) for multiple hypothesis testing by Bonferroni correction, considering one heterogeneity trend test (for levels of *pks*+ *E. coli*) and the three stratum-specific (high, low, and negative *pks*+ *E. coli*) statistical trend tests.

The western-style diet was derived by principal component analyses of the extensive diet data in the HPFS and the NHS, as previously described and validated[1,2]. Each participant was assigned a factor score, determined by adding the reported frequencies of food item intakes weighted by the factor loadings (**Table 1**). These factor scores were standardized to have a mean of 0 with standard deviation (SD) of 1. To capture long-term habitual consumption, we calculated the cumulative mean of the western diet scores from all data-available preceding food frequency questionnaires up to each questionnaire cycle. **Table 1** shows the distribution of western diet score in each cohort.

To limit the number of primary hypotheses, our primary hypothesis testing was the assessment of heterogeneity of the association of the western dietary score (continuous values) with the incidence of colorectal cancer subclassified by tissue bacterial amount (one degree-of-freedom statistical trend test for negative vs. low vs. high). All other assessments including examinations of hazard ratio (HR) for each cancer subgroup comparing cohort-specific diet score tertile categories were secondary analyses. The statistical trend test was performed using the western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores above 90th percentile and those below 10th percentile, respectively, to eliminate outlier effects. We examined heterogeneity across the ordinal subtypes (by the one degree-of-freedom statistical trend test) in the multivariable-adjusted model by using the meta-regression method with a subtype-specific random effect term[3,4].

All incidence analyses were adjusted for total caloric intake (kilocalories per day) and stratified by age (in months), year of questionnaire return, and sex (in the analyses using combined cohorts). In multivariable analyses, we further adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity [continuous with 50 metabolic equivalent task score (METS)-hours per week ceiling], and regular use of aspirin or nonsteroidal anti-inflammatory drugs (more than two tablets per week: yes vs. no), multivitamin use (yes vs. no), alcohol consumption (continuous with 30 g/day ceiling). In the NHS (female)-only analyses, we additionally adjusted for postmenopausal hormone use (yes vs. no). In secondary analyses for the three red meat variables, we used continuous intake amounts with cohort-specific ceiling at 90th percentiles and mutually adjusted for each other in the multivariable regression model. For individuals with missing data in one questionnaire, data from preceding questionnaires were used. Any remaining individuals with missing data were assigned with a median value (for a continuous variable) or a majority category (for a categorical variable). We assessed physical activity in METS-hours/week beginning in 1986 and each activity on the questionnaire was assigned a score (METS), as previously described[5]. METS in each activity is defined as the ratio of the metabolic rate associated with the activity divided by the resting metabolic rate, and the values from individual activities were summed for total METS-hours per week.

To control for selection bias due to varying tissue data availability in the 1,175 cases, we used the 3,200 incident colorectal cancer cases and inverse probability weighting (IPW) method[6,7] combined with multivariable duplication-method Cox proportional hazards regression models that calculated the hazard ratio (HR) for colorectal cancer incidence. In the IPW method, we constructed a logistic regression model to predict the availability of tumor tissue bacterial data as an outcome variable, using the dataset of 3,200 colorectal cancer cases and following variables; sex (female vs. male), age at diagnosis (continuous; a linear term and a squared term), year of diagnosis (continuous; a linear term and a squared term), family history of colorectal cancer [present vs. absent (including <5% patients with missing data)], prediagnosis body mass index [<25 kg/m^2 vs. 25 to 29.9 kg/m^2 (including patients with missing data) vs. ≥30 kg/m^2], prediagnosis physical activity [continuous; a linear term and a squared term; assigning missing data with a median value; with a separate indicator variable for the missing data (yes vs. no)], prediagnosis regular aspirin use [yes vs. no (including <5% patients with missing data)], prediagnosis pack-years of smoking (continuous; a linear term and a squared term; assigning <5% missing data with a median value), prediagnosis total alcohol intake (continuous; a linear term and a squared term; assigning <5% missing data with a median value), prediagnosis history of endoscopy (yes vs. no vs. missing), tumor location (proximal colon vs. distal colon vs. rectum vs. missing), tumor differentiation (well to moderate vs. poor vs. unspecified), and disease stage [I vs. II vs. III vs. IV (including patients with missing data)]. Prediagnosis data were obtained from the last data-available questionnaire prior to colorectal cancer diagnosis. For individuals with missing data in one questionnaire, data from preceding questionnaires were used. Any remaining

individuals with missing data were treated as mentioned in the preceding paragraph. We used a backward elimination with a threshold *P* value of 0.1 to select variables for the final model. After the selection procedure, the final logistic regression model included sex (female vs. male), year of diagnosis (a linear term and a squared term), pack-years of smoking (a squared term), tumor location (proximal colon vs. distal colon vs. rectum vs. missing), and disease stage [I vs. II vs. III vs. IV (including patients with missing data)]. The final model could calculate a probability value (for tissue bacterial data availability) based on those variables in each patient. Each patient with available tissue bacterial data was weighted by the inverse probability (i.e., one divided by the predictive probability of microbial data availability). Weights greater than the 95th percentile were converted to the 95th percentile weight to eliminate outlier effects[6]. We confirmed that results without this weight truncation did not change substantially (data not shown). Multivariable duplication-method Cox proportional hazard regression analysis without IPW yielded similar results to the IPW-adjusted analysis (**Table S5**).

The Kaplan-Meier analysis and log-rank test for trend that integrated the inverse probability weighting (IPW) method were conducted to compare patient mortality according to the amount of *pks*+ *E. coli*. For analyses of colorectal cancer-specific mortality, deaths as a result from other causes were censored. To control for potential confounders, we used multivariable, IPW-adjusted Cox proportional hazards regression model, which initially included following variables: sex (female vs. male), age at diagnosis (continuous; a linear term and a squared term), year of diagnosis (continuous; a linear term and a squared term), family history of colorectal cancer [present vs. absent (including <5% patients with missing data)], tumor location (proximal colon vs. distal colon vs. rectum vs. missing), tumor differentiation (well to moderate vs. poor vs. unspecified), and disease stage [I/II vs. III/IV (including patients with missing data)]. A backward elimination was conducted with a threshold *P* of 0.05 to select variables for the final models.

References

1. Hu FB, Rimm EB, Stampfer MJ, Ascherio A, Spiegelman D, Willett WC. Prospective study of major dietary patterns and risk of coronary heart disease in men. Am J Clin Nutr 2000;72:912-21.

2. Fung T, Hu FB, Fuchs C, et al. Major dietary patterns and the risk of colorectal cancer in women. Arch Intern Med 2003;163:309-14.

3. Wang M, Spiegelman D, Kuchiba A, et al. Statistical methods for studying disease subtype heterogeneity. Stat Med 2016;35:782-800.

4. Lunn M, McNeil D. Applying Cox regression to competing risks. Biometrics 1995;51:524-32.

5. Meyerhardt JA, Giovannucci EL, Holmes MD, et al. Physical activity and survival after colorectal cancer diagnosis. J Clin Oncol 2006;24:3527-34.

6. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013;22:278-95.

7. **Liu L**, **Nevo D**, **Nishihara R**, et al. Utility of inverse probability weighting in molecular pathological epidemiology. Eur J Epidemiol 2018;33:381-92.

Author names in bold designate shared co-first authorship.

Supplementary Figure Legend

Figure S1. Inverse probability weighting-adjusted Kaplan-Meier survival analyses of colorectal cancer patients according to the amount of *pks*+ *E. coli* DNA in tumor tissue. The *P* values were calculated using the weighted log-rank test for trend (two sided). A: Colorectal cancer-specific survival. B: Overall survival. C: The number of patients who remained alive and at risk of death at each time point after the diagnosis of colorectal cancer.

Table S1. Age-standardized Characteristics of Colorectal Cancer Patients According to Availability of *pks+ E. coli* Bacterial Data in the Health Professionals Follow-up Study and the Nurses' Health Study

	Health Professionals Follow-up Study		Nurses' Health Study	
Characteristic*	Bacterial data available	Bacterial data unavailable	Bacterial data available	Bacterial data unavailable
Participants, No.	519	877	656	1,148
Mean age, years	69.5	70.4	65.8	65.6
Mean body mass index, kg/m^2	26.1	26.3	26.1	26.1
Mean physical activity, METS-hours/week†	25.6	25.9	11.3	10.8
Mean pack-years smoked	10.8	11.3	16.7	16.5
Family history of colorectal cancer, %	24.4	20.1	27.6	24.3
History of previous endoscopy, %	15.6	17.3	17.9	17.3
Current multivitamin use, %	49.0	46.7	50.7	50.1
Regular aspirin or NSAID use, %‡	48.6	44.7	57.7	58.1
Postmenopausal, %			89.3	88.9
Current hormone use, %§			46.6	33.5
Dietary intake (means)				
Total calorie intake, kcal/day	1,980	1,960	1,710	1,660
Unprocessed red meat, servings/day	0.59	0.57	0.80	0.81
Processed red meat, servings/day	0.35	0.34	0.29	0.31
Poultry, servings/day	0.34	0.39	0.49	0.48
Fruit, servings/day	2.54	2.57	2.34	2.24
Vegetable, servings/day	3.27	3.31	2.93	2.78
Alcohol, g/day	14.0	13.0	6.4	6.5
Folate, µg/day	521	544	421	418
Calcium, mg/day	922	928	930	911
Vitamin D, IU/day	429	434	339	340
Dietary fiber, g/day	22.0	22.3	17.0	16.9

* Updated information throughout follow-up was used to calculate the mean for continuous variables and percentage for categorical variables. All variables are age-standardized except age.

† Physical activity is represented by the product sum of the metabolic equivalent task score (METS) of each specific recreational activity and hours spent on that activity per week.

‡ Regular users are defined as ≥2 standard (325-mg) tablets of aspirin or ≥2 tablets of NSAIDs per week.
§ Proportion of current menopausal hormone use is calculated among postmenopausal women only.
Abbreviations: METS, metabolic equivalent task score; NSAIDs, non-steroidal anti-inflammatory drugs.

.

Table S2. Interassay Coefficients of Variation in Quantitative PCR Assay for *pks*+ *E. coli* DNA and the Human Reference Gene *SLCO2A1*

	Targets in quantitative PCR assay			
	pks+ *E. coli* DNA		*SLCO2A1*	
	Mean cycle threshold ± SD	Interassay coefficient of variation	Mean cycle threshold ± SD	Interassay coefficient of variation
Specimen 1	No amplification*	-	26.4 ± 0.11	0.0041
Specimen 2	No amplification*	-	27.1 ± 0.13	0.0049
Specimen 3	No amplification*	-	28.1 ± 0.13	0.0045
Specimen 4	33.7 ± 0.29	0.0087	27.9 ± 0.21	0.0074
Specimen 5	37.9 ± 0.16	0.0042	23.5 ± 0.09	0.0038
Specimen 6	37.4 ± 0.25	0.0066	25.4 ± 0.05	0.0018
Mean coefficient of variation		0.0065		0.0044

Interassay coefficient of variation of cycle threshold values from the same specimen was assessed by repeating assays in five different batches using six colorectal carcinomas.
* In these specimens, *pks*+ *E. coli* DNA was not detected in any PCR reactions.
Abbreviations: PCR, polymerase chain reaction; SD, standard deviation.

Table S3. Incidence of Colorectal Cancer in Relation to Western Diet Score in the Health Professionals Follow-up Study and the Nurses' Health Study

	Western diet score			
	Tertile 1	Tertile 2	Tertile 3	*P* for trend*
Health Professionals Follow-up Study (men)				
Person-years	365,506	365,602	365,680	
Cases, No. (total n=1,396)	469	450	477	
Age-adjusted HR (95% CI)[§]	1 (referent)	1.10 (0.96-1.26)	1.36 (1.16-1.60)	<0.001
Multivariable HR (95% CI)[‖]	1 (referent)	1.01 (0.88-1.16)	1.15 (0.97-1.37)	0.079
Nurses' Health Study (women)				
Person-years	889,524	888,957	890,911	
Cases, No. (total n=1,804)	653	570	581	
Age-adjusted HR (95% CI)[§]	1 (referent)	1.04 (0.92-1.16)	1.27 (1.11-1.45)	<0.001
Multivariable HR (95% CI)[‖]	1 (referent)	0.98 (0.87-1.10)	1.12 (0.97-1.29)	0.006

* The trend test was performed using the western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores above 90th percentile and those below the 10th percentile, respectively, to eliminate outlier effects.

§ Duplication-method Cox proportional hazards model was used with total caloric intake adjusted and stratification by age (in months), sex (i.e., cohort) and year of questionnaire return.

‖ Additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity (continuous with a ceiling at 50 metabolic equivalent task score-hours per week), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling). We additionally adjusted for postmenopausal hormone use (yes vs. no) for the Nurses' Health Study analysis.

Abbreviations: CI, confidence interval; HR, hazard ratio.

Table S4. Western Diet Score and Incidence of Colorectal Cancer Using All 3,200 Incident Cases, the 1,175 Cases with Bacterial Data, and the 2,025 Cases Without Tissue Bacterial Data

	Western diet score			
	Tertile 1	Tertile 2	Tertile 3	*P* for trend*
Person-years	1,255,030	1,254,558	1,256,591	
All incident colorectal cancer cases				
Cases, No. (total n=3,200)	1,122	1,020	1,058	
Age-adjusted HR (95% CI)§	1 (referent)	1.06 (0.97-1.16)	1.31 (1.18-1.45)	<0.001
Multivariable HR (95% CI)‖	1 (referent)	1.01 (0.92-1.10)	1.16 (1.04-1.29)	<0.001
Colorectal cancer cases with bacterial data				
Cases, No. (total n=1,175)	392	391	392	
Age-adjusted HR (95% CI)§	1 (referent)	1.17 (1.01-1.36)	1.34 (1.13-1.60)	<0.001
Multivariable HR (95% CI)‖	1 (referent)	1.11 (0.96-1.29)	1.21 (1.01-1.44)	0.014
Colorectal cancer cases without bacterial data				
Cases, No. (total n=2,025)	730	629	666	
Age-adjusted HR (95% CI)§	1 (referent)	1.00 (0.90-1.12)	1.29 (1.13-1.46)	<0.001
Multivariable HR (95% CI)‖	1 (referent)	0.95 (0.85-1.06)	1.13 (0.99-1.29)	0.006

* The trend test was performed using the western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores above 90th percentile and those below the 10th percentile, respectively, to eliminate outlier effects.

§ Duplication-method Cox proportional hazards model was used with total caloric intake adjusted and stratification by age (in months), sex (i.e., cohort) and year of questionnaire return.

‖ Additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity (continuous with a ceiling at 50 metabolic equivalent task score-hours per week), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling).

Abbreviations: CI, confidence interval; HR, hazard ratio.

Table S5. Incidence of Colorectal Cancer by *pks+ E. coli* Status in Relation to Cumulative Average Western Diet Score Without Inverse Probability Weighting to Adjust for Availability of *pks+ E. coli* Data

	Western diet score				
	Tertile 1	Tertile 2	Tertile 3	*P* for trend*	*P* for heterogeneity†
Person-years	1,255,030	1,254,558	1,256,591		
Overall colorectal cancer					
Cases, No. (total n=1,175)	392	391	392		
Age-adjusted HR (95% CI)§	1 (referent)	1.17 (1.01-1.36)	1.34 (1.13-1.60)	<0.001	
Multivariable HR (95% CI)‖	1 (referent)	1.11 (0.96-1.29)	1.21 (1.01-1.44)	0.014	
pks+ E. coli status					0.017
pks+ E. coli negative					
Cases, No. (total n=1,064)	364	354	346		
Age-adjusted HR (95% CI)§	1 (referent)	1.14 (0.98-1.33)	1.28 (1.07-1.53)	0.001	
Multivariable HR (95% CI)‖	1 (referent)	1.08 (0.93-1.26)	1.15 (0.95-1.38)	0.060	
pks+ E. coli low					
Cases, No. (total n=55)	18	15	22		
Age-adjusted HR (95% CI)§	1 (referent)	0.97 (0.49-1.93)	1.48 (0.78-2.81)	0.19	
Multivariable HR (95% CI)‖	1 (referent)	0.92 (0.46-1.85)	1.32 (0.70-2.52)	0.35	
pks+ E. coli high					
Cases, No. (total n=56)	10	22	24		
Age-adjusted HR (95% CI)§	1 (referent)	2.60 (1.23-5.53)	3.47 (1.64-7.35)	<0.001	
Multivariable HR (95% CI)‖	1 (referent)	2.45 (1.15-5.22)	3.16 (1.49-6.71)	0.001	

* The trend test was performed using the western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores above 90th percentile and those below the 10th percentile, respectively, to eliminate outlier effects.

† The meta-regression method with a subtype-specific random effect term was used to test whether the association has a trend across the ordinal subtypes in the multivariable-adjusted model, where the western diet score was used as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles.

§ Duplication-method Cox proportional hazards model was used with total caloric intake adjusted and stratification by age (in months), sex (i.e., cohort) and year of questionnaire return.

‖ Additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity (continuous with a ceiling at 50 metabolic equivalent task score-hours per week), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling).

Abbreviations: CI, confidence interval; HR, hazard ratio.

Table S6. Incidence of Colorectal Cancer, Jointly Subclassified by *pks*+ *E. coli* Status and Microsatellite Instability According to Western Diet Score

	Western diet score			
	Tertile 1	Tertile 2	Tertile 3	*P* for trend*
Person-years	1,255,030	1,254,558	1,256,591	
Microsatellite instability (MSI) status				
Non-MSI-high				
pks+ *E. coli* negative				
Cases, No. (total n=840)	282	280	278	
Age-adjusted HR (95% CI)§	1 (referent)	0.97 (0.78-1.20)	1.25 (0.95-1.64)	0.083
Multivariable HR (95% CI)‖	1 (referent)	0.92 (0.74-1.14)	1.12 (0.85-1.48)	0.36
pks+ *E. coli* positive				
Cases, No. (total n=87)	22	26	39	
Age-adjusted HR (95% CI)§	1 (referent)	1.14 (0.61-2.13)	2.33 (1.28-4.26)	0.005
Multivariable HR (95% CI)‖	1 (referent)	1.08 (0.58-2.02)	2.11 (1.16-3.83)	0.012
MSI-high				
pks+ *E. coli* negative				
Cases, No. (total n=166)	62	56	48	
Age-adjusted HR (95% CI)§	1 (referent)	0.99 (0.64-1.52)	1.05 (0.66-1.66)	0.64
Multivariable HR (95% CI)‖	1 (referent)	0.95 (0.62-1.45)	0.95 (0.59-1.52)	0.95
pks+ *E. coli* positive				
Cases, No. (total n=18)	5	8	5	
Age-adjusted HR (95% CI)§	1 (referent)	2.71 (0.89-8.25)	1.74 (0.46-6.55)	0.26
Multivariable HR (95% CI)‖	1 (referent)	2.57 (0.84-7.85)	1.58 (0.42-5.99)	0.36

* The trend test was performed using the western diet score as a continuous variable with cohort-specific ceilings at 10th and 90th percentiles in the regression model. The 90th and 10th percentile values were used for scores above 90th percentile and those below the 10th percentile, respectively, to eliminate outlier effects.

§ Duplication-method Cox proportional hazards model weighted by inverse probabilities based on tissue bacterial data availability for competing risks data was used with total caloric intake adjusted and stratification by age (in months), sex (i.e., cohort) and year of questionnaire return.

‖ Additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity (continuous with a ceiling at 50 metabolic equivalent task score-hours per week), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling).

Abbreviations: CI, confidence interval; HR, hazard ratio; MSI, microsatellite instability.

Table S7. Analyses of Patient Survival in Relation to the Amount of *pks*+ *E. coli* DNA in Colorectal Cancer Tissue Using Inverse Probability Weighting

The amount of *pks*+ *E. coli* DNA	No. of cases	Colorectal cancer-specific survival			Overall survival		
		No. of events	Univariable HR (95% CI)*	Multivariable HR (95% CI)*,†	No. of events	Univariable HR (95% CI)*	Multivariable HR (95% CI)*,†
Negative	1062	347	1 (referent)	1 (referent)	719	1 (referent)	1 (referent)
Low	55	14	0.73 (0.41-1.29)	0.92 (0.53-1.59)	36	0.90 (0.64-1.29)	0.97 (0.70-1.34)
High	55	11	0.55 (0.30-1.00)	0.65 (0.35-1.18)	37	0.75 (0.53-1.08)	0.80 (0.56-1.16)
P_{trend}‡			0.028	0.16		0.096	0.24

* Inverse probability weighting was applied to reduce a bias resulting from the availability of tumor tissue bacterial data.
† The initial multivariable Cox regression model initially included age, sex, year of diagnosis, family history of colorectal cancer, tumor location, AJCC disease stage, tumor differentiation, microsatellite instability, CpG island methylator phenotype, *BRAF*, and *PIK3CA* mutations, and long-interspersed nucleotide element-1 methylation level. A backward elimination with a threshold of $P = 0.05$ was used to select variables for the final model.
‡ P_{trend} value was calculated across the ordinal categories (negative, low, and high) of the amount of *pks*+ *E. coli* DNA in colorectal cancer tissue in the inverse probability weighting adjusted Cox regression model.
Abbreviations: CI, confidence interval; HR, hazard ratio.

Table S8. Incidence of Colorectal Cancer by *pks+ E. coli* Status in Relation to Cumulative Average Total Red Meat, Unprocessed Red Meat, or Processed Red Meat in the Combined Cohorts of the Health Professionals Follow-up Study (1986-2014) and the Nurses' Health Study (1980-2014)

	Total red meat			*P* for trend*	*P* for heterogeneity†
	Tertile 1	Tertile 2	Tertile 3		
Overall colorectal cancer					
Cases, No. (total n=1,175)	377	426	372		
Age-adjusted HR (95% CI)§	1 (referent)	1.17 (1.07-1.28)	1.18 (1.06-1.30)	<0.001	
Multivariable HR (95% CI)‖	1 (referent)	1.10 (1.00-1.20)	1.06 (0.95-1.17)	0.010	
pks+ E. coli status					0.18
pks+ E. coli negative					
Cases, No. (total n=1,064)	351	380	333		
Age-adjusted HR (95% CI)§	1 (referent)	1.12 (0.95-1.32)	1.15 (0.95-1.39)	0.014	
Multivariable HR (95% CI)‖	1 (referent)	1.05 (0.89-1.24)	1.03 (0.85-1.25)	0.23	
pks+ E. coli low					
Cases, No. (total n=55)	13	27	15		
Age-adjusted HR (95% CI)§	1 (referent)	2.00 (0.89-4.49)	1.09 (0.47-2.56)	0.38	
Multivariable HR (95% CI)‖	1 (referent)	1.86 (0.83-4.18)	0.98 (0.42-2.31)	0.67	
pks+ E. coli high					
Cases, No. (total n=56)	13	19	24		
Age-adjusted HR (95% CI)§	1 (referent)	1.67 (0.77-3.59)	2.24 (1.11-4.49)	0.019	
Multivariable HR (95% CI)‖	1 (referent)	1.56 (0.73-3.33)	2.02 (1.01-4.05)	0.049	
	Processed red meat			*P* for trend*	*P* for heterogeneity†
	Tertile 1	Tertile 2	Tertile 3		
Overall colorectal cancer					
Cases, No. (total n=1,175)	360	412	403		
Age-adjusted HR (95% CI)§	1 (referent)	1.11 (1.01-1.22)	1.16 (1.06-1.28)	0.041	
Multivariable HR (95% CI)‖	1 (referent)	1.03 (0.94-1.13)	1.03 (0.93-1.14)	0.42	
pks+ E. coli status					0.059
pks+ E. coli negative					
Cases, No. (total n=1,064)	336	362	366		
Age-adjusted HR (95% CI)§	1 (referent)	1.05 (0.89-1.24)	1.13 (0.95-1.35)	0.38	
Multivariable HR (95% CI)‖	1 (referent)	0.98 (0.82-1.16)	1.00 (0.83-1.21)	0.48	

pks+ *E. coli* low					
Cases, No. (total n=55)	12	28	15		
Age-adjusted HR (95% CI)§	1 (referent)	2.03 (0.93-4.46)	1.01 (0.43-2.38)	0.66	
Multivariable HR (95% CI)‖	1 (referent)	1.85 (0.84-4.10)	0.90 (0.38-2.14)	0.70	
pks+ *E. coli* high					
Cases, No. (total n=56)	12	22	22		
Age-adjusted HR (95% CI)§	1 (referent)	2.03 (0.99-4.16)	2.61 (1.23-5.55)	0.003	
Multivariable HR (95% CI)‖	1 (referent)	1.90 (0.93-3.87)	2.34 (1.11-4.95)	0.053	
	Unprocessed red meat				
	Tertile 1	Tertile 2	Tertile 3	*P* for trend*	*P* for heterogeneity†
Overall colorectal cancer					
Cases, No. (total n=1,175)	376	409	390		
Age-adjusted HR (95% CI)§	1 (referent)	1.15 (1.05-1.26)	1.17 (1.06-1.30)	0.003	
Multivariable HR (95% CI)‖	1 (referent)	1.12 (1.01-1.25)	1.14 (0.98-1.33)	0.18	
pks+ *E. coli* status					0.37
pks+ *E. coli* negative					
Cases, No. (total n=1,064)	346	364	354		
Age-adjusted HR (95% CI)§	1 (referent)	1.09 (0.93-1.29)	1.15 (0.96-1.38)	0.14	
Multivariable HR (95% CI)‖	1 (referent)	1.07 (0.89-1.29)	1.12 (0.86-1.46)	0.49	
pks+ *E. coli* low					
Cases, No. (total n=55)	14	25	16		
Age-adjusted HR (95% CI)§	1 (referent)	2.41 (1.21-4.82)	1.37 (0.63-2.97)	0.50	
Multivariable HR (95% CI)‖	1 (referent)	2.33 (1.16-4.70)	1.31 (0.59-2.89)	0.63	
pks+ *E. coli* high					
Cases, No. (total n=56)	16	20	20		
Age-adjusted HR (95% CI)§	1 (referent)	1.32 (0.63-2.74)	1.47 (0.75-2.90)	0.12	
Multivariable HR (95% CI)‖	1 (referent)	1.29 (0.62-2.70)	1.44 (0.72-2.90)	0.16	

* The trend test was performed using the total red meat (unprocessed red meat and processed red meat) or processed red meat or unprocessed red meat as a continuous variable with cohort-specific ceilings at 90th percentiles in the regression model. The 90th percentile values were used for scores above 90th percentile to eliminate outlier effects.

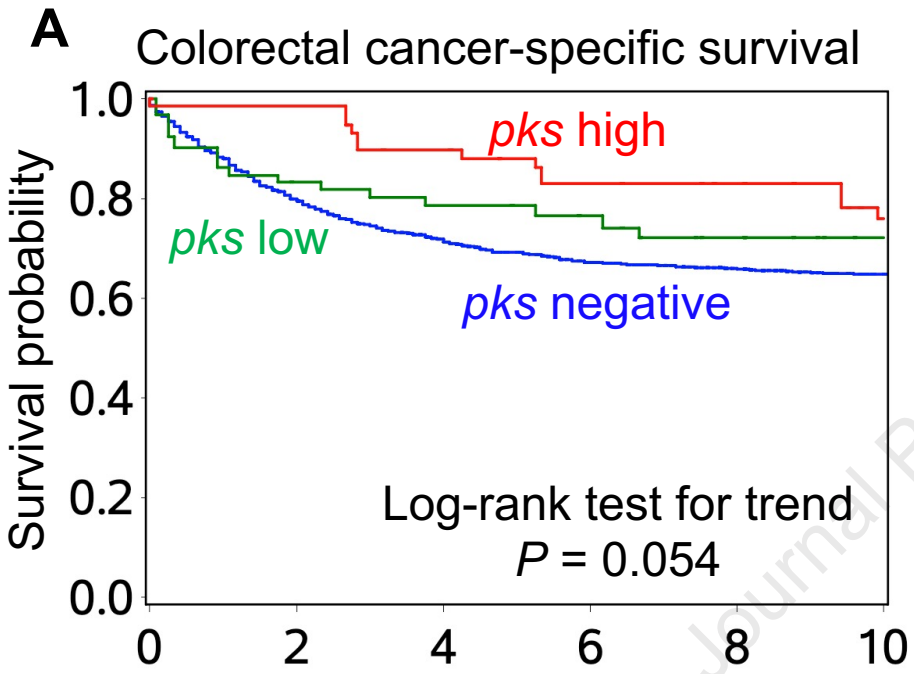
† The meta-regression method with a subtype-specific random effect term was used to test whether the association has a trend across the ordinal subtypes in the multivariable-adjusted model, where the total red meat (unprocessed red meat and processed red meat) or processed red meat or unprocessed red meat was used as a continuous variable with cohort-specific ceilings at 90th percentiles.

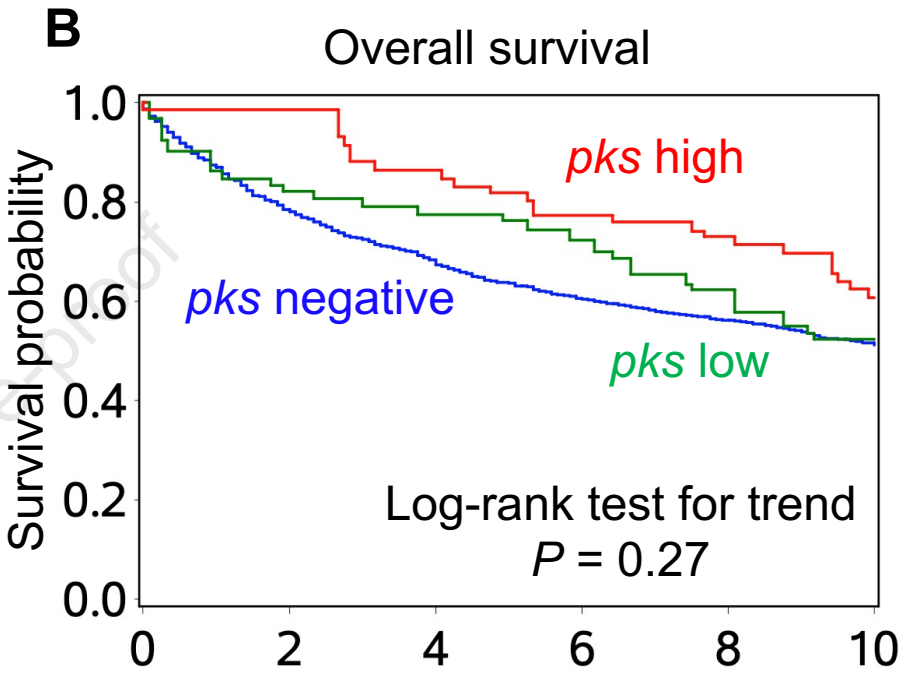
[§] Duplication-method Cox proportional hazards model weighted by inverse probabilities based on tissue bacterial data availability for competing risks data was used with total caloric intake adjusted and stratification by age (in months), sex (i.e., cohort) and year of questionnaire return.

[‖] Additionally adjusted for body mass index (continuous with 35 kg/m^2 ceiling), cumulative pack-years smoked (continuous with 50 pack-years ceiling), family history of colorectal cancer in any first-degree relative (yes vs. no), previous lower gastrointestinal endoscopy (yes vs. no), physical activity (continuous with a ceiling at 50 metabolic equivalent task score-hours per week), regular use of aspirin or nonsteroidal anti-inflammatory drugs (>2 tablets/week: yes vs. no), multivitamin use (yes vs. no), and alcohol consumption (continuous with 30 g/day ceiling), and mutually adjusted for the other type of red meat (continuous with cohort-specific ceilings at 90th percentiles).

Abbreviations: CI, confidence interval; HR, hazard ratio.

Figure S1.





C

Number at risk

The amount of *pks*+ *E. coli* DNA	Year					
	0	2	4	6	8	10
Negative	1062	868	761	682	630	571
Low	55	47	44	41	35	29
High	55	53	47	42	39	33