European Journal of Epidemiology
https://doi.org/10.1007/s10654-017-0346-8

METHODS

Utility of inverse probability weighting in molecular pathological epidemiology

Li Liu[1,2,3,4] · Daniel Nevo[5,6] · Reiko Nishihara[2,3,6,7] · Yin Cao[2,8,9] · Mingyang Song[2,8,9] · Tyler S. Twombly[1] · Andrew T. Chan[7,8,9,10] · Edward L. Giovannucci[2,6,10] · Tyler J. VanderWeele[5,6] · Molin Wang[5,6,10] · Shuji Ogino[1,3,6,7]

Received: 10 July 2017 / Accepted: 12 December 2017

Abstract

As one of causal inference methodologies, the inverse probability weighting (IPW) method has been utilized to address confounding and account for missing data when subjects with missing data cannot be included in a primary analysis. The transdisciplinary field of molecular pathological epidemiology (MPE) integrates molecular pathological and epidemiological methods, and takes advantages of improved understanding of pathogenesis to generate stronger biological evidence of causality and optimize strategies for precision medicine and prevention. Disease subtyping based on biomarker analysis of biospecimens is essential in MPE research. However, there are nearly always cases that lack subtype information due to the unavailability or insufficiency of biospecimens. To address this missing subtype data issue, we incorporated inverse probability weights into Cox proportional cause-specific hazards regression. The weight was inverse of the probability of biomarker data availability estimated based on a model for biomarker data availability status. The strategy was illustrated in two example studies; each assessed alcohol intake or family history of colorectal cancer in relation to the risk of developing colorectal carcinoma subtypes classified by tumor microsatellite instability (MSI) status, using a prospective cohort study, the Nurses' Health Study. Logistic regression was used to estimate the probability of MSI data availability for each cancer case with covariates of clinical features and family history of colorectal cancer. This application of IPW can reduce selection bias caused by nonrandom variation in biospecimen data availability. The integration of causal inference methods into the MPE approach will likely have substantial potentials to advance the field of epidemiology.

Keywords Etiologic heterogeneity · Marginal structural model · Missing at random · Neoplasm · Unique disease principle · Selection bias

Abbreviations

AUC	Area under receiver-operating characteristic curve
CCA	Complete case analysis
CI	Confidence interval
DAG	Directed acyclic graph
HR	Hazard ratio
IPW	Inverse probability weighting
MAR	Missing at random
METS	Mean metabolic equivalent task score
MCAR	Missing completely at random
MPE	Molecular pathological epidemiology
MSI	Microsatellite instability
NHS	Nurses' Health Study
ROC	Receiver-operating characteristic curve

Li Liu and Daniel Nevo have contributed equally to this work.

Molin Wang and Shuji Ogino have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-017-0346-8) contains supplementary material, which is available to authorized users.

✉ Molin Wang
stmow@channing.harvard.edu

✉ Shuji Ogino
shuji_ogino@dfci.harvard.edu

Extended author information available on the last page of the article

Published online: 20 December 2017

Introduction

Epidemiological research often aims to detect and quantify the association between an exposure and a disease, where the disease designation (name) adopted in clinical practice is used to gather individuals with similar health problems and presumably similar etiologies. With recent advances of molecular pathology, the disease classification system incorporates our improved knowledge on pathogenic mechanisms, thus allowing for better management of each individual. The field of molecular pathological epidemiology (MPE) integrates molecular pathological and epidemiological methods, and takes advantages of improved understanding of pathogenesis to generate stronger biological evidence of causality and optimize strategies for individualized prevention and treatment. MPE research typically involves studying the association of an exposure with specific subtypes of disease (most commonly, neoplastic disease), thereby clarifying the differential effects of the exposure on the development and progression of different disease subtypes [1–6]. For example, beyond epidemiological evidence for cancer preventive effects of aspirin [7–12], findings from MPE studies suggest its specific effect for certain tumor subtypes defined by tissue biomarkers [13–16]. MPE research can contribute to better understanding of the relationship between exposures and molecular pathology, leading to individualized prevention and treatment strategies [1–5].

To characterize the pathogenic heterogeneity, disease subtyping based on biomarker analysis of biospecimens is essential in MPE research. In practice, however, nearly always there exist disease cases with unavailable biomarker data. The typical strategy for dealing with the problem of missing subtype data is to perform a complete case analysis (CCA), where only cases with known subtype data are treated as outcome events and cases with missing subtype data are treated as censored at the event time. In the context of time-to-disease subtype analysis, the validity of CCA relies on the strong assumption that, among cases, the probability of missing subtype data is independent of subtype status, variables available only among cases, time of disease diagnosis and any other time-dependent covariates.

The missing at random (MAR) assumption usually means that, given covariates measured for all study participants, the probability of missing subtype data is independent of subtype status. Here we extend the MAR assumption such that the probability of missing subtype may depend on, in addition to covariates measured for all study participants, covariates that are defined and measured for cases only. In the scenarios of MAR, CCA may result in biased association estimates in MPE studies [17, 18]; this is partially because covariates measured for cases only cannot be incorporated into the model for the disease outcome as independent variables.

The causal inference field in epidemiology has advanced considerably, and developed various methodologies including inverse probability weighting (IPW) method, which is another well-established technique commonly used to deal with missing data [19–21]. In the IPW method, a statistical model for the missingness as an outcome is constructed. Then, each case with available subtype data is weighted according to the inverse of (i.e., 1 divided by) the probability of non-missing. This article aims to illustrate the use of IPW in assessing differential associations of exposures with disease subtypes in the presence of missing subtype data. We integrate IPW into a Cox proportional cause-specific hazards regression for competing risks, to address selection bias due to nonrandom availability of subtype data in MPE research [22–24]. A user-friendly SAS macro to implement this IPW method for outcome subtype analysis is publicly available.

Methods

Analysis strategy

We use colorectal cancer as a disease example; nonetheless, the IPW method for outcome subtype analysis can be applied to non-neoplastic diseases. In the context of MPE, this IPW method consists of two stages. In the first stage, the subtype data availability is modeled using logistic regression with the binary missing subtype status (subtype data available vs. missing) as the outcome variable, and covariates (including epidemiological, clinical, and pathological factors) as predictor variables. The IPW method requires that, conditional on the variables included in the logistic regression model, the probability of missing subtype data among cases is independent of subtype. The covariates included in the logistic regression model can be measured in the entire cohort or only among the disease cases. To make this assumption reasonable, it may be useful to include factors potentially associated with tumor subtype (e.g., disease stage, tumor location, tumor differentiation, age at diagnosis, etc.) [25–27] in the logistic regression model. Weights equal to the inverse of the probabilities of non-missing subtype estimated by logistic regression models are then incorporated into the second stage competing risks Cox proportional hazards regression model. Figure 1a illustrates the causal assumptions in our analysis, for a time-fixed scenario, and shows why in the MPE subtype analysis, adjusting for covariates in the model would not suffice to eliminate the selection bias. The figure describes a competing risks scenario, where Y_1 and

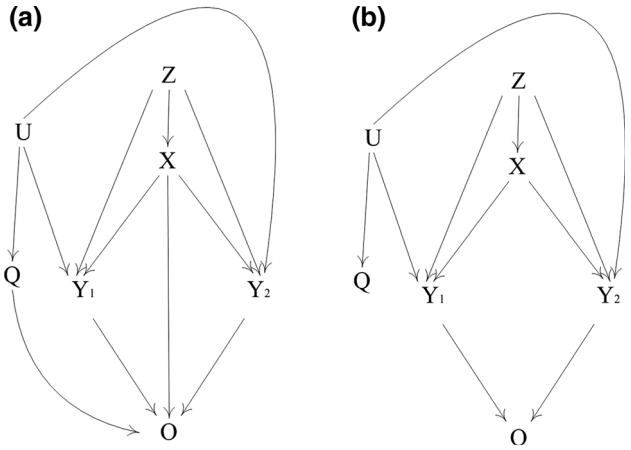

Fig. 1 Simplified diagram showing how inverse probability weighting (IPW) balances the selection bias in subtype analysis. **a** Subtype analysis without IPW; **b** subtype analysis incorporated with IPW. Y_1: subtype 1; Y_2: subtype 2; X: exposures (alcohol intake, family history of colorectal cancer, et al.); Z: confounders; Q: clinical features (tumor stage, tumor location, tumor differentiation, et al.); U: unobserved variables; O: subtype availability

Y_2 are binary indicators representing the statuses (1 = presence; 0 = absence) of different subtypes (e.g., MSI-high tumor and non-MSI-high tumor), X represents the exposure, Z represents confounders available for all study participants, U represents unobserved variables, Q represents variables only available among cases, and O is a binary indicator representing the availability of subtype data. When studying the effect of the exposure X on Y_1, the analysis is conditioned on $Y_2 = 0$. Causal interpretation of competing risks was previously discussed elsewhere [28]. Conditioning on $O = 1$, i.e., on tumor subtype availability, opens three paths. The first two are $X \rightarrow O \leftarrow Y_1$ and $X \rightarrow O \leftarrow Q \leftarrow U \rightarrow Y_1$, and the third is $X \rightarrow Y_2 \leftarrow U \rightarrow Q \rightarrow O \leftarrow Y_1$. The third path is unique in the settings of competing risks because of conditioning on $Y_2 = 0$. To eliminate the selection bias, one can apply IPW for selection bias, with the weights being the inverse of the probabilities $P(O = 1|X, Q)$. In the pseudo-population created by the weights, there are no arrows from X or Q to O, and hence no selection bias is introduced when conditioning on O (Fig. 1b). The same issues arise when studying the effect of X on Y_2.

We now turn to describe MPE analysis with time-to-disease data, before incorporating IPW. Subtype analysis concerning time-to-disease diagnosis when the disease is partitioned into K subtypes is commonly performed using the Cox model for the subtype-specific hazard function [22, 23, 29]

$$\lambda_k(t; X_i(t), Z_i(t)) = \lambda_{0k}(t) \exp\left[\beta_k X_i(t) + \gamma_k' Z_i(t)\right],$$

where $\lambda_k(t)$ and $\lambda_{0k}(t)$ are the incidence rate and baseline incidence rate, respectively, at age t for disease of subtype k, i refers to individual, X and Z are possibly time-varying exposure of interest and potential confounders, respectively, $\exp(\beta_k)$ represents the hazard ratio (HR) quantifying the exposure-subtype k association, and γ_k and Z are column vectors. For notational simplicity, we assume X is a scalar here; the method is valid for a vector-valued exposure. This model is also known as the competing risks Cox model [24, 29, 30]. Often when performing MPE analysis using this framework, the interest lies in whether X is associated with any, some, or all of the disease subtypes (in the context of the above equation, whether $\beta_k \neq 0$ for any k). Estimation of each β_k is also of interest, as it gives information on the magnitude of the association. A second goal, often of equal importance, is to determine whether the association between X and the disease is the same across subtypes (whether $\beta_1 = \beta_2 = \cdots = \beta_k$). The parameters of interest, $\beta_k, k = 1, \ldots, K$, can be estimated through fitting a single Cox model including duplicated variables X_{1i}, …, X_{ki}, Z_{1i},…,Z_{ki} in the model, stratified by subtypes, on an augmented data set, in which each block of person-time is augmented for each subtype and where X_{1i}, …, X_{ki}, Z_{1i},…,Z_{ki} are created corresponding to the K subtypes, with $X_{ki} = X_i$, $Z_{ki} = Z_i$ for subtype k and 0 otherwise, $k = 1, \ldots, K$. The estimates of β_k, $k = 1, \ldots, K$ are available as the estimated regression coefficients of X_{ki}, $k = 1, \ldots, K$. The common effect test for the null hypothesis $H_0 : \beta_1 = \cdots = \beta_K$ can also be tested in this single run of fitting the Cox model [23]. For cohort and nested case-control study designs, the tests and estimation through this data duplication method can be carried out by existing software. For example, this can be achieved by combining PROC PHREG in SAS (SAS Institute, Inc., Cary, North Carolina) with a data duplication algorithm [23, 30]. Our previous SAS macro *%subtype* [23] implements this method. This macro also works for the constrained model [23], which is the Cox model above under the assumption that the coefficients of the potential confounders are same across subtypes (i.e., $\gamma_1 = \gamma_2 = \cdots = \gamma_k$).

IPW begins with fitting of a logistic regression model among the cases for the probability

$$p_i = P(O_i = 1|\delta_i = 1, t_i, X_i, Z_i, Q_i),$$

where O_i is an indicator that equals 1 if subtype information is available and 0 otherwise, and δ_i is an indicator that equals 1 for cases and 0 for controls. This model includes t_i, the diagnosis time, the previously defined X_i and Z_i, and Q_i, a vector of measurements taken among the disease cases only, which is typically associated with the subtype and possibility with the missing subtype.

Following the illustrative Directed Acyclic Graph (DAG) for time-fixed scenario (Fig. 1a), including Q_i in the model potentially reduces the bias in estimation of hazard ratios if the Q_i's, e.g. the tumor location and stage, are associated with the missing status O_i. If X_i and/or Z_i are time-varying, they should be evaluated at the time points such that X_i and/or Z_i at those time points can best predict the missing status, or their values can be based on an appropriate metrics summarizing the history of X_i and/or Z_i over a period of time if the missing status may be related to the history of X_i and/or Z_i. Often these time points are the last available data on X_i and Z_i.

Following the model fitting, an estimated probability $\hat{p}_i$ is calculated for each of the observations with available subtype data. Standard diagnostic checks [e.g., area under receiver-operating characteristic (ROC) curves] can assist in checking the logistic regression model. In addition, one can assess the capability of inverse probability weights for selection bias in balancing the tumors with and without subtype data, by refitting the same logistic regression, while applying weights of $\frac{1}{\hat{p}_i}$ for subtype available tumors and $\frac{1}{1-\hat{p}_i}$ for subtype unavailable tumors. In the pseudo-population created by these weights, there is no association between the covariates Q_i, X_i, t_i and/or Z_i and the missing status O_i. Therefore, if the weights indeed balance tumors with and without subtype status, the coefficients in the weighted logistic regression should be close to zero. We will illustrate this balance check approach in our data example.

For the Cox model, the weights differ for controls, cases with known subtypes, and cases with missing subtypes. In the counting process data structure, which is commonly used in cohort studies to deal with the time-varying covariates and left truncation, for each control, the weight equals 1 for all time periods before the control was censored. For each case with a known subtype, the weight equals 1 for all time periods before diagnosis, and equals $\frac{1}{\hat{p}_i}$ for the period corresponding to the time of diagnosis. For each case without subtype information, the weight equals 1 for all time periods before the age of diagnosis, and equals 0 for the period corresponding to the time of diagnosis. It is possible to extend the described methodology for scenarios where IPW is used to account for missing exposure or missing confounders.

Upon calculation of the weights, the competing-risks Cox model is fitted by maximizing the following partial likelihood as a function of $\{\beta_k\}$ and $\{\gamma_k\}$,

$$L = \prod_{k=1}^{K} \prod_{i=1}^{n} \left[\frac{\exp\left[\beta_k X_i(t_i) + \gamma_k' Z_i(t_i)\right]}{\sum_{j \in R_i} w_j(t_i) \exp\left[\beta_k X_j(t_i) + \gamma_k' Z_j(t_i)\right]} \right]^{w_i(t_i)\delta_{ik}}$$

where δ_{ik} is an indicator that equals 1 if a disease of subtype k was diagnosed for participant i, t_i is the diagnosis time for the participant, R_i is the risk set at time t_i and $w_i(t)$ is the weight for participant i at time t with values described above. In SAS, this model can be easily fitted by running PROC PHREG on the duplicated dataset described above [23] together with a WEIGHT statement. We created a SAS macro for the application of this method which invokes the COVS option to get approximately correct standard deviations. Heterogeneity testing and an association test can be carried out using a Wald test procedure. The implementation of inverse probability weighted analysis is almost as simple as the unweighted CCA analysis in SAS.

Illustration of inverse probability weighted MPE analysis

We have previously conducted MPE studies to illustrate the associations between various exposures and risk of colorectal cancer subtypes, using the CCA approach. For example, we did not observe differential associations between alcohol consumption and risk of colorectal cancer subtypes according to microsatellite instability (MSI) status, which is a well-recognized tumor molecular biomarker of colorectal carcinoma [31]. In contrast, family history of colorectal cancer was associated with a greater increase in risk for MSI-high colorectal cancer than in non-MSI-high colorectal cancer [32]. Here, we use these examples to illustrate the methodology and applicability of the inverse probability weighted data duplication-method Cox proportional cause-specific hazards regression in studying disease subtype heterogeneity.

Study population

The example data were derived from the Nurses' Health Study (NHS). Details of the cohort study have been previously described [13, 15, 16, 33]. Briefly, 121,701 U.S. registered female nurses aged from 30 to 55 years were recruited into the NHS in 1976. Follow-up questionnaires were administrated at baseline and every 2 years thereafter, to collect updated lifestyle, medical and other health-related information. Incident colorectal cancer cases were ascertained and confirmed, using data from biennial questionnaires, the National Death Index, and medical records. Deaths, including lethal unreported colorectal cancer cases, were identified through National Death Index and next of kin, and cause of death in each case was determined by medical record review. A single pathologist (S.O.) performed a centralized review of hematoxylin and eosin stained tissue sections, and recorded pathological features

in all colorectal cancer cases included in the current study. This study was approved by the Institutional Review Board at the Brigham and Women's Hospital and the Harvard T.H. Chan School of Public Health.

Assessment of microsatellite instability (MSI)

DNA was extracted from formalin-fixed paraffin-embedded archival tissue of colorectal carcinoma and normal colon. The status of MSI was determined by analyzing variability in the length of the microsatellite markers from tumor DNA compared to normal DNA. MSI status was determined using 10 microsatellite markers: BAT25, BAT26, BAT40, D2S123, D5S346, D17S250, D18S55, D18S56, D18S67 and D18S487 [34]. MSI-high was defined as the presence of instability in $\geq 30\%$ of the markers, while non-MSI-high was defined as $< 30\%$ unstable markers.

Estimation of probability of observing MSI status

Logistic regression was used to compute the probability of observing MSI status. We considered two possible logistic regression models for tumor subtype availability. In Model 1, we included the clinicopathological covariates potentially related with the subtype, including disease stage (stage I, stage II, stage III, stage IV), tumor location (proximal colon, distal colon, rectum) and tumor differentiation (well, moderate, poor, unspecified). These were marked as Q in Fig. 1 and in our models. In addition, we included age at diagnosis (continuous variable) and year of diagnosis (1976–1995, 1996–2000, 2001–2012). All of these clinicopathological features have been demonstrated to be associated with MSI status previously [25–27]. The computed probability based on Model 1 was referred to as $\hat{p}^{(1)}$. Given the potential influence of family history of colorectal cancer on tumor size [35], which may associate with the availability of tumor tissue, we also considered pre-diagnosis family history of colorectal cancer in Model 2. This corresponds to the arrow between X and O in Fig. 1, for our second example, where family history of colorectal cancer is the exposure. The goodness of fit of each model was assessed by ROC curves [36]. In addition, as described in the Methods section, we rerun the logistic regression models weighted by the inverse probability of MSI status availability for cases with MSI status available and inverse probability of MSI status unavailability for cases with missing MSI status. If the estimates in the weighted logistic regression models are very close to zero, it implies that the weights balance cases with and without MSI information, with respect to the variables we assume affecting the missing status.

In primary Cox regression analyses to detect the association of the exposure (alcohol intake or family history of colorectal cancer) with risk of colorectal cancer subtypes classified by MSI, the weight for colorectal cancer cases without estimated probability of observing MSI status (non-missing) due to lack of disease stage, tumor location, tumor differentiation, year of diagnosis or age at diagnosis was assigned as 1. In secondary Cox regression analyses, cases without the aforementioned clinicopathological features were deleted. The corresponding weights using $\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ were $w^{(1)}$ and $w^{(2)}$, respectively.

Inverse probability weighted MPE analysis

We incorporated the weights into data duplication-method Cox proportional cause-specific hazards regression to estimate the associations between the exposures (alcohol intake/family history of colorectal cancer) and the risk of colorectal cancer subtypes according to MSI status in the NHS by using the SAS macro *%subtype_weights* that we developed. The macro is an extension of the macro *%subtype* that we developed previously [23]. For comparison, we also estimated the relationship between the exposures and risk of colorectal cancer subtypes classified by MSI without any weighting by using the macro *%subtype*.

Pre-diagnosis regular aspirin intake, physical activity, body mass index, pack-years of smoking, total energy intake, history of endoscopy, and either family history of colorectal cancer or alcohol intake if it was not considered as the main exposure were set as covariates when we conducted Cox regression analyses. The raw data form of the examples along with the definitions of variables is provided in the Table S1 (Online Resource). Table S2 (Online Resource) contains the augmented data set for participant id of 1. We stratified the analyses jointly by age in months at start of follow-up and calendar year of current questionnaire cycle in the NHS. The time scale for the analysis was then measured in months since the start of the current questionnaire cycle, which is equivalent to age in months because of the way we structured the data and formulated the model for analysis. The SAS code used for these examples is provided in the online supplement (Online Resource).

Results

During 32 years of follow-up in the NHS (June 1980–June 2012), we documented 2541 cases of incident colorectal carcinoma. Among these cases, 551 lacked at least one clinicopathological covariate potentially associated with tumor tissue availability, such as disease stage, tumor

location, tumor differentiation, age at diagnosis or year of diagnosis. Since these clinicopathological features were core covariates in the logistic regression models, and 93% cases (514 of 551) without complete clinicopathological features were MSI data-unavailable, the missing indicator method might be not suitable. The missing indicator method for the missingness outcome model would result in overfitting caused by quasi-complete separation of the data. Therefore, we excluded the 551 cases from the data used for fitting the models for the MSI missing status. Finally, 1990 cases were included in the logistic regression models, including 699 MSI data-available cases and 1291 MSI data-unavailable cases. Table S3 (Online Resource) shows the characteristics of cases by availability of MSI status in the NHS. Clinical and pathological features differed between MSI data-available cases and MSI data-unavailable cases.

Table 1 shows the distribution of weights for MSI data-available cases based on the aforementioned two logistic regression models: Model 1 based on clinicopathological features and model 2 based on the clinicopathological features plus the pre-diagnosis family history of colorectal cancer. The corresponding areas under the receiver-

Table 1 The distribution of weights among cases with microsatellite instability data available for the time period of colorectal cancer diagnosis in the Nurses' Health Study

Weights	No.	Median	Mean	Min	Max	Lower quartile	Upper quartile	5th percentile	95th percentile	1th percentile	99th percentile	AUC
w(1)[a]	699	2.64	2.85	1.62	10.30	2.00	3.27	1.68	4.84	1.63	7.82	0.66
w(2)[b]	699	2.64	2.85	1.55	10.23	1.99	3.25	1.68	4.93	1.59	7.45	0.66

AUC area under receiver-operating characteristic curve

[a]Weights from the logistic regression model based on disease stage, tumor location, tumor differentiation, age at diagnosis and year of diagnosis

[b]Weights from the logistic regression model based on aforementioned clinical features and pre-diagnosis family history of colorectal cancer

Table 2 Summary of the fitted and refitted logistic regression models for the availability of microsatellite instability status in the Nurses' Health Study

Parameter	Model 1		Model 2		Model 1[a]		Model 2[a]	
	β[b]	*P*-value	β[b]	*P*-value	β[b]	*P*-value	β[b]	*P*-value
Intercept	− 8.40	< 0.0001	− 8.45	< 0.0001	0.02	0.99	− 0.10	0.93
Stage II[c]	0.24	0.06	0.25	0.06	− 0.04	0.62	− 0.04	0.66
Stage III	0.17	0.21	0.18	0.19	− 0.04	0.68	− 0.03	0.73
Stage IV	− 0.13	0.44	− 0.12	0.48	− 0.07	0.54	− 0.06	0.60
Distal colon cancer[c]	− 0.13	0.27	− 0.12	0.28	0.003	0.97	0.0007	0.99
Rectal cancer	− 0.12	0.35	− 0.12	0.38	0.01	0.91	0.01	0.89
Moderate differentiation[c]	0.25	0.10	0.25	0.11	0.06	0.56	0.06	0.53
Poor differentiation	0.28	0.13	0.28	0.13	0.09	0.46	0.09	0.45
Unspecified differentiation	− 0.46	0.05	− 0.47	0.04	0.08	0.58	0.09	0.51
Age at diagnosis (continuous)	0.22	0.0006	0.22	0.0006	− 0.002	0.95	0.0008	0.98
Square of age at diagnosis (continuous)	− 0.002	0.0009	− 0.002	0.0008	3×10^{-5}	0.91	7×10^{-6}	0.98
Year of diagnosis (1996–2000)[c]	1.00	< 0.0001	1.00	< 0.0001	− 0.03	0.74	− 0.03	0.72
Year of diagnosis (2001–2012)	0.34	0.02	0.34	0.02	− 0.05	0.60	− 0.05	0.57
Family history of CRC (Yes)			0.14	0.27			0.009	0.91

Model 1: The logistic regression model based on disease stage, tumor location, tumor differentiation, age at diagnosis and year of diagnosis

Model 2: The logistic regression model based on disease stage, tumor location, tumor differentiation, age at diagnosis, year of diagnosis and family history of colorectal cancer

[a]The refitted logistic regression models weighted by the inverse probability of microsatellite instability (MSI) status availability for cases with MSI status available and the inverse probability of MSI status unavailability for cases with missing MSI status

[b]Estimate represents the estimated regression coefficient in the logistic regression model

[c]The reference groups for disease stage, tumor location, tumor differentiation and year of diagnosis are stage I, proximal colon cancer, well differentiation and year of diagnosis between 1976 and 1995, respectively

Table 3 Alcohol intake and risk of colorectal cancer according to microsatellite instability in the Nurses' Health Study[a]

	Alcohol intake (g/day)	$\log(\widehat{HR})(SE)^b$	$\widehat{HR}$ (95% CI)[b]	*P*-value	$P^c_{\text{heterogeneity}}$
Unweighted model					
Microsatellite instability (MSI)					
Non-MSI-high	0	Ref.			0.42
	1–14	0.22 (0.109)	1.25 (1.009–1.546)	0.04	
	≥ 15	0.15 (0.152)	1.16 (0.863–1.568)	0.32	
MSI-high	0	Ref.			
	1–14	0.23 (0.219)	1.25 (0.815–1.925)	0.30	
	≥ 15	0.46 (0.281)	1.58 (0.912–2.742)	0.10	
Model weighted by $w^{(1)}$ [d]					
Microsatellite instability (MSI)					
Non-MSI-high	0	Ref.			0.36
	1–14	0.17 (0.113)	1.19 (0.952–1.483)	0.13	
	≥ 15	0.21 (0.162)	1.24 (0.901–1.702)	0.19	
MSI-high	0	Ref.			
	1–14	0.25 (0.229)	1.29 (0.821–2.018)	0.27	
	≥ 15	0.56 (0.287)	1.75 (0.996–3.068)	0.05	
Model weighted by $w^{(2)}$ [e]					
Microsatellite instability (MSI)					
Non-MSI-high	0	Ref.			0.35
	1–14	0.17 (0.113)	1.19 (0.950–1.481)	0.13	
	≥ 15	0.21 (0.163)	1.24 (0.900–1.702)	0.19	
MSI-high	0	Ref.			
	1–14	0.25 (0.230)	1.29 (0.820–2.019)	0.27	
	≥ 15	0.56 (0.288)	1.75 (0.997–3.080)	0.05	

CI confidence interval; *HR* hazard ratio; *MSI* microsatellite instability

[a]For each control, the weight equals 1 for all time periods before the control was censored. For each case with a known subtype, the weight equals 1 for all time periods before diagnosis, and equals the inverse probability of MSI data availability for the period corresponding to the time of diagnosis. If the probability of MSI availability is not available for a case with a known subtype due to lack of data on disease stage, tumor location, tumor differentiation, age at diagnosis or year of diagnosis, the weight equals 1 for the period corresponding to the time of diagnosis. For each case without subtype information, the weight equals 1 for all periods before the time of diagnosis, and equals 0 for the period corresponding to the time of diagnosis

[b]The HRs were adjusted for body mass index (< 25 vs. 25–29.9 vs. ≥ 30 kg/m^2), pack-year of smoking (0 vs. 1–19 vs. 20–39 vs. ≥ 40 pack-years), family history of colorectal cancer (yes/no), history of endoscopy (yes/no), physical activity level [quintiles of mean metabolic equivalent task score (METS)–hours per week], total calorie intake (quintiles of kcal/day) and regular aspirin use (yes/no). The Cox models were also conditioned on age in months and calendar years of the questionnaire cycles

[c]$P_{\text{heterogeneity}}$ from the Wald Test

[d]Weights from the logistic regression model based on disease stage, tumor location, tumor differentiation, age at diagnosis and year of diagnosis

[e]Weights from the logistic regression model based on aforementioned clinical features and pre-diagnosis family history of colorectal cancer

operating characteristic curves (AUCs) were 0.66 for both models.

Table 2 presents the estimates for the fitted and refitted logistic regression models for the availability of MSI data. In the fitted models, tumor differentiation, age at diagnosis and year of diagnosis were the main factors associated with the availability of MSI status. After refitting the logistic regression models by including the weights we mentioned in the Methods section, the estimated regression coefficients were now all close to zero, which implied a good performance of the weights in balancing cases with and without MSI data with respect to the covariates in the logistic regression models.

The primary results of alcohol intake with risk of colorectal cancer subtypes according to MSI status are presented in Table 3. For the heavy drinkers (≥ 15 g/day) versus nondrinkers, weighting by the inverse probability of MSI availability resulted in more profound associations between alcohol intake and colorectal cancer risk for both MSI subtypes (*P* value changed from 0.10 to 0.05 for MSI-high tumors), and the adverse effect on the MSI-high subtype [HR = 1.75; 95% confidence interval (CI), 0.997–3.080 from Model 2] tended to be stronger than that on the non-MSI-high subtype (HR = 1.24; 95% CI, 0.900–1.702). For the middle drinking group (1–14 g/day) versus nondrinkers, from the unweighted analysis, the estimated HRs were approximately the same between the two subtypes (MSI-high tumors: HR = 1.25; 95% CI, 0.815–1.925; non-MSI-high tumors: HR = 1.25; 95% CI, 1.009–1.546). However, the IPW analysis led to stronger association for the MSI-high subtype and attenuated association for the non-MSI-high subtype (*P*-value changed from 0.04 to 0.13), and this resulted in the same trend as that mentioned above for the heavy drinker group, which was that the adverse effect on the MSI-high subtype tended to be stronger than that on the non-MSI-high subtype (MSI-high tumors: HR = 1.29; 95% CI, 0.820–2.019; non-MSI-high tumors: HR = 1.19; 95% CI, 0.950–1.481). When we further deleted cases with MSI status but lacking data on either disease stage, tumor location, tumor differentiation, age at diagnosis, or year of diagnosis as sensitivity analyses, the results were similar to those from the primary analyses (Online Resource, Table S4).

Table 4 Family history of colorectal cancer and risk of colorectal cancer according to microsatellite instability in the Nurses' Health Study[a]

	First-degree relatives with colorectal cancer	$\widehat{\log(HR)}\,(SE)^b$	$\widehat{HR}$ (95% CI)[b]	*P*-value	$P^c_{\text{heterogeneity}}$
Unweighted model					
Microsatellite instability (MSI)					
Non-MSI-high	No	Ref.			0.02
	Yes	0.33 (0.109)	1.40 (1.129–1.729)	0.002	
MSI-high	No	Ref.			
	Yes	0.71 (0.187)	2.04 (1.416–2.949)	0.0001	
Model weighted by $w^{(1)}$ [d]					
Microsatellite instability (MSI)					
Non-MSI-high	No	Ref.			0.05
	Yes	0.38 (0.120)	1.47 (1.159–1.859)	0.001	
MSI-high	No	Ref.			
	Yes	0.70 (0.185)	2.01 (1.398–2.888)	0.0002	
Model weighted by $w^{(2)}$ [e]					
Microsatellite instability (MSI)					
Non-MSI-high	No	Ref.			0.05
	Yes	0.30 (0.120)	1.35 (1.068–1.709)	0.01	
MSI-high	No	Ref.			
	Yes	0.63 (0.186)	1.87 (1.299–2.688)	0.008	

CI confidence interval; *HR* hazard ratio; *MSI* microsatellite instability

[a]For each control, the weight equals 1 for all time periods before the control was censored. For each case with a known subtype, the weight equals 1 for all time periods before diagnosis, and equals the inverse probability of MSI data availability for the period corresponding to the time of diagnosis. If the probability of MSI availability is not available for a case with a known subtype due to lack of data on disease stage, tumor location, tumor differentiation, age at diagnosis or year of diagnosis, the weight equals 1 for the period corresponding to the time of diagnosis. For each case without subtype information, the weight equals 1 for all periods before the time of diagnosis, and equals 0 for the period corresponding to the time of diagnosis

[b]The HRs were adjusted for body mass index (< 25 vs. 25–29.9 vs. ≥ 30 kg/m^2), pack-years of smoking (0 vs. 1–19 vs. 20–39 vs. ≥ 40 pack-years), history of endoscopy (yes/no), physical activity level [quintiles of mean metabolic equivalent task score (METS)–hours per week], total calorie intake (quintiles of kcal/day), total alcohol intake (0 vs. 1–14 vs. ≥ 15 g/day) and regular aspirin use (yes/no). The Cox models were also conditioned on age in months and calendar years of the questionnaire cycles

[c]$P_{\text{heterogeneity}}$ from the Wald Test

[d]Weights from the logistic regression model based on disease stage, tumor location, tumor differentiation, age at diagnosis and year of diagnosis

[e]Weights from the logistic regression model based on aforementioned clinical features and pre-diagnosis family history of colorectal cancer

The primary analyses revealed potentially differential associations between family history of colorectal cancer and risk of colorectal cancer subtypes by MSI status in unweighted and weighted Cox proportional hazards regression models (Table 4). The *P*-values for heterogeneity in the weighted and unweighted models were 0.05 and 0.02, respectively. Table S5 (Online Resource) presents the results from the secondary analyses of family history of colorectal cancer and risk of colorectal cancer subtypes according to MSI status by further deletion of cases with MSI status but lacking data on disease stage, tumor location, tumor differentiation, age at diagnosis, or year of diagnosis as sensitivity analyses. The results did not change substantially.

Discussion

In this study, we illustrated the use of inverse probability weighting in molecular pathological epidemiology (MPE) research to assess etiological heterogeneity across disease subtypes while addressing selection bias due to biospecimen data availability. The integrative MPE field has been growing for recent years [4, 6, 37–40]. The usefulness of the MPE approach has been discussed widely in the literature [41–51]. To fully leverage the potential of the MPE approach, further development of statistical methodologies is essential [6, 23, 52–56].

The method described in this study aims to correct for potential bias arising from differences between cases with and without available subtyping biomarker data. It should be noted that the purpose of the integration of the IPW method into MPE analyses is to address selection bias, but not to improve statistical power or efficiency. We developed a user-friendly SAS macro *%subtype_weights* that implements the IPW method. The macro is freely available at https://www.hsph.harvard.edu/molin-wang/software/.

In typical MPE studies, a complete case analysis (CCA) including only cases with relevant biospecimen data is conducted where cases without relevant biospecimen data are excluded based on the assumption that the included and excluded cases are exchangeable. When this assumption holds true, results from CCA are expected to be similar to results from analysis of the excluded cases (if relevant biospecimen data status would have been known and used) [33, 57, 58]. However, often there is a selection bias for included cases due to nonrandomness of biomarker data availability, and generalizability of findings based on the included cases need to be carefully evaluated [59]. Note that in the subtype data analysis considered in this paper, the method that treats all the cases with missing subtype data as an additional subtype is equivalent to the CCA method. In contrast to the plain CCA strategy, inverse probability weighting (IPW) can be used as a method of dealing with nonrandomly missing biomarker data, especially when there are substantial differences between biomarker (subtype) data available and unavailable cases. Our application of IPW attempts to utilize information from all cases (including cases with and without subtype data) in order to produce unbiased estimations of the associations of an exposure with outcome subtypes [20]. In the colorectal cancer analysis example, the probability of missing tissue specimens is higher in stage IV cases (due to the presence of unresectable advanced cancers) than in stage I to III cases. When we treat tumors with different stages equally, we fail to fully account the missing contribution of some stage IV patients. With the IPW method, we can correct the contribution of stage IV patients by giving them larger weights (inverse to the probability of biospecimen data availability).

Various factors can influence the probability of missing biospecimen data. In our colorectal cancer data example, using data on epidemiological and clinicopathological factors, we constructed a statistical model to estimate the probability of available subtyping biomarker data in each case. Our example can provide useful information in the application of IPW method to MPE research though its general applicability needs to be further tested. Because the best possible model of biospecimen data missingness is likely not only cohort-specific but also disease-specific, cohort- and disease-specific models need to be constructed when integrating the IPW method into MPE research.

In a large-scale cohort study, data on a predictor of missing biospecimen data are themselves missing in some individuals. Here, we provide some suggestions for dealing with missing data in such a predictor. For missing lifestyle variable data in some cases, exclusion of the cases from the missingness model is preferred when the proportion of the missing cases is small. Alternative strategies include imputation of data and the missing indicator method. For cases with missing data on critical clinicopathological features, exclusion may be a better option, especially when the availability of these features independently predicts the probability of missingness of biospecimen data [20]. In our example, 93% of disease cases with missing data in either of disease stage, tumor location, tumor differentiation, age at diagnosis or year of diagnosis were MSI data-unavailable; thus, we excluded the cases with missing those data from estimation of probability of MSI data availability. An alternative approach to deal with missing data of these clinicopathological features is to estimate the probability of observing MSI data by averaging the estimated probabilities over the distribution of the feature with missing data. For example, if tumor location datum is missing for a case, we can calculate the estimated probability for each possible value of tumor location using the fitted logistic regression

model while plugging in the values of the other variables as observed for this case, and then estimate the probability of observing MSI status for this case by the sum of the products of the obtained estimated probabilities for each tumor location and the percentage of the corresponding tumor location among the cases.

In an alternative method published recently [18], a model for Q conditional on the subtype, the time of diagnosis and other covariates was integrated with a partial likelihood based on the cause-specific Cox model for time to disease subtype data. This method and the IPW method illustrated in this paper rely on different model assumptions, and it is arguably easier to understand and implement the IPW method.

In summary, we have presented data duplication-method Cox proportional hazards regression weighted by the inverse probability of availability of subtyping biomarker data in MPE research to assess differential associations of an exposure with disease subtypes classified by the biomarker. This method is helpful for reduction of selection bias resulting from nonrandom missingness of disease subtype information. In the near future, the integration of causal inference methodologies into the MPE approach will likely have substantial potentials to advance the field of epidemiology, and this integrative area should be further explored.

Acknowledgements We would like to thank the participants and staff of 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.

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; R01 CA137178 to A.T.C.; K24 DK098311 to A.T.C.; R01 CA151993 to S.O.; R35 CA197735 to S.O.; K07 CA190673 to R.N.]; and Nodal Award (to S.O.) from the Dana-Farber Harvard Cancer Center. L.L. is supported by the grant from National Natural Science Foundation of China No. 81302491, a scholarship grant from Chinese Scholarship Council and a fellowship grant from Huazhong University of Science and Technology. 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.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Ogino S, Lochhead P, Chan AT, Nishihara R, Cho E, Wolpin BM, Meyerhardt JA, Meissner A, Schernhammer ES, Fuchs CS, Giovannucci E. Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease. Mod Pathol. 2013;26(4):465–84.
2. Ogino S, Nishihara R, VanderWeele TJ, Wang M, Nishi A, Lochhead P, Qian ZR, Zhang X, Wu K, Nan H, Yoshida K, Milner DA Jr, Chan AT, Field AE, Camargo CA Jr, Williams MA, Giovannucci EL. Review article: the role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic diseases in the era of precision medicine. Epidemiology. 2016;27(4):602–11.
3. Nishihara R, VanderWeele TJ, Shibuya K, Mittleman MA, Wang M, Field AE, Giovannucci E, Lochhead P, Ogino S. Molecular pathological epidemiology gives clues to paradoxical findings. Eur J Epidemiol. 2015;30(10):1129–35.
4. Nishi A, Milner DA Jr, Giovannucci EL, Nishihara R, Tan AS, Kawachi I, Ogino S. Integration of molecular pathology, epidemiology and social science for global precision medicine. Expert Rev Mol Diagn. 2016;16(1):11–23.
5. Ogino S, Chan AT, Fuchs CS, Giovannucci E. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. Gut. 2011;60(3):397–411.
6. Richiardi L, Barone-Adesi F, Pearce N. Cancer subtypes in aetiological research. Eur J Epidemiol. 2017;32(5):353–61.
7. Drew DA, Cao Y, Chan AT. Aspirin and colorectal cancer: the promise of precision chemoprevention. Nat Rev Cancer. 2016;16(3):173–86.
8. Chia WK, Ali R, Toh HC. Aspirin as adjuvant therapy for colorectal cancer–reinterpreting paradigms. Nat Rev Clin Oncol. 2012;9(10):561–70.
9. Tougeron D, Sha D, Manthravadi S, Sinicrope FA. Aspirin and colorectal cancer: back to the future. Clin Cancer Res. 2014;20(5):1087–94.
10. Umar A, Steele VE, Menter DG, Hawk ET. Mechanisms of nonsteroidal anti-inflammatory drugs in cancer prevention. Semin Oncol. 2016;43(1):65–77.
11. Jiang MJ, Dai JJ, Gu DN, Huang Q, Tian L. Aspirin in pancreatic cancer: chemopreventive effects and therapeutic potentials. Biochim Biophys Acta. 2016;1866(2):163–76.
12. Coyle C, Cafferty FH, Langley RE. Aspirin and colorectal cancer prevention and treatment: is it for everyone? Curr Colorectal Cancer Rep. 2016;12:27–34.
13. Liao X, Lochhead P, Nishihara R, Morikawa T, Kuchiba A, Yamauchi M, Imamura Y, Qian ZR, Baba Y, Shima K, Sun R, Nosho K, Meyerhardt JA, Giovannucci E, Fuchs CS, Chan AT, Ogino S. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. N Engl J Med. 2012;367(17):1596–606.
14. Nishihara R, Lochhead P, Kuchiba A, Jung S, Yamauchi M, Liao X, Imamura Y, Qian ZR, Morikawa T, Wang M, Spiegelman D, Cho E, Giovannucci E, Fuchs CS, Chan AT, Ogino S. Aspirin use and risk of colorectal cancer according to BRAF mutation status. JAMA. 2013;309(24):2563–71.

15. Chan AT, Ogino S, Fuchs CS. Aspirin and the risk of colorectal cancer in relation to the expression of COX-2. N Engl J Med. 2007;356(21):2131–42.
16. Cao Y, Nishihara R, Qian ZR, Song M, Mima K, Inamura K, Nowak JA, Drew DA, Lochhead P, Nosho K, Morikawa T, Zhang X, Wu K, Wang M, Garrett WS, Giovannucci EL, Fuchs CS, Chan AT, Ogino S. Regular aspirin use associates with lower risk of colorectal cancers with low numbers of tumor-infiltrating lymphocytes. Gastroenterology. 2016;151(5):879–92.
17. Lu K, Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics. 2001;57(4):1191–7.
18. Nevo D, Nishihara R, Ogino S, Wang M. The competing risks Cox model with auxiliary case covariates under weaker missing-at-random cause of failure. Lifetime Data Anal. 2017. https://doi.org/10.1007/s10985-017-9401-8.
19. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995;142(12):1255–64.
20. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278–95.
21. Graff RE, Pettersson A, Lis RT, Ahearn TU, Markt SC, Wilson KM, Rider JR, Fiorentino M, Finn S, Kenfield SA, Loda M, Giovannucci EL, Rosner B, Mucci LA. Dietary lycopene intake and risk of prostate cancer defined by ERG protein expression. Am J Clin Nutr. 2016;103(3):851–60.
22. Cox DR. Regression models and life-tables. J R Stat Soc Ser B (Methodological). 1972;34(2):187–220.
23. Wang M, Spiegelman D, Kuchiba A, Lochhead P, Kim S, Chan AT, Poole EM, Tamimi R, Tworoger SS, Giovannucci E, Rosner B, Ogino S. Statistical methods for studying disease subtype heterogeneity. Stat Med. 2016;35(5):782–800.
24. Lunn M, McNeil D. Applying Cox regression to competing risks. Biometrics. 1995;51(2):524–32.
25. Ballester V, Rashtak S, Boardman L. Clinical and molecular features of young-onset colorectal cancer. World J Gastroenterol. 2016;22(5):1736–44.
26. Ogino S, Nosho K, Kirkner GJ, Kawasaki T, Meyerhardt JA, Loda M, Giovannucci EL, Fuchs CS. CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. Gut. 2009;58(1):90–6.
27. Lochhead P, Kuchiba A, Imamura Y, Liao X, Yamauchi M, Nishihara R, Qian ZR, Morikawa T, Shen J, Meyerhardt JA, Fuchs CS, Ogino S. Microsatellite instability and BRAF mutation testing in colorectal cancer prognostication. J Natl Cancer Inst. 2013;105(15):1151–6.
28. Hernán MA, Robins JM. Causal survival analysis. In: Causal inference. Boca Raton: Chapman & Hall/CRC, forthcoming; 2018. p. 69–78. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.
29. Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978;34(4):541–54.
30. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol. 2009;170(2):244–56.
31. Schernhammer ES, Giovannuccci E, Fuchs CS, Ogino S. A prospective study of dietary folate and vitamin B and colon cancer according to microsatellite instability and KRAS mutational status. Cancer Epidemiol Biomark Prev. 2008;17(10):2895–8.
32. Ogino S, Nishihara R, Lochhead P, Imamura Y, Kuchiba A, Morikawa T, Yamauchi M, Liao X, Qian ZR, Sun R, Sato K, Kirkner GJ, Wang M, Spiegelman D, Meyerhardt JA, Schernhammer ES, Chan AT, Giovannucci E, Fuchs CS. Prospective study of family history and colorectal cancer risk by tumor LINE-1 methylation level. J Natl Cancer Inst. 2013;105(2):130–40.
33. Song M, Nishihara R, Wu K, Qian ZR, Kim SA, Sukawa Y, Mima K, Inamura K, Masuda A, Yang J, Fuchs CS, Giovannucci EL, Ogino S, Chan AT. Marine omega-3 polyunsaturated fatty acids and risk of colorectal cancer according to microsatellite instability. J Natl Cancer Inst. 2015;107(4):djv007.
34. Ogino S, Brahmandam M, Cantor M, Namgyal C, Kawasaki T, Kirkner G, Meyerhardt JA, Loda M, Fuchs CS. Distinct molecular features of colorectal carcinoma with signet ring cell component and colorectal carcinoma with mucinous component. Mod Pathol. 2006;19(1):59–68.
35. Lynch KL, Ahnen DJ, Byers T, Weiss DG, Lieberman DA. First-degree relatives of patients with advanced colorectal adenomas have an increased prevalence of colorectal cancer. Clin Gastroenterol Hepatol. 2003;1(2):96–102.
36. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
37. Ogino S, Campbell PT, Nishihara R, Phipps AI, Beck AH, Sherman ME, Chan AT, Troester MA, Bass AJ, Fitzgerald KC, Irizarry RA, Kelsey KT, Nan H, Peters U, Poole EM, Qian ZR, Tamimi RM, Tchetgen Tchetgen EJ, Tworoger SS, Zhang X, Giovannucci EL, van den Brandt PA, Rosner BA, Wang M, Chatterjee N, Begg CB. Proceedings of the second international molecular pathological epidemiology (MPE) meeting. Cancer Causes Control. 2015;26(7):959–72.
38. Campbell PT, Rebbeck TR, Nishihara R, Beck AH, Begg CB, Bogdanov AA, Cao Y, Coleman HG, Freeman GJ, Heng YJ, Huttenhower C, Irizarry RA, Kip NS, Michor F, Nevo D, Peters U, Phipps AI, Poole EM, Qian ZR, Quackenbush J, Robins H, Rogan PK, Slattery ML, Smith-Warner SA, Song M, VanderWeele TJ, Xia D, Zabor EC, Zhang X, Wang M, Ogino S. Proceedings of the third international molecular pathological epidemiology (MPE) meeting. Cancer Causes Control. 2017;28(2):167–76.
39. Hamada T, Keum N, Nishihara R, Ogino S. Molecular pathological epidemiology: new developing frontiers of big data science to study etiologies and pathogenesis. J Gastroenterol. 2017;52(3):265–75.
40. Gao C. Molecular pathological epidemiology in diabetes mellitus and risk of hepatocellular carcinoma. World J Hepatol. 2016;8(27):1119–27.
41. Rescigno T, Micolucci L, Tecce MF, Capasso A. Bioactive nutrients and nutrigenomics in age-related diseases. Molecules. 2017;22(1):E105.
42. Bishehsari F, Mahdavinia M, Vacca M, Malekzadeh R, Mariani-Costantini R. Epidemiological transition of colorectal cancer in developing countries: environmental factors, molecular pathways, and opportunities for prevention. World J Gastroenterol. 2014;20(20):6055–72.
43. Martinez-Useros J, Garcia-Foncillas J. Obesity and colorectal cancer: molecular features of adipose tissue. J Transl Med. 2016;14:21.
44. Serafino A, Sferrazza G, Colini Baldeschi A, Nicotera G, Andreola F, Pittaluga E, Pierimarchi P. Developing drugs that target the Wnt pathway: recent approaches in cancer and neurodegenerative diseases. Expert Opin Drug Discov. 2017;12(2):169–86.
45. Patil H, Saxena SG, Barrow CJ, Kanwar JR, Kapat A, Kanwar RK. Chasing the personalized medicine dream through biomarker validation in colorectal cancer. Drug Discov Today. 2017;22(1):111–9.
46. Alnabulsi A, Murray GI. Integrative analysis of the colorectal cancer proteome: potential clinical impact. Expert Rev Proteomics. 2016;13(10):917–27.

47. Kuroiwa-Trzmielina J, Wang F, Rapkins RW, Rapkins RW, Ward RL, Buchanan DD, Win AK, Clendenning M, Rosty C, Southey MC, Winship IM, Hopper JL, Jenkins MA, Olivier J, Hawkins NJ, Hitchins MP. SNP rs16906252C > T is an expression and methylation quantitative trait locus associated with an increased risk of developing MGMT-methylated colorectal cancer. Clin Cancer Res. 2016;22(24):6266–77.
48. Slattery ML, Lee FY, Pellatt AJ, Mullany LE, Stevens JR, Samowitz WS, Wolff RK, Herrick JS. Infrequently expressed miRNAs in colorectal cancer tissue and tumor molecular phenotype. Mod Pathol. 2017;30(8):1152–69.
49. Hughes LA, Khalid-de Bakker CA, Smits KM, van den Brandt PA, Jonkers D, Ahuja N, Herman JG, Weijenberg MP, van Engeland M. The CpG island methylator phenotype in colorectal cancer: progress and problems. Biochim Biophys Acta. 2012;1825(1):77–85.
50. Campbell PT, Newton CC, Newcomb PA, Phipps AI, Ahnen DJ, Baron JA, Buchanan DD, Casey G, Cleary SP, Cotterchio M, Farris AB, Figueiredo JC, Gallinger S, Green RC, Haile RW, Hopper JL, Jenkins MA, Le Marchand L, Makar KW, McLaughlin JR, Potter JD, Renehan AG, Sinicrope FA, Thibodeau SN, Ulrich CM, Win AK, Lindor NM, Limburg PJ. Association between body mass index and mortality for colorectal cancer survivors: overall and by tumor molecular phenotype. Cancer Epidemiol Biomark Prev. 2015;24(8):1229–38.
51. Gray RT, Loughrey MB, Bankhead P, Cardwell CR, McQuaid S, O'Neill RF, Arthur K, Bingham V, McGready C, Gavin AT, James JA, Hamilton PW, Salto-Tellez M, Murray LJ, Coleman HG. Statin use, candidate mevalonate pathway biomarkers, and colon cancer survival in a population-based cohort study. Br J Cancer. 2017;116(12):1652–9.
52. Begg CB, Orlow I, Zabor EC, Arora A, Sharma A, Seshan VE, Bernstein JL. Identifying etiologically distinct sub-types of cancer: a demonstration project involving breast cancer. Cancer Med. 2015;4(9):1432–9.
53. Begg CB, Seshan VE, Zabor EC, Furberg H, Arora A, Shen R, Maranchie JK, Nielsen ME, Rathmell WK, Signoretti S, Tamboli P, Karam JA, Choueiri TK, Hakimi AA, Hsieh JJ. Genomic investigation of etiologic heterogeneity: methodologic challenges. BMC Med Res Methodol. 2014;14:138.
54. Begg CB, Zabor EC, Bernstein JL, Bernstein L, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. Stat Med. 2013;32(29):5039–52.
55. Chatterjee N, Sinha S, Diver WR, Feigelson HS. Analysis of cohort studies with multivariate and partially observed disease classification data. Biometrika. 2010;97(3):683–98.
56. Wang M, Kuchiba A, Ogino S. A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. Am J Epidemiol. 2015;182(3):263–70.
57. Inamura K, Song M, Jung S, Nishihara R, Yamauchi M, Lochhead P, Qian ZR, Kim SA, Mima K, Sukawa Y, Masuda A, Imamura Y, Zhang X, Pollak MN, Mantzoros CS, Harris CC, Giovannucci E, Fuchs CS, Cho E, Chan AT, Wu K, Ogino S. Prediagnosis plasma adiponectin in relation to colorectal cancer risk according to KRAS mutation status. J Natl Cancer Inst. 2016;108(4):djv363.
58. Song M, Nishihara R, Wang M, Chan AT, Qian ZR, Inamura K, Zhang X, Ng K, Kim SA, Mima K, Sukawa Y, Nosho K, Fuchs CS, Giovannucci EL, Wu K, Ogino S. Plasma 25-hydroxyvitamin D and colorectal cancer risk according to tumour immunity status. Gut. 2016;65(2):296–304.
59. Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples LA. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Stat Med. 2003;22(4):545–57.

Affiliations

Li Liu[1,2,3,4] · Daniel Nevo[5,6] · Reiko Nishihara[2,3,6,7] · Yin Cao[2,8,9] · Mingyang Song[2,8,9] · Tyler S. Twombly[1] · Andrew T. Chan[7,8,9,10] · Edward L. Giovannucci[2,6,10] · Tyler J. VanderWeele[5,6] · Molin Wang[5,6,10] · Shuji Ogino[1,3,6,7]

[1] Department of Oncologic Pathology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA

[2] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[3] Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 450 Brookline Ave., Room SM1036, Boston, MA 02215, USA

[4] Department of Epidemiology and Biostatistics, and the Ministry of Education Key Lab of Environment and Health, School of Public Health, Huazhong University of Science and Technology, Wuhan, Hubei, People's Republic of China

[5] Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Ave., Boston, MA 02215, USA

[6] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[7] Broad Institute of MIT and Harvard, Cambridge, MA, USA

[8] Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA

[9] Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[10] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA