The Practice of Epidemiology

A Meta-Regression Method for Studying Etiologic Heterogeneity across Disease Subtypes Classified by Multiple Biomarkers

Molin Wang, Aya Kuchiba, Shuji Ogino

Correspondence to Molin Wang, Departments of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, and Channing Division of Network Medicine, Brigham and Women’s Hospital, Harvard Medical School, 677 Huntington Ave., Boston, Massachusetts (email: stmow@channing.harvard.edu)

Abbreviations: CIMP, CpG island methylator phenotype; HPFS, Health Professionals Follow-up Study; MPE, molecular pathological epidemiology; MSI, microsatellite instability; MSS, microsatellite stable; NHS, the Nurses’ Health Study; OR, odds ratio; RR, relative risk; RRR, ratio of relative risk.

We use standardized official symbol *BRAF*, which is described at www.genenames.org.

Word count main abstract: 200

Word count text: 3615

N.Figures: 0 N.Tables: 2 Appendices: 0

Abstract

In interdisciplinary biomedical, epidemiological, and population research, it is increasingly necessary to consider pathogenesis and inherent heterogeneity of any given health condition and outcome. As the unique disease principle implies, no single biomarker can perfectly define disease subtypes. The complex nature of molecular pathology and biology necessitates biostatistical methodologies to simultaneously analyze multiple biomarkers and subtypes. To analyze and test for heterogeneity hypothesis across subtypes defined by multiple categorical and/or ordinal markers, the authors developed a meta-regression method that can utilize existing statistical software for mixed model analysis. This method can be used to assess whether the exposure-subtype associations are different across subtypes defined by one marker while controlling for other markers, and to evaluate whether the difference in exposure-subtype association across subtypes defined by one marker depends on any other markers. To illustrate this method in molecular pathological epidemiology research, the authors examined the associations between smoking status and colorectal cancer subtypes defined by three correlated tumor molecular characteristics (CpG island methylator phenotype, microsatellite instability and *BRAF* mutation), in the Nurses' Health Study and the Health Professionals Follow-up Study. This method can be widely useful as molecular diagnostics and genomic technologies become routine in clinical medicine and public health.

Key words: causal inference; genomics; heterogeneity test; molecular diagnosis; omics; transdisciplinary.

Epidemiologic research typically aims to investigate the relationship between exposure and disease, based on the underlying premise that individuals with the same disease name have similar etiologies and disease evolution. With the advancement of biomedical sciences, it is increasingly evident that many human disease processes comprise of a range of heterogeneous molecular pathologic processes, modified by the exposome (1). Molecular classification can be utilized in epidemiology because individuals with similar molecular pathologic processes likely share similar etiologies (2). Pathogenic heterogeneity has been considered in various neoplasms such as endometrial (3), colorectal (3-20), and lung cancers (21-24), as well as non-neoplastic diseases such as stroke (25), cardiovascular disease (26), autism (27), infectious disease (28), autoimmune disease (29), glaucoma (30), and obesity (31).

New statistical methodologies to address disease heterogeneity are useful not only in molecular pathological epidemiology (MPE) (32) with bona fide molecular subclassification, but also in epidemiologic research which takes other features of disease heterogeneity (e.g., lethality, disease severity) into consideration. There are statistical methods for evaluating whether the association of an exposure with disease varies by subtypes which are defined by categorical (33-36) or ordinal (33-35) subclassifiers (reviewed by Wang *et al*., unpublished); the published methods by Chatterjee (33), Chatterjee *et al*. (34), and Rosner *et al*. (35) apply to cohort studies, and the method by Begg *et al*. (36) focuses on case-control studies. For simplicity, we use the term “categorical variable” (or the adjective “categorical”) referring to “non-ordinal categorical variable” throughout this paper. Given the complexity of molecular pathology and pathogenesis indicated by the unique disease principle (1), no single biomarker can perfectly subclassify any disease entity. Notably, molecular disease markers are often correlated (37). For example, in colorectal cancer, there is a strong association between high-level microsatellite instability (MSI-high) and high-level CpG island methylator phenotype (CIMP-high), and between CIMP-high and the *BRAF* mutation (38). Cigarette smoking has been associated with the risk of MSI-high colorectal cancer (16-18, 20, 39-42), CIMP-high colorectal cancer (17, 20, 42, 43), and *BRAF* mutated colorectal cancer (17, 19, 20, 42). Given the correlations between these molecular markers, the association of smoking with a subtype defined by one marker may solely (or in part) reflect the association with a subtype defined by another marker. Thus, it remains unclear which molecular marker subtypes are primarily differentially associated with smoking, and how it can confound the association between smoking and subtypes defined by other markers. Although the published methods (33-35) are useful to analyze the exposure-subtype associations according to multiple subtyping markers in cohort studies using existing statistical software, analysis using those methods can become computationally infeasible in large datasets. In this article, we present an intuitive and computationally efficient biostatistical method for the analysis of disease and etiologic heterogeneity when there are multiple disease subtyping markers (categorical and/or ordinal), which are possibly but not necessarily correlated.

METHODS

Cohort and nested case-control studies

In cohort studies where age at disease onset is available, a commonly used statistical model for evaluating subtype-specific exposure-disease associations is the cause-specific hazards model (44, 45):

$$\lambda_j(t|\boldsymbol{X}_i(t), \boldsymbol{W}_i(\boldsymbol{t})) = \lambda_{0j}(t)\exp\{\boldsymbol{\beta}_{1j}\boldsymbol{X}_i(t) + \boldsymbol{\beta}_{2j}\boldsymbol{W}_i(\boldsymbol{t})\}, \qquad [1]$$

where $\lambda_j(t)$ is the incidence rate at age t for subtype j, $\lambda_{0j}(t)$ is the baseline incidence rate for subtype j, $\boldsymbol{X}_i(t)$ is a possibly time-varying column vector of exposure variables for the ith individual, $\boldsymbol{W}_i(t)$ is a possibly time-varying column vector of potential confounders, and $\boldsymbol{\beta}_{1j}$ and $\boldsymbol{\beta}_{2j}$ are row vector-valued log relative risks (RRs) for the corresponding covariates for subtype j. Model 1 can be estimated in cohort studies and incidence density-sampled case-control studies (46). Assume J subtypes are resulted from cross classification of multiple categorical and/or ordinal markers. We create binary indicators for categorical markers; thus, hereafter, we treat the marker variables as either binary or ordinal. Let s_{pj} denote the level of the pth marker variable corresponding to the jth subtype; it is 1 or 0 if the pth marker variable is binary, and is the ordinal or median score of the marker level corresponding to the jth subtype if the pth marker is an ordinal marker, $p = 1, \dots, P, j = 1, \dots J$.

One-stage method. The method developed by Rosner, *et al*. (35), Chatterjee (33) and Chatterjee, *et al*. (34) can be usefully applied in cohort studies to study multiple markers. In that method, $\boldsymbol{\beta}_{1j}$ in model 1 is modeled using the marker variables, for example, by $\boldsymbol{\beta}_{1j}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}_0 + \sum_{p=1}^{P} \boldsymbol{\gamma}_p s_{pj}$, where some interaction terms of marker variables can be added. Model 1 then becomes

$$\lambda_j(t|\boldsymbol{X}_i(t), \boldsymbol{W}_i(t)) = \lambda_{0j}(t) \exp\{\boldsymbol{\gamma}_0 \boldsymbol{X}_i(t) + \sum_{p=1}^{P} \boldsymbol{\gamma}_p s_{pj} \boldsymbol{X}_i(t) + \boldsymbol{\beta}_{2j} \boldsymbol{W}_i(t)\}. \quad [2]$$

To distinguish this method from the proposed two-stage one below, we name it "one-stage method". The parameters of interest, $\boldsymbol{\gamma}_0$ and each $\boldsymbol{\gamma}_p$, which have the same dimension as $\boldsymbol{\beta}_{1j}$, characterize how the levels of multiple markers are associated with differential exposure associations. We can obtain the maximum partial likelihood estimate (33, 34) of $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_p, p = 1, \dots, P\}$ using the existing statistical software for the Cox model analysis, such as PROC PHREG in SAS, through the data duplication method (47), which is based on the following transformation of model 2:

$$\lambda_j(t|\boldsymbol{X}_i(t), \boldsymbol{W}_i(t)) = \lambda_{0j}(t) \exp\{\boldsymbol{\gamma}_0 \boldsymbol{X}_i(t) + \sum_{p=1}^{P} \boldsymbol{\gamma}_p \widetilde{\boldsymbol{X}}_{pji}(t) + \sum_{l=1}^{J} \boldsymbol{\beta}_{2l} \boldsymbol{W}_{li}(t)\},$$

where $\widetilde{\boldsymbol{X}}_{pji}(t) = s_{pj} \boldsymbol{X}_i(t)$, $\boldsymbol{W}_{li}(t) = \boldsymbol{W}_i(t)$ for $l = j$, and $\boldsymbol{W}_{li}(t) = \mathbf{0}$ for $l \neq j$. In this data duplication method, model 2 can be fit using the stratified Cox regression (stratified by subtype) on an augmented data set, in which, each block of person-time is augmented for each subtype, and variables $\widetilde{\boldsymbol{X}}_{pji}$ and $\boldsymbol{W}_{ji}$ are created for $p = 1, \dots P$, $j = 1, \dots, J$. Rosner, *et al*. (35) also proposed an adjusted RR for the exposure-disease association for a disease subtype defined by one or some marker(s) while adjusting for other markers. The data duplication method may become computationally infeasible when the augmented dataset becomes too large; this can easily happen when the original data set is sizable and the number of subtypes cross-classified from the multiple markers is large. For example, in our colorectal example, there are $3{,}099{,}586$ rows in our original data set. With $P = 3$ and $J = 8$, in the augmented data set, there will be about $3{,}099{,}586 \times 8 = 24{,}796{,}688$ rows, $P \times J = 24$ new variables created for each exposure variable, and $J = 8$ variables created for each confounding variable. If considering more markers, the large augmented dataset can easily make the Cox model analysis computationally infeasible.

Two-stage method. When subtypes are defined by multiple categorical and/or ordinal markers, we propose a meta-regression method that is intuitive, does not need augmentation of the dataset and can be easily implemented using existing statistical software for the mixed model analysis. We first assume the exposure variable $X_i(t)$ in Model 1 is scalar. This includes the situations in which the exposure is continuous or binary, and the trend analysis for categorical exposure in which a new continuous variable, median level in each exposure category, is included in model 1. The meta-regression method includes two stages of analysis. The first stage is to conduct the subtype-specific analysis for each cross-classified subtype from the multiple markers. For the cohort and nested case-control study, this analysis can be based on Model 1. Typically, a standard competing risks framework can be used, where it is assumed that only one disease subtype can be observed in each individual. The occurrence of a disease subtype that is different from the subtype for which the exposure association is studied is censored at the date of diagnosis. The model for the second stage analysis is

$$\hat{\beta}_{1j} = \gamma_0 + \sum_{p=1}^{P} \gamma_p s_{pj} + e_j, \qquad [3]$$

where $\hat{\beta}_{1j}$, the estimated $\log(RR)$ representing the exposure association with the jth subtype, is obtained in the first stage analysis, and e_j are within-study sampling errors; that is, $Var(e_j) = \widehat{Var}(\hat{\beta}_{1j})$. Since, in the competing risk framework, the relative risks for distinct tumor subtypes are asymptotically uncorrelated (45), this meta-regression for J subtypes is the same as the standard meta-regression for J independent studies. Interactions of s_{pj} can be included as covariates in model 3 if appropriate. We can use the Wald test to test the hypothesis $H_0: \gamma_p = 0,$ for each p. This null hypothesis implies that the exposure-subtype association does not change over the level of the pth marker variable while controlling for the other marker variables. For a categorical marker, we can also test whether $\gamma_p = 0$ for all p's corresponding to the binary marker variables created for this categorical marker; the null hypothesis implies that the categorical marker does not contribute to the possible etiologic heterogeneity. Note that the difference between this two-stage method with a fixed effects meta-regression model and the one-stage method is essentially only in the estimation method, not the model.

We can also add subtype-specific random effects in model 3 to account for heterogeneity between subtypes that cannot be explained by variables in model 3. Below is a random effects meta-regression model (48),

$$\hat{\beta}_{1j} = \gamma_0 + \sum_{p=1}^{P} \gamma_p s_{pj} + b_j + e_j, \qquad [4]$$

where $b_j \sim N(0, \sigma_b^2)$ are subtype-specific random effects accounting for heterogeneity between the subtypes that cannot be explained by variables $s_{pj,}$ and e_j, assumed independent of b_j, has the same definition as in model 3. This random effects two-stage method uses a different model from the fixed effects two-stage and the one-stage methods. It has the advantage over both the fixed effects two-stage method and the one-stage method that it can incorporate additional heterogeneity between subtypes that cannot be explained by the given marker variables. If $\sigma_b^2 = 0$, where model 4 agrees with model 3, the random effects meta-regression model method is typically less efficient than the fixed effects method, and since the one-stage method is a maximum likelihood method, it should be the most efficient among the three methods. In the random effects model, the test $H_0\colon \sigma_b^2 = 0$ assesses the significance of the random effects term. Note that when the number of subtypes is small, this test may be underpowered and the estimate of σ_b^2 may be imprecise. When the test rejects $H_0\colon \sigma_b^2 = 0$ or when we believe there is heterogeneity in addition to those explained by the marker variables, we may use the random effects model in the two-stage method.

Unmatched case-control study

In the unmatched case-control design, the first-stage model of the two-stage method can be the nominal polytomous logistic regression

$$P(Y_i = j|X_i, \boldsymbol{W}_i)/P(Y_i = 0|X_i, \boldsymbol{W}_i) = \exp\left(\beta_{0j} + \beta_{1j}X_i + \boldsymbol{\beta}_{2j}\boldsymbol{W}_i\right),\ j = 1, \dots, J,$$

where $Y = j$ represents subtype j cases, $Y = 0$ represents controls, and β_{1j} represents the subtype-specific log odds ratio (OR), assumed to be a scalar. The scenarios where the exposure is a vector will be considered in a later section. If the disease is rare, $\exp\left(\beta_{1j}\right)$ approximates RR. In this design, the subtype-specific association estimates, $\hat{\beta}_{11,} \dots, \hat{\beta}_{1J}$, are typically correlated. The second stage model of the two-stage method is the meta-regression model 3 or 4 with an additional condition: $Cov(e_{j_1}, e_{j_2}) = \widehat{Cov}(\hat{\beta}_{1j_1,}\hat{\beta}_{1j_2})$. R function rma.mv() can be used to estimate $\hat{\gamma}_p, p = 1, \dots P,$ in models 3 and 4 and the variance of $\hat{\gamma}_p$ (49). We can then use the Wald test to test the hypothesis $H_0: \gamma_p = 0$ for each p, or test whether $\gamma_p = 0$ for all p's corresponding to the binary marker variables created for a categorical marker.

Interaction between markers

The adjusted $\widehat{RR}$ proposed by Rosner, *et al*. (35) can also be estimated in models 3 and 4. For example, if there are two binary markers, cross-classification of which defines 4 subtypes, and the second stage model of the fixed effects meta-regression method is $\hat{\beta}_{1j} = \gamma_0 + \gamma_1 s_{1j} + \gamma_2 s_{2j} + e_j,\ j = 1, \dots, 4,$ where γ_p represents the difference in exposure-disease subtype associations between the two subtypes defined by the pth marker while the level of the other marker is the same, $p = 1, 2$. The meta-regression method can also be used to evaluate whether the difference in exposure-disease subtype association across the subtypes defined by one marker depends on the level of another marker by including appropriate interaction terms for these markers in the meta-regression model. For example, in the second stage fixed effects model $\hat{\beta}_{1j} = \gamma_0 + \gamma_1 s_{1j} + \gamma_2 s_{2j} + \gamma_3 s_{1j} \times s_{2j} + e_j,$ rejection of the null hypothesis $H_0: \gamma_3 = 0$ implies that the difference in exposure-disease subtype associations across the subtypes defined by the first marker depends on the level of the second marker. The discussion above, which is for the fixed effects two-stage method, can be easily extended to the random effects method.

Categorical exposures and multiple exposures

Let $\boldsymbol{\beta}_{1j} = (\beta_{1j1}, \ldots, \beta_{1jK}), K > 1$, represent the subtype-specific exposure-disease association corresponding to binary indicators created for a categorical exposure with $K + 1$ levels, or multiple exposures, one or more of which could be categorical exposures, for which binary indicators are created. The first stage analysis of the two-stage method, which is the subtype-specific analysis for each cross-classified subtype, is the same as in the cases when β_{1j} is a scalar. At the second stage, one strategy is to conduct the meta-regression analysis for each element of $\boldsymbol{\beta}_{1j}$ separately. For the kth element of $\boldsymbol{\beta}_{1j}$, the random effects meta-regression model $\hat{\beta}_{1jk} = \gamma_{0k} + \sum_{p=1}^{P} \gamma_{pk} s_{pj} + b_{jk} + e_{jk}$, or the fixed effects meta-regression model, which does not include the random effects term b_{jk}, may be used to characterize the relationship between β_{1jk} and levels of the multiple markers. For an any given k, in cohort and nested case-control studies, e_{jk}'s, $j = 1, \ldots J$, are independent, and in unmatched case-control studies, $cov(e_{j_1k}, e_{j_2k}) = cov(\hat{\beta}_{1j_1k}, \hat{\beta}_{1j_2k})$.

Alternatively, the second stage model can be a random effects multivariate meta-regression model (50, 51)

$$\begin{pmatrix} \hat{\beta}_{1j1} \\ \cdots \\ \hat{\beta}_{1jK} \end{pmatrix} = \begin{pmatrix} r_{01} \\ \cdots \\ r_{0K} \end{pmatrix} + \sum_{p=1}^{P} \begin{pmatrix} r_{p1} \\ \cdots \\ r_{pK} \end{pmatrix} s_{pj} + \boldsymbol{b}_j + \boldsymbol{e}_j, \qquad [5]$$

where the error term $\boldsymbol{e}_j = (e_{j1}, \ldots, e_{jK})$ is a K- dimension normal distribution with $cov(e_{jk_1}, e_{jk_2}) = \widehat{Cov}(\hat{\beta}_{1jk_1}, \hat{\beta}_{1jk_2})$ for $k_1 \neq k_2$, and $var(e_{jk}) = \widehat{Var}(\hat{\beta}_{1jk})$. In cohort and nested case-control studies, $cov\left(e_{j_1k_1}, e_{j_2k_2}\right) = 0$, and for unmatched case-control studies, $cov\left(e_{j_1k_1}, e_{j_2k_2}\right) = cov\left(\ \hat{\beta}_{1j_1k_1},\ \hat{\beta}_{1j_2k_2}\right)$, for $j_1 \neq j_2, k_1, k_2 = 1, ..K$. The random effects term $\boldsymbol{b}_j$ is a K- dimension normal distribution with mean zero, independent from $\boldsymbol{e}_j$. The fixed effects multivariate meta-regression model is model 5 with $\boldsymbol{b}_j$ excluded. As pointed out in (50, 51), the estimator of r_{pk} using the multivariate random effects meta-regression method is more efficient than that from the univariate random effects meta-regression method presented above. Presumably the same conclusion can be made on the fixed effects models. R function rma.mv() can be used to estimate $\hat{\gamma}_{pk}$ in the random effects and fixed effects multivariate meta-regression models.

EXAMPLE

To illustrate the proposed meta-regression method for multiple markers, we examine the associations between smoking status (never, former, current) and 8 possible colorectal cancer subtypes defined by three binary markers, CIMP (high vs. low/negative), MSI (high vs. microsatellite stable (MSS)) and *BRAF* (mutant vs. wild-type). The smoking status is coded as 0 for never, 1 for former, and 2 for current, and the trend association is examined. The analysis includes 88,620 women in the Nurses' Health Study (NHS) and 46,251 men in the Health Professionals Follow-Up Study (HPFS), with 3,099,586 person-years of follow-up. In each cohort, one subtype with fewer than 5 cases (CIMP-low/negative, MSI-high, *BRAF*-mutated) was excluded, leading to a total of 1118 colorectal cancer cases (654 women in NHS, and 464 men in HPFS) in the remaining 7 subtypes.

In the first stage of the two-stage meta-regression approach, a subtype-specific multivariate Cox model analysis, stratified by age (months) and calendar year of the questionnaire cycle, and adjusted for potential confounders, was performed for each cohort. Table 1 contains subtype definitions, subtype-specific case numbers, and the estimated smoking status - colorectal cancer subtype associations in the NHS and HPFS. In the second stage analysis, we modeled the subtype and cohort-specific log(RR) using the three markers considered, MSI, CIMP and *BRAF,* and cohort (NHS vs. HPFS), and compared the results with those from the one-stage method (33-35); in the one-stage method, we conducted the Cox model analysis for each cohort using the data duplication method, and then combined the estimates from NHS and HPFS by the fixed effects meta-analysis approach. Table 2 shows inferences for exponential of the coefficients of the marker variables in the model for log(RR) which represent the ratios of RRs (RRR) between marker levels. For example, based on the meta-regression method, the estimated ratio of the RR for the association of smoking with CIMP-high colorectal cancer over the RR for CIMP-low/negative colorectal cancer, while the MSI and *BRAF* levels stay the same, was 1.23 (95% confidence interval: 0.84, 1.82). As shown in Table 2, the results from these two methods were consistent. These analysis results suggest that we do not have sufficient statistical evidence to conclude that the smoking - colorectal cancer subtype associations are different across subtypes defined by any one of the biomarkers (MSI, CIMP and *BRAF*) while controlling for the other two biomarkers.

In a second analysis for illustrating the proposed meta-regression method, the first stage analysis was the same as before, but in the second stage, we started from a model with all three markers, two-way interactions of the markers, and cohort, and then used stepwise model selection with a cutoff p-value of 0.05 for entering or removing the variables. This analysis was for selecting covariates in the meta-regression model that are important for characterizing the subtype-specific exposure-disease association. Only MSI was in the final model (RRR for MSI-high versus MSS = 1.38; 95% confidence interval: 1.07, 1.79; p-value = 0.015).

DISCUSSION

When subtypes are defined by multiple categorical and/or ordinal markers, we propose a meta-regression method that is intuitive, does not need augmentation of the dataset and can be easily implemented using existing statistical software such as SAS procedures for the mixed model analysis. This meta-regression method can be used to test for etiologic heterogeneity across multiple disease subtypes classified by multiple markers, to assess whether the exposure-disease subtype associations are different across subtypes defined by one marker while controlling for other markers, and to evaluate whether the difference in exposure-disease subtype association across subtypes by one marker depends on any of other markers.

Addressing etiologic heterogeneity by MPE research has relevance to disease prevention. As an example, we herein discuss smoking, colonoscopy and colorectal cancer risk. Colonoscopy has been associated with lower colorectal cancer risk for up to 10 years after the procedure in individuals with average risk for developing colorectal cancer (52); however, it remains to be determined whether colonoscopy every 10 years is also effective for colorectal cancer prevention in high-risk individuals. A recent MPE study suggests that preventive effect of colonoscopy may be weaker for MSI-high colorectal cancer than for non-MSI-high colorectal cancer (52). MPE studies (16-18, 20, 39-42) have also shown that smokers are susceptible to developing MSI-high colorectal cancer. Taken together, it is implied that preventive effect of colonoscopy is not as effective for smokers compared to non-smokers. Hence, MPE research can help us towards more personalized disease prevention strategies.

In addition to heterogeneity between tumors across individuals, accumulating evidence has indicated heterogeneity within one tumor in one individual. An integrative concept ("the unique tumor principle") on intra- and inter-tumor heterogeneity along with epidemiologic exposures has been discussed in detail (53). Though our current paper primarily addresses inter-tumor (or inter-individual) heterogeneity, it is of our interest to develop new statistical methodologies to address both intra- and inter-tumor heterogeneity in the future.

As advancements of biomedical technologies, molecular pathology tests are increasingly common in clinical practice as well as epidemiologic studies (54-56). The MPE approach is useful not only for assessment of risk of developing disease but also for evaluation of predictive biomarkers for intervention in a disease population (57). In the future, routine clinical molecular pathology data may be integrated into population-based disease registries and databases, and large-scale MPE studies can be routine research practice (58). Thus, our methodology will be widely useful.

We developed a user-friendly SAS macro %stepmetareg implementing this meta-regression method. It includes a stepwise selection procedure to select covariates considered in the meta-regression model that are important for characterizing the subtype-specific exposure-disease association, represented by $\widehat{\boldsymbol{\beta}}_{1j}$. The SAS macro can be obtained at the website, http://www.hsph.harvard.edu/donna-spiegelman/software/

This meta-regression method will be most useful in situations where the number of subtypes is relatively low; otherwise, the number of cases for each unique tumor subtype defined by cross-classification of the multiple markers may be too small to obtain stable estimates of each $\boldsymbol{\beta}_{1j}$. The minimum number of cases required for each tumor subtype for obtaining stable estimates of each $\boldsymbol{\beta}_{1j}$ depends on the number of covariates in the first-stage model. A rule of thumb for the minimum events per covariate is 5 to 10. An advantage of the proposed two-stage method for cohort studies is that $\widehat{\boldsymbol{\beta}}_{1j}$, $j = 1, \dots, J$, can be estimated separately, without using the data duplication method, which becomes computationally infeasible when the augmented dataset becomes too large. In addition, the random effects model has the advantage that it can incorporate additional heterogeneity between subtypes that cannot be explained by the given marker variables.

Disease subtype data are often missing in some proportion of cases. Chatterjee, *et al*. (34) developed an estimating function method based on model 2 which can be used to handle missing subtype data under a missing-at-random assumption. That method can be used directly to handle missing subtype data for estimating $\boldsymbol{\beta}_{1j}$ in the first stage of the two-stage models. Statistical methods for handling missing marker data, which are covariates data now, in the second stage model of the two-stage method may be developed through extension of existing methods for missing covariates data problems in the mixed model analysis; this is a topic of future research. Alternatively, we may use the conventional method of creating missing indicators for missing markers data, and the method of imputing the missing marker data based on regression models that link the marker data and covariates that contain information about the marker data. While using these methods, the two-stage method with a random effect meta-regression model could have the advantage of partially taking into account additional variability due to using missing indicators or using imputed marker data through the random effect term; future research is needed for this topic.

In conclusion, in consideration of pathogenesis and etiologic heterogeneity of disease, we developed a meta-regression method to study etiologic heterogeneity across disease subtypes defined by multiple biomarkers. This method is useful in the emerging interdisciplinary field of molecular pathological epidemiology (32, 59). There is an increasing need to integrate molecular pathology and epidemiology to better understand disease etiologies and causalities (59-62). Our meta-regression method can be widely useful, as use of molecular pathology and genomic technologies is increasingly common in clinical medicine and public health.

ACKNOWLEDGMENTS

We deeply thank hospitals and pathology departments throughout the U.S. for generously providing us with tissue specimens. We also would like to thank 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.

This work was supported by U.S. National Institutes of Health grants (R01 CA151993 to S.O.; P01 CA87969 to M.J. Stampfer; UM1 CA186107 to M.J. Stampfer; P01 CA55075 to W.C. Willett; UM1 CA167552 to W.C. Willett; P50 CA127003 to C.S. Fuchs), grants from The Paula and Russell Agrusa Fund for Colorectal Cancer Research (to C.S. Fuchs), and the Friends of the Dana-Farber Cancer Institute (to S.O.).

Conflict of interest: none declared.

Author affiliations: Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Molin Wang, Shuji Ogino); Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Molin Wang); Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts (Molin Wang); Biostatistics Division, Center for Research Administration and Support, National Cancer Center, Tokyo, Japan (Aya Kuchiba); Department of Pathology, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts (Shuji Ogino); Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts (Shuji Ogino)

REFERENCES

1. Ogino S, Lochhead P, Chan AT, et al. Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease.

Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2013;26(4):465-84.

2. Ogino S, Chan AT, Fuchs CS, et al. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. *Gut* 2011;60(3):397-411.
3. Chen H, Taylor NP, Sotamaa KM, et al. Evidence for heritable predisposition to epigenetic silencing of MLH1. *Int J Cancer* 2007;120(8):1684-8.
4. Allan JM, Shorto J, Adlard J, et al. MLH1 -93G>A promoter polymorphism and risk of mismatch repair deficient colorectal cancer. *Int J Cancer* 2008;123(10):2456-9.
5. Campbell PT, Curtin K, Ulrich CM, et al. Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle factors. *Gut* 2009;58(5):661-7.
6. Raptis S, Mrkonjic M, Green RC, et al. MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. *J Natl Cancer Inst* 2007;99(6):463-74.
7. Samowitz WS, Curtin K, Wolff RK, et al. The MLH1 -93 G>A promoter polymorphism and genetic and epigenetic alterations in colon cancer. *Genes Chromosomes Cancer* 2008;47(10):835-44.
8. Ogino S, Hazra A, Tranah GJ, et al. MGMT germline polymorphism is associated with somatic MGMT promoter methylation and gene silencing in colorectal cancer. *Carcinogenesis* 2007;28(9):1985-90.
9. Hawkins NJ, Lee JH, Wong JJ, et al. MGMT methylation is associated primarily with the germline C>T SNP (rs16906252) in colorectal cancer and normal colonic mucosa.

Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2009;22(12):1588-99.

10. Slattery ML, Curtin K, Anderson K, et al. Associations between cigarette smoking, lifestyle factors, and microsatellite instability in colon tumors. *J Natl Cancer Inst* 2000;92(22):1831-6.

11. Satia JA, Keku T, Galanko JA, et al. Diet, lifestyle, and genomic instability in the north Carolina colon cancer study. *Cancer Epidemiol Biomarkers Prev* 2005;14(2):429-36.

12. Slattery ML, Curtin K, Sweeney C, et al. Diet and lifestyle factor associations with CpG island methylator phenotype and BRAF mutations in colon cancer. *Int J Cancer* 2007;120(3):656-63.

13. Campbell PT, Jacobs ET, Ulrich CM, et al. Case-control study of overweight, obesity, and colorectal cancer risk, overall and by tumor microsatellite instability status. *J Natl Cancer Inst* 2010;102(6):391-400.

14. Kuchiba A, Morikawa T, Yamauchi M, et al. Body mass index and risk of colorectal cancer according to fatty acid synthase expression in the nurses' health study. *J Natl Cancer Inst* 2012;104(5):415-20.

15. Wu AH, Shibata D, Yu MC, et al. Dietary heterocyclic amines and microsatellite instability in colon adenocarcinomas. *Carcinogenesis* 2001;22(10):1681-4.

16. Chia VM, Newcomb PA, Bigler J, et al. Risk of microsatellite-unstable colorectal cancer is associated jointly with smoking and nonsteroidal anti-inflammatory drug use. *Cancer Res* 2006;66(13):6877-83.

17. Samowitz WS, Albertsen H, Sweeney C, et al. Association of smoking, CpG island methylator phenotype, and V600E BRAF mutations in colon cancer. *J Natl Cancer Inst* 2006;98(23):1731-8.

18. Poynter JN, Haile RW, Siegmund KD, et al. Associations between smoking, alcohol consumption, and colorectal cancer, overall and by tumor microsatellite instability status. *Cancer Epidemiol Biomarkers Prev* 2009;18(10):2745-50.

19. Rozek LS, Herron CM, Greenson JK, et al. Smoking, gender, and ethnicity predict somatic BRAF mutations in colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 2010;19(3):838-43.

20. Limsui D, Vierkant RA, Tillmans LS, et al. Cigarette Smoking and Colorectal Cancer Risk by Molecularly Defined Subtypes. *J Natl Cancer Inst* 2010;102(14):1012-22.

21. Leng S, Bernauer AM, Hong C, et al. The A/G Allele of Rs16906252 Predicts for MGMT Methylation and Is Selectively Silenced in Premalignant Lesions from Smokers and in Lung Adenocarcinomas. *Clin Cancer Res* 2011;17(7):2014-23.

22. Ahrendt SA, Decker PA, Alawi EA, et al. Cigarette smoking is strongly associated with mutation of the K-ras gene in patients with primary adenocarcinoma of the lung. *Cancer* 2001;92(6):1525-30.

23. Riely GJ, Kris MG, Rosenbaum D, et al. Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. *Clin Cancer Res* 2008;14(18):5731-4.

24. Riely GJ, Marks J, Pao W. KRAS mutations in non-small cell lung cancer. *Proc Am Thorac Soc* 2009;6(2):201-5.

25. Julin B, Bergkvist C, Wolk A, et al. Cadmium in diet and risk of cardiovascular disease in women. *Epidemiology* 2013;24(6):880-5.

26. Jeong I, Rhie J, Kim I, et al. Working Hours and Cardiovascular Disease in Korean Workers: A Case-control Study. *J Occup Health* 2013.

27. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. *Pediatrics* 2013.

28. Dandri M, Locarnini S. New insight in the pathobiology of hepatitis B virus infection. *Gut* 2012;61 Suppl 1:i6-17.

29. Perez OD. Appreciating the heterogeneity in autoimmune disease: multiparameter assessment of intracellular signaling mechanisms. *Ann N Y Acad Sci* 2005;1062:155-64.

30. Takamoto M, Kaburaki T, Mabuchi A, et al. Common variants on chromosome 9p21 are associated with normal tension glaucoma. *PLoS One* 2012;7(7):e40107.

31. Field AE, Camargo JCA, Ogino S. The merits of subtyping obesity: one size does not fit all. *The Journal of American Medicial Association* 2013;310(20):2147-8.

32. Ogino S, Stampfer M. Lifestyle factors and microsatellite instability in colorectal cancer: the evolving field of molecular pathological epidemiology. *J Natl Cancer Inst* 2010;102(6):365-7.

33. Chatterjee N. A Two-Stage Regression Model for Epidemiological Studies With Multivariate Disease Classification Data. *Journal of the American Statistical Association* 2004;99(465):127-38.

34. Chatterjee N, Sinha S, Diver WR, et al. Analysis of cohort studies with multivariate and partially observed disease classification data. *Biometrika* 2010;97(3):683-98.

35. Rosner B, Glynn RJ, Tamimi RM, et al. Breast Cancer Risk Prediction with Heterogeneous Risk Profiles According to Breast Cancer Tumor Markers. *Am J Epidemiol* 2013;15:296-308.

36. Begg CB, Zabor EC, Bernstein JL, et al. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med* 2013;32(29):5039-52.

37. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. *J Mol Diagn* 2008;10(1):13-27.

38. 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(11):e3698.

39. Lindor NM, Yang P, Evans I, et al. Alpha-1-antitrypsin deficiency and smoking as risk factors for mismatch repair deficient colorectal cancer: A study from the colon cancer family registry. *Mol Genet Metab* 2010;99(2):157-9.

40. Phipps AI, Baron J, Newcomb PA. Prediagnostic smoking history, alcohol consumption, and colorectal cancer survival: the Seattle Colon Cancer Family Registry. *Cancer* 2011;117(21):4948-57.

41. Eaton AM, Sandler R, Carethers JM, et al. 5,10-methylenetetrahydrofolate reductase 677 and 1298 polymorphisms, folate intake, and microsatellite instability in colon cancer. *Cancer Epidemiol Biomarkers Prev* 2005;14(8):2023-9.

42. Nishihara R, Morikawa T, Kuchiba A, et al. A prospective study of duration of smoking cessation and colorectal cancer risk by epigenetics-related tumor classification. *Am J Epidemiol* 2013;178(1):84-100.

43. Curtin K, Samowitz WS, Wolff RK, et al. Somatic alterations, metabolizing genes and smoking in rectal cancer. *Int J Cancer* 2009;125(1):158-64.

44. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. Wiley; 1980.

45. Prentice RL, Kalbfleisch JD, Peterson AV, Jr., et al. The analysis of failure times in the presence of competing risks. *Biometrics* 1978;34(4):541-54.

46. Prentice RL. On the design of synthetic case-control studies. *Biometrics* 1986;42(2):301-10.

47. Lunn M, McNeil D. Applying Cox regression to competing risks. *Biometrics* 1995;51(2):524-32.

48. Stram DO. Meta-analysis of published data using a linear mixed-effects model. *Biometrics* 1996;52(2):536-44.

49. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 2010;36(3):1-48.

50. Ritz J, Demidenko E, Spiegelman D. Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. *J Stat Plann Inference* 2008;138(7):1919-33.

51. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002;21(4):589-624.

52. Nishihara R, Wu K, Lochhead P, et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. *The New England journal of medicine* 2013;369(12):1095-105.

53. Ogino S, Fuchs CS, Giovannucci E. How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. *Expert review of molecular diagnostics* 2012;12(6):621-8.

54. Colussi D, Brandi G, Bazzoli F, et al. Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. *International journal of molecular sciences* 2013;14(8):16365-85.

55. Phipps AI, Limburg PJ, Baron JA, et al. Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology* 2015;148(1):77-87 e2.

56. Caiazza F, Ryan EJ, Doherty G, et al. Estrogen receptors and their implications in colorectal carcinogenesis. *Frontiers in Oncology* 2015;5(19):doi:10.3389/fonc.2015.00019.

57. Liao X, Lochhead P, Nishihara R, et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. *The New England journal of medicine* 2012;367(17):1596-606.

58. Ogino S, Lochhead P, Giovannucci E, et al. Discovery of colorectal cancer PIK3CA mutation as potential predictive biomarker: power and promise of molecular pathological epidemiology. *Oncogene* 2014;33(23):2949-55.

59. Ogino S, King EE, Beck AH, et al. Interdisciplinary education to integrate pathology and epidemiology: towards molecular and population-level health science. *Am J Epidemiol* 2012;176(8):659-67.

60. Campbell PT, Deka A, Briggs P, et al. Establishment of the Cancer Prevention Study II Nutrition Cohort Colorectal Tissue Repository. *Cancer Epidemiol Biomarkers Prev* 2014;23(12):2694-702.

61. Wild CP, Bucher JR, de Jong BW, et al. Translational cancer research: balancing prevention and treatment to combat cancer globally. *J Natl Cancer Inst* 2015;107(1):353.

62. Kuller LH, Bracken MB, Ogino S, et al. The role of epidemiology in the era of molecular epidemiology and genomics: Summary of the 2013 AJE-sponsored Society of Epidemiologic Research Symposium. *Am J Epidemiol* 2013;178(9):1350-4.

Table 1. Subtype definitions, subtype-specific case numbers and estimated smoking status - colorectal cancer subtype associations.[a]

Subtype	CIMP	MSI	*BRAF*	N of cases	RR	95% CI[b]	P value[b]
1	L/N	MSS	Wild-type	832	1.12	1.01,1.25	0.039
2	L/N	MSS	Mutant	47	0.86	0.54,1.37	0.53
3	L/N	High	Wild-type	42	1.35	0.80,2.25	0.26
4	High	MSS	Wild-type	34	1.28	0.71,2.32	0.41
5	High	MSS	Mutant	31	1.00	0.57,1.78	0.99
6	High	High	Wild-type	43	1.93	1.18,3.14	0.008
7	High	High	Mutant	95	1.45	1.05,2.00	0.026

Abbreviations: CI, confidence interval; CIMP, CpG island methylator phenotype; L/N, low/negative; MSI, microsatellite instability; MSS, microsatellite stable; RR, relative risk.

[a] The analysis includes only subtypes with ≥ 5 cases. The subtype-specific analyses were controlled for body mass index (<25, 25-29.9, ≥30 kg/m^2), family history of colorectal cancer (yes/no), physical activity in metabolic equivalent of tasks (quintiles), red meat intake (quintiles of servings per day), alcohol consumption (0, quartiles of grams per day), total caloric intake (quintiles of calories per day) , regular aspirin use (2 or more tablets per week or at least 2 times per week/less) and stratified by age (month), calendar year. Postmenopausal hormone use (never/ever) is also adjusted in NHS.

[b] The cohort-specific estimates were combined using a fixed effects meta-analysis method.

Table 2. Results from modeling the smoking status - colorectal cancer subtype association using three markers

	Two-stage approach			One–stage approach		
Subtype by	RRR	95% CI	P value	RRR	95% CI	P value
CIMP	1.23	0.84 – 1.82	0.29	1.28	0.87 – 1.88	0.21
MSI	1.34	0.93 – 1.91	0.11	1.31	0.92 – 1.87	0.13
BRAF	0.78	0.55 – 1.09	0.14	0.78	0.56 – 1.10	0.16

Abbreviations: CI, confidence interval; CIMP, CpG island methylator phenotype; MSI, microsatellite instability; RRR, ratio of relative risks.