HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2017 February 28.

Published in final edited form as:

Stat Med. 2016 February 28; 35(5): 782–800. doi:10.1002/sim.6793.

Statistical Methods for Studying Disease Subtype Heterogeneity

Molin Wang[a,b,e,†,*], **Donna Spiegelman**[a,b,c,d], **Aya Kuchiba**[f], **Paul Lochhead**[h], **Sehee Kim**[g], **Andrew T. Chan**[e,h], **Elizabeth M. Poole**[e], **Rulla Tamimi**[b,e], **Shelley S. Tworoger**[b,e], **Edward Giovannucci**[b,c,e], **Bernard Rosner**[a,e], **Shuji Ogino**[b,i,j,*,†]

[a]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A

[b]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A

[c]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A

[d]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A

[e]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, U.S.A

[f]Department of Biostatistics, National Cancer Center, Tokyo, Japan

[g]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, U.S.A

[h]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, U.S.A

[i]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, U.S.A

[j]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, U.S.A

Abstract

A fundamental goal of epidemiologic research is to investigate the relationship between exposures and disease risk. Cases of the disease are often considered a single outcome, and assumed to share a common etiology. However, evidence indicates that many human diseases arise and evolve through a range of heterogeneous molecular pathologic processes, influenced by diverse exposures. Pathogenic heterogeneity has been considered in various neoplasms such as colorectal, lung, prostate, and breast cancers, leukemia and lymphoma, as well as non-neoplastic diseases, including obesity, type II diabetes, glaucoma, stroke, cardiovascular disease, autism and autoimmune disease. In this article, we discuss analytic options for studying disease subtype heterogeneity, emphasizing methods for evaluating whether the association of a potential risk factor with disease varies by disease subtype. Methods are described for scenarios where disease subtypes are categorical and ordinal, and for cohort studies, matched and unmatched case-control studies, and case-case study designs. For illustration, we apply the methods to a molecular

[*]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, MA, U.S.A. Shuji Ogino, Department of Epidemiology, Harvard T.H. Chan School of Public Health, and Department of Medical Oncology, Dana-Farber Cancer Institute, 450 Brookline Ave., Room M422, Boston, MA 02215 U.S.A.
[†]stmow@channing.harvard.edu; shuji_ogino@dfci.harvard.edu.

pathological epidemiology study of alcohol intake and colon cancer risk by tumor LINE-1 methylation subtypes. User-friendly software to implement the methods is publicly available.

Keywords

heterogeneity test; molecular pathologic epidemiology; omics; pathogenesis; pathology

1. Introduction

Epidemiology is the study of the distribution and determinants of health and disease in human populations [1]. There is a general underlying premise that patients with a given disease share similar etiologies. However, advances in biomedical sciences are revealing inherent heterogeneity of pathogenesis and disease processes between individuals, leading to a shift from the conventional epidemiologic paradigm where a given disease is typically treated as a uniform entity [2–10]. Research that elucidates specific relationships between putative etiologic factors and cellular molecular alterations (i.e., molecular pathological epidemiology) can provide evidence for causality [8, 9]. Studies have identified links between exposures and specific molecular subtypes of neoplastic diseases, including endometrial [11], colorectal [11–28] and lung cancers [29–32]. Disease subtyping schemes have also been in development for non-neoplastic diseases, e.g., stroke [33], cardiovascular disease [34], autism [35], infectious disease [36], autoimmune disease [37], glaucoma [38] and obesity [39]. Disease subtypes may be defined by molecular characteristics or other features; for example, tumor sub-classification systems include lethality [40, 41], anatomical location, histopathologic features [42, 43], disease stage, tumor dominance [44], and tumor aggressiveness [45]. Because a more personalized preventive strategy may be feasible for individuals with a susceptibility to a specific disease subtype [8, 9], the statistical analysis of disease subtype heterogeneity may contribute to more effective translation of epidemiologic findings into disease prevention.

This paper does not address methods for the discovery of subtypes to be formed from high-dimensional data, but rather, restricts its consideration to a relatively small number of *a priori* subtypes, as illustrated in the examples above. Study of the following hypotheses can be used to assess the evidence in the data for the presence of heterogeneity in associations of risk factors with the incidence of disease A between two disease subtypes, A_1 and A_2. The first null hypothesis is that the exposure is not associated with disease subtype A_1. Letting β_j be the log-relative risk (RR) representing the exposure-subtype A_j association, where $j = 1,2$, this null hypothesis is H_0: $\beta_1 = 0$. Rejecting this null hypothesis supports the alternative that the exposure is associated with disease subtype A_1. The second hypothesis tests whether an exposure is associated with disease subtype A_2. These are the *subtype-specific hypothesis tests*. The third null hypothesis is that the relationship between the exposure and subtype A_1 is not different from that for subtype A_2, that is, H_0: $\beta_1 = \beta_2$. Rejecting this third hypothesis, which we call the *disease subtype heterogeneity hypothesis* or *common effect hypothesis*, provides evidence that the exposure contributes to risk of a specific pathogenic pathway (e.g., to subtype A1) in a significantly different manner than another pathogenic pathway (e.g., to subtype A2). In addition to testing this hypothesis, it is of interest to estimate the

extent of any difference in exposure-subtype associations by $\widehat{RRR} = \exp(\hat{\beta}_1)/\exp(\hat{\beta}_2)$, which is the ratio of the two subtype-specific RRs. Subtype research is usually motivated by having previously observed an overall association between the exposure and the disease, and is typically undertaken after the overall association is well-established, as in our motivating example of alcohol intake in relation to the risk of colon cancer [46–50]. The overall test of the exposure effect on the outcome has usually been conducted previously in a separate study, often with a large body of evidence accumulating before investigations into disease subtype heterogeneity are undertaken. Thus, the heterogeneity test, which is a focus of this paper, is typically not treated as part of a sequential testing procedure that begins with the overall test but rather treated as a separate test.

The purpose of this paper is to discuss statistical methods for the analysis of disease heterogeneity, propose recommendations for data analysis, and identify future areas for methodological innovation. Interest in etiological research is typically in the exposure effect on incident disease, and, therefore, we can generally assume that only one subtype will occur within a single subject. We will consider statistical methods to study disease heterogeneity with categorical and ordinal subtypes, for cohort, matched and unmatched case-control, and case-case studies. Although many of these methods have been available for use in other contexts, their application to the study of disease subtype heterogeneity has not been well elucidated.

2. Study design consideration for research on disease subtype heterogeneity

Research on subtype heterogeneity has been based on three primary design schemes: the prospective cohort design, the case-control design and the case-case design [8]. The prospective cohort design makes it possible to form case-cohort, nested case-control and case-case studies within it. The strengths and weaknesses of each of these designs as applied to disease subtype heterogeneity research have been discussed previously [8].

The judicious utilization of the platform made possible by an ongoing prospective cohort study or experimental trial can be a cost-effective approach [8, 51–56]. In contrast to case-control and case-case designs, experimental and observational prospective studies can facilitate research on multiple diseases. Once the infrastructure of a prospective study is established, numerous diseases can be studied at relatively low cost and effort, compared to the cost and effort required for establishing case-control studies for each disease, one by one. In addition to this compelling efficiency consideration, the benefits of the prospective cohort design strategy from the point of view of validity is strong – the ‘controls’ and alternative case subtypes indisputably arise from the same study base that has produced the cases of the subtype of interest, and hence the formidable challenge of overcoming selection bias due to non-comparable controls or cases of other subtypes is eliminated. In addition, the prospective nature of exposure data collection in most cohort studies eliminates recall bias as a potential source of bias and allows the investigator to optimize the quality of exposure information collected, thereby minimizing the extent to which differential measurement error, which may occur in case-control studies where measurements are taken after the

outcome is realized, thus allowing for the error to be correlated with the outcome, as well as non-differential measurement error, where the error is independent of the outcome, occurs.

The case-control study design is useful in the study of disease subtype heterogeneity for rare diseases or as a preliminary study where little is known about the association between risk factors and disease subtypes. Compared to prospective cohort studies, case-control studies tend to be less costly and shorter in duration. The controls are shared by all subtypes in unmatched case-control studies, while in matched case-control studies, some cases are matched to controls on one or more important potential confounders. For example, in a 1:m matched case-control study, the controls in the matched sets for Subtype 1 cases are typically not shared with the other subtypes. Heterogeneity tests can also be conducted in a case-case design, which is discussed in Section 3.4.2. The case-case design will, in general, be the cheapest and fastest approach for investigating disease subtype heterogeneity, at the cost of not being able to estimate the subtype-specific relative risks, only the ratio of these between the subtypes.

3. Statistical methods for disease heterogeneity analyses for categorical subtypes

3.1 Cohort studies

3.1.1 Statistical model—Since we assume that only one subtype will occur within a single subject, a standard competing risks framework can thus be used. In cohort studies, a commonly used statistical model is the cause-specific proportional hazards model [57, 58]

$$\lambda_j(t \mid \boldsymbol{X}_i, \boldsymbol{W}_i) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}_j \boldsymbol{X}_i + \boldsymbol{\theta}_j \boldsymbol{W}_i), \quad (1)$$

for subtype j, j =1, …, J, where $\lambda_j(t)$ is the incidence rate at time t for subtype j, $\lambda_{0j}(t)$ is the baseline incidence rate for subtype j, $\boldsymbol{X}_i$ is a column vector of possibly time-varying exposure variables for the ith participant, $\boldsymbol{W}_i$ is a column vector of possibly time-varying potential confounders of the relationship between $\boldsymbol{X}$ and the incidence of at least one of the J subtypes, and $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j$ are row vector-valued log relative risks (RRs) for the corresponding covariates for subtype j. Suppose $\boldsymbol{X}_i$ is K-dimention ($K \geq 1$), representing either one single exposure or the cross-classification of levels of multiple exposures; for a $m + 1$ level categorical exposure, m indicator variables will be created and included as elements of $\boldsymbol{X}_i$. For presentational simplicity, we sometimes assume K =1, although the methods discussed are easily extended to situations where K >1. In Model (1), it is advisable that the time scale is age [59, 60] and survival is left-truncated [61]. The methods discussed in this paper apply to other time scales as well, and can be easily extended to situations where the parameters of interest also include the coefficients of the $\boldsymbol{X} - \boldsymbol{W}$ interaction by treating the interaction terms as new variables.

Sometimes, it is reasonable to assume that the covariates, other than the exposure, have the same association with each of the J subtypes; that is, $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_J$. We name the models with the same regression coefficients for the covariates across subtypes *constrained* models.

Model (1) is written as an *unconstrained* model. Under the constrained model, Model (1) becomes

$$\lambda_j(t \mid \boldsymbol{X}_i, \boldsymbol{W}_i) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}_j \boldsymbol{X}_i + \boldsymbol{\theta} \boldsymbol{W}_i).$$

3.1.2 Methods for estimation and inference—Estimation and inference for parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j$, $j = 1, \ldots, J$, in Model (1) can be obtained by maximizing the partial likelihood [58],

$l(\boldsymbol{\beta}_j, \boldsymbol{\theta}_j, j = 1, \ldots, J) = \prod_{j=1}^{J} l_j(\boldsymbol{\beta}_j, \boldsymbol{\theta}_j)$, where l_j is a partial likelihood for the jth subtype, with

$$l_j(\boldsymbol{\beta}_j, \boldsymbol{\theta}_j) = \prod_{i=1}^{n} \left\{ \frac{\lambda_j(t_i \mid \boldsymbol{X}_i, \boldsymbol{W}_i)}{\sum_{l=1}^{n} I(t_l \geq t_i) \lambda_j(t_i \mid \boldsymbol{X}_l, \boldsymbol{W}_l)} \right\}^{I(Y_i = j)}$$
$$= \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}_j \boldsymbol{X}_i + \boldsymbol{\theta}_j \boldsymbol{W}_i)}{\sum_{l=1}^{n} I(t_l \geq t_i) \exp(\boldsymbol{\beta}_j \boldsymbol{X}_l + \boldsymbol{\theta}_j \boldsymbol{W}_l)} \right\}^{I(Y_i = j)},$$

where t_i is the minimum of time at disease occurrence and time at end of follow-up, for the ith participant, $i = 1, \ldots, n$, and Y_i is the observed subtype for the ith participant, if the ith participant became a case by the end of follow-up, and 0 otherwise. From standard likelihood method theory, the estimates of $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j$, $j = 1, \ldots, J$, are consistent [58]. As noted in Prentice *et al.* [58] and consistent with the standard assumptions made when the Cox model is used for analysis, we assume an independent censoring mechanism, which means that, at any given time point and given $\boldsymbol{X}$ and $\boldsymbol{W}$, individuals are not selectively censored on the basis of a relatively good or relatively poor subtype-specific prognosis. Also, as discussed by Prentice *et al.* [58], we note that, although no assumption is required about any interrelation between the subtypes for valid estimation and inference, the same inferences for subtype-specific effects would not necessarily prevail under a new set of study conditions in which, for example, certain subtypes have been eliminated.

Under the unconstrained model: Under the unconstrained model, the unknown parameters for the jth subtype, $(\boldsymbol{\beta}_j, \boldsymbol{\theta}_j)$, are not involved in l_q for $q \neq j$, but involved in l_j only, and thus inference for $\{\boldsymbol{\beta}_j, \boldsymbol{\theta}_j, j = 1, \ldots, J\}$ based on maximizing l is equivalent to maximizing l_j for each j, separately. It also follows that, in the matrix of the second derivative of $\log(l)$ with respect to parameters $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_J)$, the elements corresponding to the second derivative with respect to two parameters for different subtypes are zero; that is, $\frac{\partial^2 log(l)}{\partial \beta_{k_1 j_1} \partial \beta_{k_2 j_2}} = 0$ for $j_1 \neq j_2$, $k_1 = 1, \ldots, K$, $k_2 = 1, \ldots, K$, where β_{kj} is the kth element of $\boldsymbol{\beta}_j$. This implies that the estimated relative risks for distinct tumor subtypes are asymptotically uncorrelated. It follows that, for inference on the exposure effects for the jth subtype, standard Cox proportional hazards model analysis can be performed in which the participants experiencing another subtype will be censored at the time when the other subtype occurs [58]. This analysis will be referred as 'separate analysis' hereafter. Left

truncation and time-varying covariates can be handled by converting the data into an Anderson-Gill (counting process) data structure [62]. To test whether the exposure-disease association differs among the J subtypes, the contrast test statistic, given by $Z^2 = (\boldsymbol{C}\hat{\boldsymbol{\beta}})^T(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^T)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}})$, can be used, where, for presentational simplicity we assume K=1, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_J)^T$, which can be obtained from the separate analysis above, $\hat{\Sigma}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, with the estimated variance of $\hat{\beta}_j$, $j = 1, \ldots, J$, on the main diagonal and 0 everywhere else, $\boldsymbol{C}$ is a $(J-1)\times J$ contrast matrix [63], and superscript T denotes transpose. For example, if J=3, the rows of $\boldsymbol{C}$ are (1, −1, 0) and (1, 0, −1). This Z^2 statistic has an approximate χ^2 distribution with $J-1$ degrees of freedom under the null, H_0:$\boldsymbol{C\beta} = 0$ [63].

Alternatively, the disease subtype heterogeneity test can be conducted by the *duplication method* for Cox regression [64], which is based on the following transformation of Model (1) for subtype j:

$$\lambda_j(t \mid \boldsymbol{X}_i, \boldsymbol{W}_i) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}_1\boldsymbol{X}_{1i} + \cdots + \boldsymbol{\beta}_J\boldsymbol{X}_{Ji} + \boldsymbol{\theta}_1\boldsymbol{W}_{1i} + \cdots + \boldsymbol{\theta}_J\boldsymbol{W}_{Ji}), \quad (2)$$

where $\boldsymbol{X}_{ji} = \boldsymbol{X}_i$, $\boldsymbol{W}_{ji} = \boldsymbol{W}_i$, and $\boldsymbol{X}_{qi} = \boldsymbol{0}$, $\boldsymbol{W}_{qi} = \boldsymbol{0}$ for $q \neq j$, $j = 1, \ldots, J$. The parameters for subtype-specific exposure effects for all of the subtypes are now in the same model. In order to use standard software for analysis in this setting, this model can be fit using the Cox model stratified by subtype on an augmented data set, in which, each block of person-time is augmented for each subtype, and variables $\boldsymbol{X}_{1i}, \ldots, \boldsymbol{X}_{Ji}, \boldsymbol{W}_{1i}, \ldots, \boldsymbol{W}_{Ji}$ are duplicated variables based on $\boldsymbol{X}_i, \boldsymbol{W}_i$. See Section 5.1 for an example of how to construct an augmented data set. In model (2), the estimates of the log relative risks for exposure in relation to each subtype are available as regression coefficients. Another transformation of Model (1) is

$$\lambda_j(t \mid \boldsymbol{X}_i, \boldsymbol{W}_i) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}_1\boldsymbol{X}_i + \boldsymbol{\alpha}_2\boldsymbol{X}_{2i} \ldots + \boldsymbol{\alpha}_J\boldsymbol{X}_{Ji} + \boldsymbol{\theta}_1\boldsymbol{W}_{1i} + \cdots + \boldsymbol{\theta}_J\boldsymbol{W}_{Ji}), \quad (3)$$

where $\boldsymbol{\alpha}_j = \boldsymbol{\beta}_j - \boldsymbol{\beta}_1$ for $j = 2, \ldots, J$. In Model (3), the estimates of the log ratio of relative risks, $\boldsymbol{\alpha}_j$, $j = 2, \ldots, J$, are available as regression coefficients, and the heterogeneity test H_0:$\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_J$ is simply to test whether the regression coefficients, $\boldsymbol{\alpha}_2 = \cdots = \boldsymbol{\alpha}_J = 0$. In either Model (2) or (3), the heterogeneity hypothesis is best assessed through a likelihood ratio test (LRT) that compares the model that allows for separate associations for each disease subtype with the model that assumes a common association across subtypes. The Wald test, which is an alternative, has been shown to have poorer finite sample properties compared to the LRT [65].

If the unconstrained model is used in the duplication method, the maximum partial likelihood method-based point estimates of $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$, and their asymptotic variances obtained from the duplication approach will be the same as those obtained from fitting separate Cox models. If the unconstrained model is feasible given the data available, separate analysis with the contrast method has the advantage of avoiding the creation of the augmented data set required by the duplication method. The augmented data set can become very large if the original study is sizeable. In addition, the contrast method applies even when data are not available, and, instead, only the $\widehat{RR}s$ and their variance estimates for the

exposure-subtype associations are available from the literature. An advantage of the duplication method is that both the subtype-specific effect and the common effect test are available in a single analysis.

Under the constrained model: Under the constrained model, $\boldsymbol{\theta}_j$ in each l_j is replaced by the common parameter, $\boldsymbol{\theta}$, and thus inference for $\{\boldsymbol{\theta}, \boldsymbol{\beta}_j, j = 1, \ldots, J\}$ have to be made simultaneously for all j by maximizing l as a function of $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{\theta})$, and $\hat{\beta}_{k_1 j_1}$ and $\hat{\beta}_{k_2 j_2}$ are typically correlated for $j_1 \neq j_2$, $k_1 = 1, \ldots, K$, $k_2 = 1, \ldots, K$. Due to the unknown nonzero off-diagonal elements in Σ, the contrast test method, following the Cox proportional hazards model analysis separately for each subtype, is typically not suitable for the constrained model analysis. The duplication method enables the use of standard Cox regression software for constrained models and partially constrained models, in which the covariate-disease associations for some covariates are allowed to differ by subtype and some are forced to be the same across subtypes. Under the fully constrained model, Model (2) becomes

$$\lambda_j(t \mid \boldsymbol{X}_i, \boldsymbol{W}_i) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}_1 \boldsymbol{X}_{1i} + \cdots + \boldsymbol{\beta}_J \boldsymbol{X}_{Ji} + \boldsymbol{\theta} \boldsymbol{W}_i),$$

and Model (3) becomes

$$\lambda_j(t \mid \boldsymbol{X}_i, \boldsymbol{W}_i) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}_1 \boldsymbol{X}_i + \boldsymbol{\alpha}_2 \boldsymbol{X}_{2i} \ldots + \boldsymbol{\alpha}_J \boldsymbol{X}_{Ji} + \boldsymbol{\theta} \boldsymbol{W}_i).$$

The only difference in the duplication method here from that of the unconstrained analysis is that, in the augmented data set, duplicated variables are not created for the confounders to be constrained in the augmented data set. In the duplication method for the constrained model, since $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_J)$ in Model (2), or $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\alpha}}_2, \ldots, \hat{\boldsymbol{\alpha}}_J)$ if using Model (3), are estimated in the same Cox model through data augmentation, their variance-covariance matrices is also estimated, and the disease subtype heterogeneity across the subtypes can be evaluated using the Wald and likelihood ratio tests through standard methods.

In some applications, it is of interest not only to test whether there is overall heterogeneity, but also to identify which specific subtypes are responsible for any differences detected through a pairwise comparison of each pair of subtypes. When conducting these pairwise heterogeneity tests, multiple comparisons should be considered for controlling the Type I error rate. When there are multiple exposures of interest, disease subtype heterogeneity can be tested for each exposure separately, treated the others as potential confounders, or we can test the null hypothesis that the exposure-subtype associations are the same across subtypes for a group of cross-classified exposures, with the alternative hypothesis being that the disease subtype heterogeneity exists for at least one exposure.

3.1.3. Model selection: constrained vs. unconstrained model—In this section, we propose a new automated variable selection method for choosing between the constrained and unconstrained models. Using a scalar confounder, W_i as an example, in the *duplication method,* the term for this confounder in Model (2) can be re-written as $\theta_1 W_{1i} + \cdots + \theta_J W_{Ji} = \theta_1 W_i + \tau_2 W_{2i} + \cdots + \tau_J W_{Ji}$, where W_{qi} is defined as in Model (2), and $\tau_q = \theta_q - \theta_1$, $q = 2$,

… *J*. Since the constrained model implies $\theta_1 = \cdots = \theta_J$, and equivalently $\tau_2 = \cdots = \tau_J = 0$, the question of whether to allow the coefficient of W_i to differ by subtype is equivalent to the question of whether the entire group of variables, $W_{2i}, \ldots, W_{Ji}$, should be included in the model. Therefore, in the absence of *a priori* information, standard variable selection methods, such as stepwise regression, can be used to allow the data to select which covariates should not be constrained. Otherwise, biological knowledge should be used to determine which variables should be constrained. If the constrained model is true, the unconstrained analysis may be less efficient than the constrained analysis.

3.1.4. Evaluating the overall association—When the subtype-specific effects are in opposite direction, ignoring subtypes when evaluating the overall association of the exposure with the disease outcome may have reduced power due to the cancellation of the opposite effects of the different subtypes. To account for disease subtype heterogeneity to improve the power of the initial detection of the underlying susceptibility markers, we can test the null hypothesis that all subtype-specific effects are zero, i.e., $H_0: \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_J = 0$ through the data duplication method if either the constrained or unconstrained model is assumed, or through the separate analysis method if the unconstrained model is assumed. In this subtype specific test, opposite effects of the exposure on different subtypes will not cancel each other out. Bhattacharjee, *et al*. [66] have presented a subset-based analysis method that explores all possible subsets of subtypes and assesses significance of the association based on the best subset after adjusting for multiple comparisons. An advantage of this method is that it identifies the subset of subtypes that drives the overall association. A comparison of the power of this method and that of testing the null hypothesis that all subtype-specific effects are zero is a subject of future research.

3.1.5. Multiple comparison consideration—It is important to consider the context of the analysis (e.g., hypothesis generating versus hypothesis confirming) as well as the number of exposures and subtype levels in deciding how to best deal with multiple comparisons. If the purpose of the analysis is hypothesis confirming, a multiple comparison adjustment for subtype-specific tests may not be necessary [67–72]. If there is a large number of exposures under consideration for associations with a *J* level subtype (e.g., a large number of germline polymorphisms potentially associated with gene expression-based tumor subtypes), we may need to adjust for multiple comparisons among the exposure-specific overall heterogeneity tests. Good reviews of multiple comparison methods exist [73–75].

3.2. Nested case-control studies

Conditional logistic regression analysis is typically used to analyze data arising from nested case-control studies. As follows from the paradigm of the nested case-control study as risk-set sampled from the underlying study base that produced the cases [76], the regression coefficients estimate the log-RRs, and not just the log-odds ratios. Similar to the analysis of cohort studies, if the confounder-disease associations are unconstrained, the overall likelihood can be written as $\prod_{j=1}^{J} l_j(\boldsymbol{\beta}_j, \theta_j)$, where l_j is the likelihood for $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j$ the log-RRs for the *j*th subtype, and for the same reasons as explained in Section 3.1.2, $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_J$ are asymptotically uncorrelated. Therefore, we can conduct conditional logistic regression

analysis for each subtype separately and perform the heterogeneity test using the contrast method.

However, if all or some of the confounder-disease associations are constrained to be the same across subtypes, the models for different subtypes should be fit together in order to share parameter estimates among the models and appropriately account for the correlations between $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_J$ in constructing the heterogeneity tests. This can be done using the duplication method described in Section 3.1.2. Since controls are matched to cases on age and other characteristics here, unlike in the analysis of cohort studies in which a case of one subtype is included as a censored subject in the analysis for the other subtypes, the case(s) and control(s) in a matched set are typically not included as controls in the analysis of the other subtypes in other matched sets. Therefore, data augmentation with respect to rows (i.e. individuals) is not needed. We still need to create the subtype-specific covariates, $\boldsymbol{X}_{1i}, \ldots, \boldsymbol{X}_{Ji}$, in the augmented data set; for example, if the subtype of the case(s) in the matched set of individual i is j, then $\boldsymbol{X}_{ji} = \boldsymbol{X}_i$ and $\boldsymbol{X}_{qi} = 0$ for $q \neq j$. In unconstrained or partially constrained models, the unconstrained elements of $\boldsymbol{W}_i$ can be duplicated similarly. See Section 5.2 for an example of how to create an augmented data set. Statistical methods for matched case-control studies are similar to those discussed above for the nested case-control studies.

3.3. Unmatched case-control studies

3.3.1 Statistical model—In an unmatched case-control study design, there is a single control group shared by multiple case subtypes, as shown in Table 1, which presents data for an unmatched case-control study for a binary exposure and two subtypes. As previously, for presentational simplicity, we assume there is one single categorical exposure in Sections 3.3 and 3.4, although these methods can be easily extended to incorporate multiple continuous and/or categorical exposures. The unmatched case-control study data can be analyzed using the nominal polytomous logistic regression model

$$P(\boldsymbol{Y}_i = j \mid \boldsymbol{X}_i, \boldsymbol{W}_i)/P(\boldsymbol{Y}_i = 0 \mid \boldsymbol{X}_i, \boldsymbol{W}_i) = \exp(\beta_{0j} + \boldsymbol{\beta}_j \boldsymbol{X}_i + \boldsymbol{\theta}_j \boldsymbol{W}_i), j = 1, \ldots, \boldsymbol{J}, \quad (4)$$

where $\boldsymbol{Y}_i = j$ if the ith individual is a subtype j case, $\boldsymbol{Y}_i = 0$ if the ith individual is a control, $\boldsymbol{X}_i$ is a vector of indicator variables for the $K + 1$-level categorical exposure, X_i, $\boldsymbol{W}_i$, is defined in Section 3.1.1, and $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{Kj})$ $K \geq 1$, with β_{kj} the subtype-specific log odds ratio (OR) of the $(k + 1)$th level of the exposure relative to the reference level, $X = 1$. Note that $\exp(\beta_{0j})$, corresponding to $P(Y = j|\boldsymbol{X} = \boldsymbol{0}, \boldsymbol{W} = \boldsymbol{0})/P(Y = 0|\boldsymbol{X} = \boldsymbol{0}, \boldsymbol{W} = \boldsymbol{0})$ in the case-control sample, does not reflect the true ratio of these two probabilities in the population from which the case-control study was drawn whenever the sampling proportions are different between cases and controls, as is almost always the case[77]. From (4), we have

$$\exp(\beta_{kj}) = \frac{P(Y = j \mid X = k + 1, \boldsymbol{W})}{P(Y = 0 \mid X = k + 1, \boldsymbol{W})} \Big/ \frac{P(Y = j \mid X = 1, \boldsymbol{W})}{P(Y = 0 \mid X = 1, \boldsymbol{W})}. \quad (5)$$

If the disease is rare, $P(Y = 0|X = k + 1, \boldsymbol{W})$ and $P(Y = 0|X = 1, \boldsymbol{W})$ are approximately 1, and thus $\exp(\beta_{kj})$ approximates $P(Y = j|X = k + 1, \boldsymbol{W})/P(Y = j|X = 1, \boldsymbol{W})$, which is the RR for

the (k + 1) th level of exposure relative to the 1st level (the reference level) for the jth subtype, k = 1,…,K. Based on (5), the ratio of the ORs (ROR) for subtype j relative to subtype 1 (the reference subtype), $\exp(\beta_{kj} - \beta_{k1}) = \left\{\frac{P(Y=j \mid X=k+1, \boldsymbol{W})}{P(Y=0 \mid X=k+1, \boldsymbol{W})} / \frac{P(Y=j \mid X=1, \boldsymbol{W})}{P(Y=0 \mid X=1, \boldsymbol{W})}\right\}$, can be simplified as $/\left\{\frac{P(Y=1 \mid X=k+1, \boldsymbol{W})}{P(Y=0 \mid X=k+1, \boldsymbol{W})} / \frac{P(Y=1 \mid X=1, \boldsymbol{W})}{P(Y=0 \mid X=1, \boldsymbol{W})}\right\}$ $\frac{P(Y=j \mid X=k+1, \boldsymbol{W})}{P(Y=j \mid X=1, \boldsymbol{W})} / (\frac{P(Y=1 \mid X=k+1, \boldsymbol{W})}{P(Y=1 \mid X=1, \boldsymbol{W})})$, which is equivalent to the ratio of the RRs (RRR) for the jth subtype versus the 1st subtype.

3.3.2 Methods for inference—To conduct the heterogeneity test, we estimate β_{k1}, …, β_{kJ}, k = 1, … K, and the variance-covariance matrix of these estimates from the nominal polytomous regression model (4), and test the overall hypothesis $H_0 : \beta_{k1} = \cdots = \beta_{kJ}$ and/or pairwise hypotheses $H_0 : \beta_{kj_1} = \beta_{kj_2}$ for each subtype pair of (j_1,j_2) across all exposure levels overall or for each exposure level separately, using the LRT or Wald test. The unconstrained model analysis can be conducted using standard statistical software such as SAS PROC LOGISTIC; however, this SAS procedure cannot handle constrained models.

We propose a new data duplication method, through which the constrained unmatched case-control study models can be run in standard software such as SAS PROC LOGISTIC with STRATA statement. We first create an augmented data set in which controls are duplicated in each (subtype-specific) stratum. Since the controls are shared by all the subtypes, and cases of one subtype cannot be treated as controls for other subtypes, in the augmented data set here only the controls are augmented for each disease subtype, and not the cases. Let $\boldsymbol{X}_{1i}$, …, $\boldsymbol{X}_{Ji}$ be the augmented versions of the covariates; for subtype j strata, $\boldsymbol{X}_{ji} = \boldsymbol{X}_i$ and $\boldsymbol{X}_{qi} = 0$ if $q \neq j$. See Section 5.3 for an illustration of this augmented data set. The conditional logistic regression for this augmented dataset, with $\boldsymbol{X}_{1i}$, …, $\boldsymbol{X}_{Ji}$, $\boldsymbol{W}_i$ as covariates and is stratified by subtype, is

$$\frac{P(\boldsymbol{Y}_i > 0 \mid \textit{in subtype } j \textit{ stratum}, \boldsymbol{X}_i, \boldsymbol{W}_i)}{P(\boldsymbol{Y}_i = 0 \mid \textit{in subtype } j \textit{ stratum}, \boldsymbol{X}_i, \boldsymbol{W}_i)} = \exp(\beta_{0j} + \boldsymbol{\beta}_1 \boldsymbol{X}_{1i} + \cdots + \boldsymbol{\beta}_J \boldsymbol{X}_{Ji} + \boldsymbol{\theta} \boldsymbol{W}_i), \quad (6)$$

j = 1, …, J. Next we show that Model (6) is equivalent to Model (4) if $\boldsymbol{\theta}_j$ in Model (4) is replaced by $\boldsymbol{\theta}$. Since controls are duplicated in each stratum in the augmented data set, we have $P(\textit{in subtype } j \textit{ stratum} \mid \boldsymbol{X}, \boldsymbol{W}) = p(Y = j \textit{ or } Y = 0 \mid \boldsymbol{X}, \boldsymbol{W})$. It follows that $P(Y > 0 \mid \textit{in subtype } j \textit{ stratum}, \boldsymbol{X}, \boldsymbol{W}) = P(Y = j \mid \boldsymbol{X}, \boldsymbol{W}) / p(Y = j \textit{ or } Y = 0 \mid \boldsymbol{X}, \boldsymbol{W})$ and $P(Y = 0 \mid \textit{in subtype } j \textit{ stratum}, \boldsymbol{X}, \boldsymbol{W}) = P(Y = 0 \mid \boldsymbol{X}, \boldsymbol{W}) / p(Y = j \textit{ or } Y = 0 \mid \boldsymbol{X}, \boldsymbol{W})$, and thus the left hand side of Model (6) can be rewritten as $P(Y_i = j \mid \boldsymbol{X_i}, \boldsymbol{W_i}) / P(Y_i = 0 \mid \boldsymbol{X_i}, \boldsymbol{W_i})$. The right hand side is equivalent to $\exp(\beta_{0j} + \boldsymbol{\beta}_j \boldsymbol{X}_i + \boldsymbol{\theta} \boldsymbol{W}_i)$. Therefore, Model (6) is equivalent to Model (4) if $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_J$ in Model (4). Further research is needed to quantify the relative efficiency of the unconditional and conditional methods for the heterogeneity test in the setting.

3.4. Case-case studies

3.4.1. Statistical model—In a case-case study, for estimation and inference about the ROR and to perform the heterogeneity test, the data can be analyzed using the nominal polytomous logistic regression model

$$P(Y_i = j \mid \boldsymbol{X_i}, \boldsymbol{W_i}) / P(Y_i = 1 \mid \boldsymbol{X_i}, \boldsymbol{W_i}) = \exp(\alpha_{0j} + \boldsymbol{\alpha_j X_i} + \tau_j \boldsymbol{W_i}), j = 2, \ldots, \boldsymbol{J}, \quad (7)$$

where $\boldsymbol{\alpha_j} = (\alpha_{1j}, \ldots, \alpha_{kj})$ with $\exp(\alpha_{kj}) = \frac{P(Y = j \mid X = k+1, \boldsymbol{W})}{P(Y = j \mid X = 1, \boldsymbol{W})} / \frac{P(Y = 1 \mid X = k+1, \boldsymbol{W})}{P(Y = 1 \mid X = 1, \boldsymbol{W})}$, which is the *ROR* or equivalently the RRR as discussed in Section 3.3.1 for subtype j relative to subtype 1, for exposure level $k + 1$ relative to the reference exposure level. Dividing Model (4) for the unmatched case-control study, $P(Y = j|\boldsymbol{X_i}, \boldsymbol{W_i}) / P(Y = 0|\boldsymbol{X_i}, \boldsymbol{W_i}) = \exp(\boldsymbol{\beta}_{0j} + \boldsymbol{\beta_j X_i} + \boldsymbol{\theta_j W_i})$ by the same model with $j = 1$, leads to Model (7). It follows that $\boldsymbol{\alpha_j} = \boldsymbol{\beta_j} - \boldsymbol{\beta_1}$ and $\boldsymbol{\tau_j} = \boldsymbol{\theta_j} - \boldsymbol{\theta_1}$. Therefore, the heterogeneity test for comparing subtypes j and 1 tests the null hypothesis $H_0 : \alpha_{kj} = 0, k = 1, \ldots, K$, and can be conducted using the LRT, score test, or Wald test. This nominal polytomous logistic regression model in the case-case design allows for testing differences in exposure-subtype associations in any pairwise or multi-way comparison of the J subtypes.

We show in the Appendix that the case-case model (7) is a parametric transformation of the fundamental model (1).

Since $\boldsymbol{\tau_j} = \boldsymbol{\theta_j} - \boldsymbol{\theta_1}$, only those confounders that themselves have disease subtype heterogeneity with respect to two or more of the subtypes under consideration need to be controlled for in case-case analysis using Model (7). Therefore, the constrained model which assumes that the covariates, $\boldsymbol{W}$, have the same effects across all subtypes is the model that excludes $\boldsymbol{W}$. In other words, under the constrained model, although $\boldsymbol{W}$ may be a confounder for the exposure-outcome relationship represented by $\boldsymbol{\beta_j}$, $\boldsymbol{W}$ will not be a confounder of the subtype-heterogeneity effect represented by $\boldsymbol{\alpha_j}$ in Model (7).

3.4.2. Utility of case-case studies—Typically, investigators are not only interested in testing and estimation of the heterogeneity hypothesis, but are also interested in testing and estimation of the subtype-specific exposure-disease associations, which cannot be assessed in a case-case study. Nevertheless, a cost-effective strategy would be to test the heterogeneity hypothesis in an adequately powered case-case design, and then, for those exposures for which there is sufficient evidence to reject the null, proceed to the second stage, to conduct a case-control study or a prospective cohort study with adequate power for assessing subtype-specific exposure-disease associations. Usually the subtype research has been motivated by having previously observed an overall association; thus, if the null hypothesis of the heterogeneity test is not rejected, no further analysis is needed.

4. Statistical methods for disease heterogeneity analyses for subtypes defined by ordinal markers

Many disease biomarkers derive from continuous measurements. Here, we present methods for evaluating subtype heterogeneity when there are ordinal subtype categories; analytic methods for continuous markers will be reported in a separate paper. As an example, colon cancer can be sub-classified into ordinal categories according to high, medium or low DNA methylation levels of the long interspersed nucleotide element-1 (LINE-1) [78].

The method of Chatterjee [79], Chatterjee, *et al.* [80], and Rosner, *et al.* [81] can be usefully applied in cohort studies to study ordinal markers. We will call this method the "one-stage method". Let variable *score* be the ordinal or median score for each subtype. An ordinal score assigns value j to variable $score_j$, while a median score assigns the median value of the continuous biomarker in category j to $score_j$. In the one stage method, $\boldsymbol{\beta}_j$ in Model (1) is replaced by $\boldsymbol{\beta}_j(\gamma_0, \gamma_1) = \gamma_0 + \gamma_1 \times score_j$. The maximum partial likelihood estimate given in [79, 80] of (γ_0, γ_1) can be obtained using the data duplication method described in Section 3.1.2.

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 large. Alternatively, the two-stage approach described by Wang *et al.* [82] can be used to analyze ordinal subtypes. We first assume the exposure variable X_i is scalar. This includes the situations in which the exposure is continuous or binary, and in a trend analysis for a categorical exposure in which a new continuous variable, the median level in each exposure category, is included in Model (1). In the first stage of the two-stage method, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_J$ are obtained. Then, in the second stage, we fit a fixed effects meta-regression [83],

$$\hat{\beta}_j = \gamma_0 + \gamma_1 \times score_j + e_j,$$

where $e = (e_1, \ldots, e_J)$ are the within-subtype sampling errors, distributed as $N(0,\Sigma)$, with the (j_1,j_2)th element of Σ, $\Sigma_{j_1 j_2}$ being $\widehat{Cov}(\hat{\beta}_{j_1}, \hat{\beta}_{j_2})$ for $j_1 \neq j_2$, and $\sum_{jj} = \widehat{var}(\hat{\beta}_j)$, or a random effects meta-regression model [83],

$$\hat{\beta}_j = \gamma_0 + \gamma_1 \times score_j + b_j + e_j,$$

where b_j are the random subtype-specific effects, distributed as $N(0, \sigma_b^2)$, independent from e, representing the heterogeneity between the estimated effects of the exposures on the subtypes that is not explained by variable *score* and covariates already considered in the subtype-specific analysis. As discussed in [82], the fixed effects meta-regression method and the one-stage method under the unconstrained model fit the same model. The random effects meta-regression method has an advantage over both the fixed effects method and the one-stage method in that it can incorporate additional heterogeneity between subtypes that cannot be explained by the ordinal score of the marker. When additional heterogeneity can reasonably be anticipated, as would often be the case, the random effects model may be useful. When $\sigma_b^2 = 0$, the random effects method will be reduced to, or will be closed to, the fixed effect method [84]. Additional sources of variation among the $\hat{\beta}_j$ can be the result of other unmeasured subtype-specific determinants which are correlated with the marker under study.

The test of the hypothesis $H_0: \gamma_1 = 0$ tests whether the subtype-specific exposure-disease association has a monotone increasing or decreasing trend across the ordinal or median-scored disease subtypes. As discussed earlier, in unconstrained analyses of a cohort or

nested case-control study, these subtype-specific estimates are asymptotically uncorrelated and can be obtained by subtype-specific regression analysis; i.e., $\Sigma_{j_1 j_2}$ can be set to 0 when $j_1 \neq j_2$. For the unmatched case-control design and/or constrained analyses, the subtype-specific effect estimates are correlated, and we have $\Sigma_{j_1 j_2} \neq 0$; methods for estimating the parameters in the meta-regression model have been given [85, 86].

In the cases of $\boldsymbol{\beta}_{1j} = (\beta_{1j}, \ldots, \beta_{jK})$ $K > 1$, with the first stage analysis of the two-stage method is the same as in the cases when β_j is a scalar. At the second stage, we can conduct the meta-regression analysis for each element of $\boldsymbol{\beta}_j$ separately.

In a case-case study, with subtype 1 as the reference subtype, we can fit the resulting random effects meta-regression model: $\hat{\alpha_j} = \gamma_1 \times Dscore_j + b_j + e_j$, $j = 2, \ldots, J$, or a fixed effects meta-regression model $\hat{\alpha_j} = \gamma_1 \times Dscore_j + e_j$, with $Dscore_j = score_j - score_1$, $\sum_{j_1 j_2} = \widehat{Cov}(\hat{\alpha}_{j_1}, \hat{\alpha}_{j_2})$ for $j_1 \neq j_2$, and $\sum_{jj} = \widehat{var}(\hat{\alpha}_j)$. The test of $H_0: \gamma_1 = 0$ tests whether the subtype-specific exposure-disease association has a monotone increasing or decreasing trend across the ordinal or median-scored subtypes. A comparison of the efficiency of the disease heterogeneity trend test from the case-case study to that based on the other study types is a topic for future research.

5. Case study

We illustrate the methods in Sections 3 and 4 in a molecular pathological epidemiology study of the association of alcohol consumption with colon cancer subtypes defined by LINE-1 methylation level, categorized as high, medium or low, in the Health Professionals Follow-up Study (HPFS), where 51,529 men were followed between 1986 and 2000 [87], and 99 LINE-1 methylation-high, 102 LINE-1 methylation-medium and 67 LINE-1 methylation-low colon cancer cases were observed. Alcohol intake was assessed at baseline in 1986 and categorized into 4 groups, 0g/day, >0 and ≤5 g/day, >5 and ≤15g/day, >15g/day, and the median level of alcohol intake in each group is treated as a continuous exposure in this analysis.

5.1. Analysis of the cohort study

Table 2 shows the original data set, cohort, in the Anderson-Gill data format [62] for this study. In this data set, id is the study participant's unique identifier; time is months from the start of the questionnaire cycle until colon cancer incidence, date of next questionnaire return, death or end of study, whichever happens first; cancer is the outcome variable (1 for high LINE-1, 2 for medium LINE-1, 3 for low LINE-1, and 0 for censored or controls); period represents the every-two-year questionnaire cycle; agemo is age in months at the beginning of each questionnaire cycle; alcohol is the median level of alcohol intake in units of 12g/day for the alcohol categories described above and asp denotes current aspirin use (0 for < 2 tablets per week and 1 for ≥2 tablets per week), which is a potential confounder and updated in each questionnaire cycle. For presentational simplicity, we will only discuss one confounder, current aspirin use, in this analysis. Given the data set above, the rows of which are ordered by id and period, the first step of the duplication method is to re-format the data set such that each block of person-time is augmented for each of the three cancer subtypes.

Table 3 is the augmented data set, cohort_aug, for the participant id=1. In cohort_aug, the variable censor is a censoring indicator for each disease subtype (specified by variable type); it is 0 if censored and 1 if the specific disease subtype is diagnosed in the corresponding block of person-time. The variables alcohol_1, alcohol_2 and alcohol_3 are the augmented exposure variables for the three cancer subtypes, respectively. Below is the SAS program for the constrained analysis in which the effect of asp is assumed to be the same across subtypes. Note that, to control as finely as possible for confounding by age, calendar time and any possible two-way interactions between these two time scales, we stratified the analysis jointly by age in months at start of follow-up and calendar year of the current questionnaire cycle. The time scale for the analysis is then measured as 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.

```
proc phreg data=cohort_aug;
model time*censor(0)=alcohol_1 alcohol_2 alcohol_3 asp;
strata type agemo period;
```

For an unconstrained analysis using the duplication method, augmented variables for asp will need to be created and included in the model. Note that, although it is always necessary to augment the data set by additional observations for each person corresponding to each subtype under consideration, there is an alternative method of coding which avoids the creation of the augmented variables described. For example, alcohol_1 in the SAS statement above can be replaced by the interaction of alcohol and an indicator variable for the first level of type.

5.2. Analysis of the nested case-control study

To illustrate the use of the methods in the nested case-control studies, we created a nested case-control study with 1:2 matching by risk-set sampling [76] from the original cohort study. The data set ncaco in Table 4 shows the standard format for three matched sets, where matchid indexes the stratum defined by matching variables, and age is age in years when the cancer was diagnosed. Table 5 is the augmented data set, ncaco_aug, for these matched sets using in the duplication method for a nested case-control study. This augmented data set corresponds to the constrained analysis with respect to aspirin use and thus no augmented variable is created for this variable. The standard conditional logistic regression with covariates alcohol_1, alcohol_2 and alcohol_3 was used to analyze this nested case-control study. For an unconstrained analysis using the duplication method, as in the cohort study, augmented variables for asp need to be created and included in the model.

5.3. Analysis of the unmatched case-control study

To illustrate our analysis methods for the unmatched case-control study, we treated the data set ncaco from Section 5.2 as having arisen from an unmatched case-control study after excluding the three controls who developed cancer after they were sampled as the matched controls; we name this resulting data set uncaco. Table 6 is the augmented data set, uncaco_aug, for the first three ids; this data set was used in the conditional logistic

regression analysis described in Section 3.3.2 for the constrained analysis for the unmatched case control design. Variables censor and type in uncaco_aug have the same definitions as those in cohort_aug. Below is the SAS code for the constrained conditional logistic regression in which the effect of asp is assumed to be the same across subtypes.

```
proc logistic data=uncaco_aug;
strata type;
model censor(event='1')=alcohol_1 alcohol_2 alcohol_3 asp;
```

Standard nominal polytomous logistic regression on the original data uncaco is used for the unconstrained analysis of the unmatched case-control study.

5.4. Analysis of the case-case study

The analysis of the case-case study can use standard nominal polytomous logistic regression on the original data.

5.5. Results

Table 7 summarizes the results of the subtype-specific $\widehat{RR}$s and the heterogeneity tests for alcohol intake in relation to the incidence of LINE-1 methylation subtypes for each of the cohort, nested case-control, unmatched case-control and case-case designs. The case-case study data included all 268 cases. The following variables were adjusted for in all analyses: current aspirin use (≥2 tablets/week or less), body mass index (kg/m^2) (<21, 21–22.9, 23–24.9, 25–29.9, 30+), history of colorectal screening (yes/no), physical activity in metabolic equivalent of tasks (quintiles), history of colorectal polyps (yes/no), family history of colon cancer (yes/no), smoking (pack-years), red meat intake (quintiles), multivitamin use (yes/no), calcium intake (quintiles) and folate intake (quintiles). We used the 2-degree of freedom Wald test for categorical subtypes ignoring ordering, to test whether the alcohol-colon cancer association is the same across the three subtypes of colon cancer, and also used the fixed effects meta-regression method for ordinal subtypes with score=0,1,2 for subtypes low, medium and high. In the unconstrained model, the associations of each potential confounder with the subtypes of colon cancer were allowed to be different among subtypes; in the constrained model, the associations were assumed to be the same across subtypes for all the potential confounders.

As shown in Table 7, the heterogeneity test p-values were similar between the case-case design and the full cohort design. The nested case-control study, which has 12 (4%) non-informative strata, is subject to a finite sample efficiency loss. For estimating subtype-specific *RR*s, the asymptotic efficiency of a nested case-control study with 1:2 matching should be about 67% of that of the full cohort [88, 89]. In this example, the unmatched case-control study is more efficient than the nested case-control study. This could be mainly due to the fact that in the unmatched design, the 533 controls are available for the analysis of all of the subtypes, while in the matched design, the numbers of controls available for the analysis of each subtype are 198, 204 and 134, respectively. Further research on the relative efficiency of these designs for the heterogeneity test is of interest.

As seen in Table 7, the constrained and unconstrained models for the cohort study led to similar results. The effect of alcohol in relation to colon cancer incidence did vary by the three LINE-1 methylation colon cancer subtypes (p=0.017); higher alcohol consumption significantly increased the risk of LINE-1 medium (RR=1.57; 95% CI: 1.27–1.94) and low (RR=1.36; 95% CI: 1.05–1.77) colon cancer, but had no association with LINE-1 high colon cancer. When treating LINE-1 methylation level as ordinal, there was insufficient evidence to conclude that the colon cancer risk increased or decreased monotonically across the LINE-1 levels – the subtype-specific effects across the subtype groups appeared to be non-linear. When constructing the other study designs from the full cohort data for illustrative purposes, the results were not substantially different from the full cohort from which they originated and from each other.

6. Software

We have developed the SAS macro %subtype for cohort, matched or nested case-control, unmatched case-control and case-case studies implementing the data duplication and the nominal polytomous logistic regression methods discussed in this paper for categorical subtypes. The macro provides point and interval estimates of subtype-specific relative risks, and the test for subtype specific exposure-disease associations and for the overall and pair-wise heterogeneity hypotheses. We have also developed SAS macros %contrastTest, implementing the contrast test, and %meta_subtype_trend, which implements the fixed effects and random effects meta-regression methods for ordinal subtypes. These SAS macros, along with user friendly manuals, can be obtained at the second author's website, http://www.hsph.harvard.edu/donna-spiegelman/software/. R function rma.mv() can be used for the meta-regression analysis when the subtype-specific $\widehat{RR}$s are correlated [90].

7. Discussion

We have presented an overview of methods for the characterization and assessment of heterogeneity in the associations of exposures with more than one disease subtype. Methods are given for scenarios where disease subtypes are categorical and ordinal, appropriate for cohort studies, matched and unmatched case-control studies, and case-case study designs.

This paper is restricted to the study of disease subtype heterogeneity among a relatively small number of *a priori* subtypes. Bhattacharjee and colleagues [66] have proposed methods for identifying subtypes, as mentioned in Section 3.1.4. Begg and colleagues [3, 7, 91] have defined a measure of disease subtype heterogeneity based on the risk distributions. This measure can be used as a tool to examine different subtyping options to determine which ones correspond to the most distinctive disease subtype heterogeneities. Future research could further develop the methods for the discovery of subtypes to be formed from high-dimensional data, such as tumor genomic data.

Although the methods discussed in this paper apply to any multiple endpoint situation as long as only one of these multiple endpoints can occur in a single person (i.e. competing risks), the disease subtype heterogeneity of primary consideration in this paper was motivated by studies of the occurrence of multiple possible subtypes of a single disease

where if one subtype occurs, it is not possible for other subtypes to occur. For example, it is not possible for an incident colon tumor to have both high and low LINE-1 methylation levels. However, in some cases it may be possible for a single participant to experience more than one subtype and it could also be of interest to assess disease subtype heterogeneity across non-competing distinct diseases. For example, we may wish to assess whether the effect of cigarette smoking on lung cancer risk is the same as that for heart disease. These are examples of non-competing risks, for which the coefficients in cause-specific hazards model (1) can be validly estimated from cohort studies, and variances can also be validly estimated using a sandwich method [92]. The methods presented in Section 3.1 of this paper can be easily extended to this setting by using the data duplication method. Then, disease subtype heterogeneity across the non-competing risks subtypes can be evaluated using the Wald test using the sandwich variance-covariance matrix [92], and in the contrast test method, $\hat{\Sigma}$ is the sandwich variance-covariance matrix, the off-diagonal elements of which are no longer zero.

Disease subtype data are often missing in some proportion of cases. An estimating function method can be used to handle missing subtype data under a missing-at-random assumption [80]. Further research comparing this approach to handling missing subtype data to alternatives is of interest.

Note that all the tests introduced in this paper rely on asymptotic results, and thus may not perform well in small to median-size studies. In addition, some preliminary investigations we have conducted have indicated that to achieve any given power, the heterogeneity test requires larger sample sizes than the tests of subtype-specific main effects. Efficiency and power in this context is a topic for future research.

In addition to the analysis of disease incidence risk according to heterogeneous subtypes, the study of disease heterogeneity can help identify lifestyle factors which modify the course of a specific disease subtype after disease diagnosis [8, 9]. These findings may provide the rationale for initiating clinical trials to assess the efficacy of lifestyle or pharmacological intervention, which targets the specific subtypes.

The use of disease biomarkers is increasingly common in clinical practice and research [93–95] as exemplified by the emergence and evolution of molecular pathological epidemiology (MPE). MPE represents an integrative interdisciplinary field of molecular pathology and epidemiology, and has been increasingly recognized not only in cancer sciences [96], but also in non-neoplastic disease areas [97]. This trend is further facilitated by the recent initiative of precision medicine [98, 99]. Therefore, the statistical methods discussed in this paper will be useful analytic tools for biomedical and population health science.

Acknowledgments

We thank the Associate Editor and two referees for their helpful comments in improving this paper. We also thank Dr. Meir Stampfer for reviewing this paper and providing helpful comments. This research was supported by the NIH grants R01 CA151993, UM1 CA167552, P01 CA55075, P01 CA87969, UM1 CA186107, R01 CA137178 and R35 CA197735.

References

1. MacMahon, B, Pugh, TF. Epidemiologic Methods. Little, Brown: Boston; 1960.
2. 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; 26:465–484. [PubMed: 23307060]
3. Begg CB. A strategy for distinguishing optimal cancer subtypes. Int J Cancer. 2011; 129:931–937. [PubMed: 20949563]
4. Begg CB, Zabor EC. Detecting and exploiting etiologic heterogeneity in epidemiologic studies. American Journal of Epidemiology. 2012; 176:512–518. [PubMed: 22922440]
5. Amaral AF, Mendez-Pertuz M, Munoz A, Silverman DT, Allory Y, Kogevinas M, Lloreta J, Rothman N, Carrato A, Rivas del Fresno M, Real FX, Malats N. Plasma 25-hydroxyvitamin D(3) and bladder cancer risk according to tumor stage and FGFR3 status: a mechanism-based epidemiological study. J Natl Cancer Inst. 2012; 104:1897–1904. [PubMed: 23108201]
6. Shen H, Fridley BL, Song H, Lawrenson K, Cunningham JM, Ramus SJ, Cicek MS, Tyrer J, Stram D, Larson MC, Kobel M, Ziogas A, Zheng W, Yang HP, Wu AH, Wozniak EL, Woo YL, Winterhoff B, Wik E, Whittemore AS, Wentzensen N, Weber RP, Vitonis AF, Vincent D, Vierkant RA, Vergote I, Van Den Berg D, Van Altena AM, Tworoger SS, Thompson PJ, Tessier DC, Terry KL, Teo SH, Templeman C, Stram DO, Southey MC, Sieh W, Siddiqui N, Shvetsov YB, Shu XO, Shridhar V, Wang-Gohrke S, Severi G, Schwaab I, Salvesen HB, Rzepecka IK, Runnebaum IB, Rossing MA, Rodriguez-Rodriguez L, Risch HA, Renner SP, Poole EM, Pike MC, Phelan CM, Pelttari LM, Pejovic T, Paul J, Orlow I, Omar SZ, Olson SH, Odunsi K, Nickels S, Nevanlinna H, Ness RB, Narod SA, Nakanishi T, Moysich KB, Monteiro AN, Moes-Sosnowska J, Modugno F, Menon U, McLaughlin JR, McGuire V, Matsuo K, Adenan NA, Massuger LF, Lurie G, Lundvall L, Lubinski J, Lissowska J, Levine DA, Leminen A, Lee AW, Le ND, Lambrechts S, Lambrechts D, Kupryjanczyk J, Krakstad C, Konecny GE, Kjaer SK, Kiemeney LA, Kelemen LE, Keeney GL, Karlan BY, Karevan R, Kalli KR, Kajiyama H, Ji BT, Jensen A, Jakubowska A, Iversen E, Hosono S, Hogdall CK, Hogdall E, Hoatlin M, Hillemanns P, Heitz F, Hein R, Harter P, Halle MK, Hall P, Gronwald J, Gore M, Goodman MT, Giles GG, Gentry-Maharaj A, Garcia-Closas M, Flanagan JM, Fasching PA, Ekici AB, Edwards R, Eccles D, Easton DF, Durst M, du Bois A, Dork T, Doherty JA, Despierre E, Dansonka-Mieszkowska A, Cybulski C, Cramer DW, Cook LS, Chen X, Charbonneau B, Chang-Claude J, Campbell I, Butzow R, Bunker CH, Brueggmann D, Brown R, Brooks-Wilson A, Brinton LA, Bogdanova N, Block MS, Benjamin E, Beesley J, Beckmann MW, Bandera EV, Baglietto L, Bacot F, Armasu SM, Antonenkova N, Anton-Culver H, Aben KK, Liang D, Wu X, Lu K, Hildebrandt MA, Schildkraut JM, Sellers TA, Huntsman D, Berchuck A, Chenevix-Trench G, Gayther SA, Pharoah PD, Laird PW, Goode EL, Pearce CL. Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. Nature communications. 2013; 4:1628.
7. Begg CB, Zabor EC, Bernstein JL, Bernstein L, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. Statistics In Medicine. 2013; 32:5039–5052. [PubMed: 23857589]
8. Ogino S, Chan AT, Fuchs CS, Giovannucci E. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. Gut. 2011; 60:397–411. [PubMed: 21036793]
9. Ogino S, Stampfer M. Lifestyle factors and microsatellite instability in colorectal cancer: the evolving field of molecular pathological epidemiology. J Natl Cancer Inst. 2010; 102:365–367. [PubMed: 20208016]
10. Ogino S, Lochhead P, Giovannucci E, Meyerhardt JA, Fuchs CS, Chan A. Discovery of colorectal cancer PIK3CA mutation as potential predictive biomarker: power and promise of molecular pathological epidemiology. Oncogene. 2014; 33:2949–2955. [PubMed: 23792451]
11. Chen H, Taylor NP, Sotamaa KM, Mutch DG, Powell MA, Schmidt AP, Feng S, Hampel HL, Chapelle AD, Goodfellow PJ. Evidence for heritable predisposition to epigenetic silencing of MLH1. Int J Cancer. 2007; 120:1684–1688. [PubMed: 17230510]

12. Allan JM, Shorto J, Adlard J, Bury J, Coggins R, George R, Katory M, Quirke P, Richman S, Scott D, Scott K, Seymour M, Travis LB, Worrillow LJ, Bishop DT, Cox A. MLH1 -93G>A promoter polymorphism and risk of mismatch repair deficient colorectal cancer. Int J Cancer. 2008; 123:2456–2459. [PubMed: 18712731]
13. Campbell PT, Curtin K, Ulrich CM, Samowitz WS, Bigler J, Velicer CM, Caan B, Potter JD, Slattery ML. Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle factors. Gut. 2009; 58:661–667. [PubMed: 18523027]
14. Raptis S, Mrkonjic M, Green RC, Pethe VV, Monga N, Chan YM, Daftary D, Dicks E, Younghusband BH, Parfrey PS, Gallinger SS, McLaughlin JR, Knight JA, Bapat B. MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. J Natl Cancer Inst. 2007; 99:463–474. [PubMed: 17374836]
15. Samowitz WS, Curtin K, Wolff RK, Albertsen H, Sweeney C, Caan BJ, Ulrich CM, Potter JD, Slattery ML. The MLH1 -93 G>A promoter polymorphism and genetic and epigenetic alterations in colon cancer. Genes Chromosomes Cancer. 2008; 47:835–844. [PubMed: 18615680]
16. Ogino S, Hazra A, Tranah GJ, Kirkner GJ, Kawasaki T, Nosho K, Ohnishi M, Suemoto Y, Meyerhardt JA, Hunter DJ, Fuchs CS. MGMT germline polymorphism is associated with somatic MGMT promoter methylation and gene silencing in colorectal cancer. Carcinogenesis. 2007; 28:1985–1990. [PubMed: 17621591]
17. Hawkins NJ, Lee JH, Wong JJ, Kwok CT, Ward RL, Hitchins MP. MGMT methylation is associated primarily with the germline C>T SNP (rs16906252) in colorectal cancer and normal colonic mucosa. Mod Pathol. 2009; 22:1588–1599. [PubMed: 19734844]
18. Slattery ML, Curtin K, Anderson K, Ma KN, Ballard L, Edwards S, Schaffer D, Potter J, Leppert M, Samowitz WS. Associations between cigarette smoking, lifestyle factors, and microsatellite instability in colon tumors. J Natl Cancer Inst. 2000; 92:1831–1836. [PubMed: 11078760]
19. Satia JA, Keku T, Galanko JA, Martin C, Doctolero RT, Tajima A, Sandler RS, Carethers JM. Diet, lifestyle, and genomic instability in the north Carolina colon cancer study. Cancer Epidemiol Biomarkers Prev. 2005; 14:429–436. [PubMed: 15734969]
20. Slattery ML, Curtin K, Sweeney C, Levin TR, Potter J, Wolff RK, Albertsen H, Samowitz WS. Diet and lifestyle factor associations with CpG island methylator phenotype and BRAF mutations in colon cancer. Int J Cancer. 2007; 120:656–663. [PubMed: 17096326]
21. Campbell PT, Jacobs ET, Ulrich CM, Figueiredo JC, Poynter JN, McLaughlin JR, Haile RW, Jacobs EJ, Newcomb PA, Potter JD, Le Marchand L, Green RC, Parfrey P, Younghusband HB, Cotterchio M, Gallinger S, Jenkins MA, Hopper JL, Baron JA, Thibodeau SN, Lindor NM, Limburg PJ, Martinez ME. Registry ftCCF. Case-control study of overweight, obesity, and colorectal cancer risk, overall and by tumor microsatellite instability status. J Natl Cancer Inst. 2010; 102:391–400. [PubMed: 20208017]
22. Kuchiba A, Morikawa T, Yamauchi M, Imamura Y, Liao X, Chan AT, Meyerhardt JA, Giovannucci E, Fuchs CS, Ogino S. Body mass index and risk of colorectal cancer according to fatty acid synthase expression in the nurses' health study. J Natl Cancer Inst. 2012; 104:415–420. [PubMed: 22312135]
23. Wu AH, Shibata D, Yu MC, Lai MY, Ross RK. Dietary heterocyclic amines and microsatellite instability in colon adenocarcinomas. Carcinogenesis. 2001; 22:1681–1684. [PubMed: 11577009]
24. Chia VM, Newcomb PA, Bigler J, Morimoto LM, Thibodeau SN, Potter JD. Risk of microsatellite-unstable colorectal cancer is associated jointly with smoking and nonsteroidal anti-inflammatory drug use. Cancer Res. 2006; 66:6877–6883. [PubMed: 16818666]
25. Samowitz WS, Albertsen H, Sweeney C, Herrick J, Caan BJ, Anderson KE, Wolff RK, Slattery ML. Association of smoking, CpG island methylator phenotype, and V600E BRAF mutations in colon cancer. J Natl Cancer Inst. 2006; 98:1731–1738. [PubMed: 17148775]
26. Poynter JN, Haile RW, Siegmund KD, Campbell PT, Figueiredo JC, Limburg P, Young J, Le Marchand L, Potter JD, Cotterchio M, Casey G, Hopper JL, Jenkins MA, Thibodeau SN, Newcomb PA, Baron JA. Associations between smoking, alcohol consumption, and colorectal cancer, overall and by tumor microsatellite instability status. Cancer Epidemiol Biomarkers Prev. 2009; 18:2745–2750. [PubMed: 19755657]

27. Rozek LS, Herron CM, Greenson JK, Moreno V, Capella G, Rennert G, Gruber SB. Smoking, gender, and ethnicity predict somatic BRAF mutations in colorectal cancer. Cancer Epidemiol Biomarkers Prev. 2010; 19:838–843. [PubMed: 20200438]

28. Limsui D, Vierkant RA, Tillmans LS, Wang AH, Weisenberger DJ, Laird PW, Lynch CF, Anderson KE, French AJ, Haile RW, Harnack LJ, Potter JD, Slager SL, Smyrk TC, Thibodeau SN, Cerhan JR, Limburg PJ. Cigarette Smoking and Colorectal Cancer Risk by Molecularly Defined Subtypes. J Natl Cancer Inst. 2010; 102:1012–1022. [PubMed: 20587792]

29. Leng S, Bernauer AM, Hong C, Do KC, Yingling CM, Flores KG, Tessema M, Tellez CS, Willink RP, Burki EA, Picchi MA, Stidley CA, Prados MD, Costello JF, Gilliland FD, Crowell RE, Belinsky SA. 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; 17:2014–2023. [PubMed: 21355081]

30. Ahrendt SA, Decker PA, Alawi EA, Zhu Yr YR, Sanchez-Cespedes M, Yang SC, Haasler GB, Kajdacsy-Balla A, Demeure MJ, Sidransky D. Cigarette smoking is strongly associated with mutation of the K-ras gene in patients with primary adenocarcinoma of the lung. Cancer. 2001; 92:1525–1530. [PubMed: 11745231]

31. Riely GJ, Kris MG, Rosenbaum D, Marks J, Li A, Chitale DA, Nafa K, Riedel ER, Hsu M, Pao W, Miller VA, Ladanyi M. Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. Clin Cancer Res. 2008; 14:5731–5734. [PubMed: 18794081]

32. Riely GJ, Marks J, Pao W. KRAS mutations in non-small cell lung cancer. Proc Am Thorac Soc. 2009; 6:201–205. [PubMed: 19349489]

33. Julin B, Bergkvist C, Wolk A, Akesson A. Cadmium in diet and risk of cardiovascular disease in women. Epidemiology. 2013; 24:880–885. [PubMed: 24030503]

34. Jeong I, Rhie J, Kim I, Ryu I, Jung PK, Park YS, Lim YS, Kim HR, Park SG, Im HJ, Lee MY, Won JU. Working Hours and Cardiovascular Disease in Korean Workers: A Case-control Study. Journal of occupational health. 2013

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

36. Dandri M, Locarnini S. New insight in the pathobiology of hepatitis B virus infection. Gut. 2012; 61(Suppl 1):i6–17. [PubMed: 22504921]

37. Perez OD. Appreciating the heterogeneity in autoimmune disease: multiparameter assessment of intracellular signaling mechanisms. Annals of the New York Academy of Sciences. 2005; 1062:155–164. [PubMed: 16461798]

38. Takamoto M, Kaburaki T, Mabuchi A, Araie M, Amano S, Aihara M, Tomidokoro A, Iwase A, Mabuchi F, Kashiwagi K, Shirato S, Yasuda N, Kawashima H, Nakajima F, Numaga J, Kawamura Y, Sasaki T, Tokunaga K. Common variants on chromosome 9p21 are associated with normal tension glaucoma. PLoS One. 2012; 7:e40107. [PubMed: 22792221]

39. Field AE, Camargo JCA, Ogino S. The merits of subtyping obesity: one size does not fit all. The Journal of American Medicial Association. 2013; 310:2147–2148.

40. Nguyen PL, Ma J, Chavarro JE, Freedman ML, Lis R, Fedele G, Fiore C, Qiu W, Fiorentino M, Finn S, Penney KL, Eisenstein A, Schumacher FR, Mucci LA, Stampfer MJ, Giovannucci E, Loda M. Fatty acid synthase polymorphisms, tumor expression, body mass index, prostate cancer risk, and survival. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2010; 28:3958–3964. [PubMed: 20679621]

41. Giovannucci EL, Liu Y, Leitzmann MF, Stampfer MJ, Willett WC. A prospective study of physical activity and incident and fatal prostate cancer. Archives of internal medicine. 2005; 165:1005–1010. [PubMed: 15883238]

42. Gates MA, Rosner BA, Hecht JL, Tworoger SS. Risk factors for epithelial ovarian cancer by histologic subtype. American Journal of Epidemiology. 2010; 171:45–53. [PubMed: 19910378]

43. Yang HP, Trabert B, Murphy MA, Sherman ME, Sampson JN, Brinton LA, Hartge P, Hollenbeck A, Park Y, Wentzensen N. Ovarian cancer risk factors by histologic subtypes in the NIH-AARP Diet and Health Study. International Journal of Cancer Journal International Du Cancer. 2012; 131:938–948. [PubMed: 21960414]

44. Kotsopoulos J, Terry KL, Poole EM, Rosner B, Murphy MA, Hecht JL, Crum CP, Missmer SA, Cramer DW, Tworoger SS. Ovarian cancer risk factors by tumor dominance, a surrogate for cell of origin. International journal of cancer Journal international du cancer. 2013; 133:730–739. [PubMed: 23364849]
45. Poole EM, Merritt MA, Jordan SJ, Yang HP, Hankinson SE, Park Y, Rosner B, Webb PM, Cramer DW, Wentzensen N, Terry KL, Tworoger SS. Hormonal and reproductive risk factors for epithelial ovarian cancer by tumor aggressiveness. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2013; 22:429–437.
46. Ferrari P, McKay JD, Jenab M, Brennan P, Canzian F, Vogel U, Tjonneland A, Overvad K, Tolstrup JS, Boutron-Ruault MC, Clavel-Chapelon F, Morois S, Kaaks R, Boeing H, Bergmann M, Trichopoulou A, Katsoulis M, Trichopoulos D, Krogh V, Panico S, Sacerdote C, Palli D, Tumino R, Peeters PH, van Gils CH, Bueno-de-Mesquita B, Vrieling A, Lund E, Hjartaker A, Agudo A, Suarez LR, Arriola L, Chirlaque MD, Ardanaz E, Sanchez MJ, Manjer J, Lindkvist B, Hallmans G, Palmqvist R, Allen N, Key T, Khaw KT, Slimani N, Rinaldi S, Romieu I, Boffetta P, Romaguera D, Norat T, Riboli E. Alcohol dehydrogenase and aldehyde dehydrogenase gene polymorphisms, alcohol intake and the risk of colorectal cancer in the European Prospective Investigation into Cancer and Nutrition study. European Journal of Clinical Nutrition. 2012; 66:1303–1308. [PubMed: 23149980]
47. Goldbohm RA, Van den Brandt PA, Van't Veer P, Dorant E, Sturmans F, Hermus RJ. Prospective study on alcohol consumption and the risk of cancer of the colon and rectum in the Netherlands. Cancer Causes & Control. 1994; 5:95–104. [PubMed: 8167268]
48. Bagnardi V, Blangiardo M, La Vecchia C, Corrao G. A meta-analysis of alcohol drinking and cancer risk. British Journal of Cancer. 2001; 85:1700–1705. [PubMed: 11742491]
49. Moskal A, Norat T, Ferrari P, Riboli E. Alcohol intake and colorectal cancer risk: A dose-response meta-analysis of published cohort studies. International Journal of Cancer. 2007; 120:664–671. [PubMed: 17096321]
50. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. World Cancer Research Fund, American Institute for Cancer Research; 2007.
51. Colditz GA, Winn DM. Criteria for the evaluation of large cohort studies: an application to the nurses' health study. J Natl Cancer Inst. 2008; 100:918–925. [PubMed: 18577745]
52. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. Nat Rev Cancer. 2005; 5:388–396. [PubMed: 15864280]
53. Robison LL, Armstrong GT, Boice JD, Chow EJ, Davies SM, Donaldson SS, Green DM, Hammond S, Meadows AT, Mertens AC, Mulvihill JJ, Nathan PC, Neglia JP, Packer RJ, Rajaraman P, Sklar CA, Stovall M, Strong LC, Yasui Y, Zeltzer LK. The Childhood Cancer Survivor Study: a National Cancer Institute-supported resource for outcome and intervention research. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2009; 27:2308–2318. [PubMed: 19364948]
54. Colditz GA. Ensuring long-term sustainability of existing cohorts remains the highest priority to inform cancer prevention and control. Cancer Causes Control. 2010; 21:649–656. [PubMed: 20063074]
55. Hughes LA, Williamson EJ, van Engeland M, Jenkins MA, Giles G, Hopper J, Southey M, Young J, Buchanan D, Walsh M, Van den Brandt PA, Goldbohm RA, Weijenberg MP, English DR. Body size and risk for colorectal cancers showing BRAF mutation or microsatellite instability: a pooled analysis. Int J Epidemiol. 2012
56. Ogino S, Giovannucci E. Commentary: Lifestyle factors and colorectal cancer microsatellite instability--molecular pathological epidemiology science, based on unique tumour principle. Int J Epidemiol. 2012; 41:1072–1074. [PubMed: 22596930]
57. Kalbfleisch, JD, Prentice, RL. The statistical analysis of failure time data. Wiley; 1980.
58. 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; 34:541–554. [PubMed: 373811]
59. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. Journal of the American Statistical Association. 1983; 78:1–12.

60. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. American Journal Of Epidemiology. 1997; 145:72–80. [PubMed: 8982025]
61. Commenges D, Letenneur L, Joly P, Alioum A, Dartigues JF. Modelling age-specific risk: application to dementia. Statistics In Medicine. 1998; 17:1973–1988. [PubMed: 9777690]
62. Therneau, TM, Grambach, PM. In The Counting Process Form of a Cox Model. Springer; New York, New York: 2000. The Counting Process Form of a Cox Model.
63. Anderson, TW. Introduction to multivariate statistics. John Wiley and Sons; New York, NY: 1984.
64. Lunn M, McNeil D. Applying Cox regression to competing risks. Biometrics. 1995; 51:524–532. [PubMed: 7662841]
65. Vaeth M. On The Use Of Walds Test In Exponential-Families. International Statistical Review. 1985; 53:199–214.
66. Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, Yeager M, Chung CC, Chanock SJ, Chatterjee N. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet. 2012; 90:821–835. [PubMed: 22560090]
67. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990; 1:43–46. [PubMed: 2081237]
68. Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data - Reply. American Journal Of Epidemiology. 1997; 145:85–85.
69. Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. American Journal of Epidemiology. 1995; 142:904–908. [PubMed: 7572970]
70. Goodman SN. Multiple comparisons, explained. American Journal of Epidemiology. 1998; 147:807–812. [PubMed: 9583709]
71. Thomas DCSJ, Dewar R, Robins J, Goldberg M, Armstrong BG. The problem of multiple inference in studies designed to generate hypotheses. American Journal of Epidemiology. 1985:1080–1095. [PubMed: 4061442]
72. Thompson JR. Invited commentary: Re: “Multiple comparisons and related issues in the interpretation of epidemiologic data”. American Journal of Epidemiology. 1998; 147:801–806. [PubMed: 9583708]
73. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Annals of Statistics. 2003; 31:2013–2035.
74. Day RW, Quinn GP. Comparisons of treatments after an analysis of variance in ecology. Ecological Monographs. 1989; 59:433–463.
75. Story JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B. 2002; 64:479–498.
76. Miettinen, OS. Theoretical epidemiology : principles of occurrence research in medicine. Wiley; New York: 1985.
77. Breslow, NE, Day, NE. Statistical Methods in Cancer research. Vol I: The Analysis of Case-Control Studies. IARC Scientific Publications; Lyon: 1980.
78. 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; 105:130–140. [PubMed: 23175808]
79. Chatterjee N. A Two-Stage Regression Model for Epidemiological Studies With Multivariate Disease Classification Data. Journal of the American Statistical Association. 2004; 99:127–138.
80. Chatterjee N, Sinha S, Diver WR, Feigelson HS. Analysis of cohort studies with multivariate and partially observed disease classification data. Biometrika. 2010; 97:683–698. [PubMed: 22822252]
81. Rosner B, Glynn RJ, Tamimi RM, Chen WY, Colditz GA, Willett WC, Hankinson SE. Breast Cancer Risk Prediction with Heterogeneous Risk Profiles According to Breast Cancer Tumor Markers. American Journal of Epidemiology. 2013; 15:296–308.

82. Wang M, Kuchiba A, Ogino S. A Meta-Regression Method for Studying Etiological Heterogeneity Across Disease Subtypes Classified by Multiple Biomarkers. American Journal Of Epidemiology. 2015; 182:263– 270. [PubMed: 26116215]
83. Stram DO. Meta-analysis of published data using a linear mixed-effects model. Biometrics. 1996; 52:536–544. [PubMed: 8672702]
84. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychological Methods. 1998; 3:486–504.
85. Ritz J, Demidenko E, Spiegelman D. Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. J Stat Plann Inference. 2008; 138:1919–1933.
86. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics In Medicine. 2002; 21:589–624. [PubMed: 11836738]
87. Schernhammer ES, Giovannucci E, Kawasaki T, Rosner B, Fuchs CS, Ogino S. Dietary folate, alcohol and B vitamins in relation to LINE-1 hypomethylation in colon cancer. Gut. 59:794–799. [PubMed: 19828464]
88. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics. 1975; 31:643–649. [PubMed: 1100136]
89. Langholz B, Goldstein L. Conditional logistic analysis of case-control studies with complex sampling. Biostatistics. 2001; 2:63–84. [PubMed: 12933557]
90. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010; 36:1–48.
91. 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. [PubMed: 25532962]
92. Wei LJ, Lin DY, Weissfeld L. Regression-analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association. 1989; 84:1065–1073.
93. Colussi D, Brandi G, Bazzoli F, Ricciardiello L. Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. International journal of molecular sciences. 2013; 14:16365–16385. [PubMed: 23965959]
94. 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 journal of gastroenterology: WJG. 2014; 20:6055. [PubMed: 24876728]
95. Barrow TM, Michels KB. Epigenetic epidemiology of cancer. Biochemical and biophysical research communications. 2014; 455:70–83. [PubMed: 25124661]
96. Ogino SCP, 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 Cont Accepted. 2015
97. Shuji Ogino RN, Vander Weele Tyler J, Wang Molin, Nishi Akihiro, Paul Lochhead ZRQ, Zhang Xuehong, Wu Kana, Nan Hongmei, Yoshida Kazuki, Milner Danny A Jr, Chan Andrew T, Field Alison E, Camargo Carlos A Jr, Williams Michelle A, Giovannucci Edward. Molecular Pathological Epidemiology Is Essential in Studying Neoplastic and Non-neoplastic Diseases in the Era of Precision Medicine. Epidemiology. 2015
98. Collins FS, Varmus H. A new initiative on precision medicine. New England Journal of Medicine. 2015; 372:793–795. [PubMed: 25635347]
99. Bayer R, Galea S. Public Health in the Precision-Medicine Era. New England Journal of Medicine. 2015; 373:499–501. [PubMed: 26244305]

Appendix

Proof for the fact that the case-case model (7) is a parametric transformation of the fundamental model (1).

Replacing $\lambda_j(t)$ by its definition $\lim_{\Delta t \to 0} P(t \le T \le t + \Delta t, Y = j \mid T \ge t, \boldsymbol{X}, \boldsymbol{W})/\Delta t$, where T is the time to disease occurrence, and replacing $\lambda_{0j}(t)$ by $\lim_{\Delta t \to 0} P(t \le T \le t + \Delta t,\ Y = j \mid T \ge t, \boldsymbol{X} = \boldsymbol{0}, \boldsymbol{W} = \boldsymbol{0})/\Delta t$ in Model (1), we have

$$d_{j_1 j_2}(t, \boldsymbol{X}, \boldsymbol{W}) = d_{j_1 j_2}(\mathrm{t}, \boldsymbol{X} = 0, \boldsymbol{W} = 0) \exp\{(\boldsymbol{\beta}_{j_1} - \boldsymbol{\beta}_{j_2})\boldsymbol{X} + (\boldsymbol{\theta}_{j_1} - \boldsymbol{\theta}_{j_2})\boldsymbol{W}\},$$

where

$d_{j_1 j_2}(t, \boldsymbol{X}, \boldsymbol{W}) = \lim_{\Delta t \to 0} P(t \le T \le t + \Delta t, Y = j_1 \mid T \ge t,\ \boldsymbol{X}, \boldsymbol{W})/P(t \le T \le t + \Delta t, Y = j_2 \mid T \ge t, . \boldsymbol{X}, \boldsymbol{W})$

Since d_{j1j2} (t,$\boldsymbol{X}$,$\boldsymbol{W}$) can be written as $P(Y = j_1 | T = t, \boldsymbol{X}, \boldsymbol{W})/P(Y = j_2 | T = t, \boldsymbol{X}, \boldsymbol{W})$, Model (1) can be transformed to

$$\frac{P(Y = j_1 \mid T = t, \boldsymbol{X}, \boldsymbol{W})}{P(Y = j_2 \mid T = t, \boldsymbol{X}, \boldsymbol{W})} = d_{j_1 j_2}(t, \boldsymbol{X} = 0, \boldsymbol{W} = 0) \exp\{(\boldsymbol{\beta}_{1 j_1} - \boldsymbol{\beta}_{1 j_2})\boldsymbol{X} + (\boldsymbol{\theta}_{j_1} - \boldsymbol{\theta}_{j_2})\boldsymbol{W}\}.$$

Assuming the reference subtype is Type 1, i.e., $j_2 = 1$,$\boldsymbol{\beta}_j - \boldsymbol{\beta}_1 = \boldsymbol{\alpha}_j$, and $\boldsymbol{\theta}_j - \boldsymbol{\theta}_1 = \boldsymbol{\tau}_j$, we have $\frac{P(Y = j \mid T = t, \boldsymbol{X}, \boldsymbol{W})}{P(Y = 1 \mid T = t, \boldsymbol{X}, \boldsymbol{W})} = d_{j1}(t, \boldsymbol{X} = 0, \boldsymbol{W} = 0) \exp(\boldsymbol{\alpha}_j \boldsymbol{X} + \tau_j \boldsymbol{W})$. This is the model for nominal polytomous logistic regression for the case-case study with age at disease incidence, T, as a stratification factor. We can use the conditional logisitic regression model to eliminate the nuisance parameter d_{j1} (t,$\boldsymbol{X}$ = 0,$\boldsymbol{W}$ = 0), the log-transformation of which is the intercept for a stratum with $T = t$. If T is treated as a covariate instead of a stratification factor, and assuming T is an element of $\boldsymbol{W}$, the model above becomes Model (7), in which successful adjustment for T will rely on including an appropriate form for T in the model; for example, a spline function of T could be included to model a non-linear relationship between T and $\log \frac{P(Y = j \mid T = t, \boldsymbol{X}, \boldsymbol{W})}{P(Y = 1 \mid T = t, \boldsymbol{X}, \boldsymbol{W})}$.

Table 1

Data for the unmatched case-control study for a binary exposure and two subtypes

	E	$\bar{E}$	**Total**
Subtype 1	a_1	b_1	m_{11}
Subtype 2	a_2	b_2	m_{12}
Controls	c	d	

Notations: E is exposed group; $\bar{E}$ is unexposed group.

Table 2

Original data set (cohort) for the colon cancer example

id	time	cancer	period	agemo	alcohol	asp
1	20	0	1	560	0.15	0
1	23	0	2	580	0.15	1
1	16	1	3	603	0.15	0
...						
2	23	0	1	606	0	0
2	21	0	2	623	0	0
2	19	0	3	644	0	1
2	25	0	4	663	0	1
...						

Table 3

Augmented data set (cohort_aug) for id=1 in the colon cancer example

id	time	cancer	period	agemo	alcohol	asp	censor	type	alcohol_1	alcohol_2	alcohol_3
1	20	0	1	560	0.15	0	0	1	0.15	0	0
1	20	0	1	560	0.15	0	0	2	0	0.15	0
1	20	0	1	560	0.15	0	0	3	0	0	0.15
1	23	0	2	580	0.15	1	0	1	0.15	0	0
1	23	0	2	580	0.15	1	0	2	0	0.15	0
1	23	0	2	580	0.15	1	0	3	0	0	0.15
1	16	1	3	603	0.15	0	1	1	0.15	0	0
1	16	1	3	603	0.15	0	0	2	0	0.15	0
1	16	1	3	603	0.15	0	0	3	0	0	0.15

Table 4

Original data set (ncaco) for three matched sets for a nested case-control study

id	matchid	cancer	age	alcohol	asp
1	1	0	56	0.15	0
2	1	0	56	2.29	1
3	1	1	56	0.85	0
4	2	0	71	2.29	1
5	2	0	71	0.15	1
6	2	2	71	2.29	1
7	3	0	53	0.85	0
8	3	0	53	0	1
9	3	3	53	0.85	0

Table 5

Augmented data set (ncaco_aug) for three matched sets for a nested case-control study

id	matchid	cancer	age	alcohol	alcohol_1	alcohol_2	alcohol_3	asp
1	1	0	56	0.15	0.15	0	0	0
2	1	0	56	2.29	2.29	0	0	1
3	1	1	56	0.85	0.85	0	0	0
4	2	0	71	2.29	0	2.29	0	1
5	2	0	71	0.15	0	0.15	0	1
6	2	2	71	2.29	0	2.29	0	1
7	3	0	53	0.85	0	0	0.85	0
8	3	0	53	0	0	0	0	1
9	3	3	53	0.85	0	0	0.85	0

Table 6

Augmented data set (uncaco_aug) for the first three ids for an unmatched case-control study

id	censor	type	age	alcohol	alcohol_1	alcohol_2	alcohol_3	asp
1	0	1	56	0.15	0.15	0	0	0
1	0	2	56	0.15	0	0.15	0	0
1	0	3	56	0.15	0	0	0.15	0
2	0	1	56	2.29	2.29	0	0	1
2	0	2	56	2.29	0	2.29	0	1
2	0	3	56	2.29	0	0	2.29	1
3	1	1	56	0.85	0.85	0	0	0

Table 7

Alcohol intake in relation to LINE-1 methylation colon cancer subtype incidence and subtype heterogeneity evaluations by study design

Study design [Sample size]	Unconstrained model		Constrained model	
	Subtype-specific RR/OR (95% CI); LINE-1 high, medium, low	P value for heterogeneity (categorical, ordinal)	Subtype-specific RR/OR (95% CI); LINE-1 high, medium, low	P value for heterogeneity (categorical, ordinal)
Prospective cohort [268 cases; 47,363 men; 701,119 person- years]	1.00 (0.79–1.26)		1.00 (0.80–1.25)	
	1.57 (1.27–1.94)	0.017	1.54 (1.25–1.89)	0.017
	1.36 (1.05–1.77)	0.124	1.36 (1.06–1.76)	0.109
Nested case-control [268 matched sets; 1:2 matching]	1.01 (0.74–1.39)		0.98 (0.73–1.31)	
	1.51 (1.07–2.15)	0.083	1.43 (1.07–1.92)	0.139
	1.86 (1.09–3.17)	0.031	1.39 (0.97–1.99)	0.099
Unmatched case-control [268 cases; 533 controls]	0.96 (0.74–1.24)		0.94 (0.72–1.22) *	
	1.55 (1.21–1.99)	0.014	1.56 (1.21–2.01) *	0.023
	1.30 (0.96–1.77)	0.180	1.30 (0.95–1.78) *	0.169
Case-case [268 cases]	1.00		1.00	
	1.68 (1.18–2.39) **	0.014	1.55 (1.15–2.08) **	0.016
	1.39 (0.95–2.04) **	0.097	1.32 (0.95–1.84) **	0.105

* Based on the conditional logistic regression method described in Section 3.3.2.

** Ratio of the OR for LINE-1 medium or low colon cancer to the OR for LINE-1 high colon cancer. The following variables were adjusted for in all analyses: current aspirin use (≥2 tablets/week or less), body mass index (kg/m^2) (<21, 21–22.9, 23–24.9, 25–29.9, 30+), history of colorectal screening (yes/no), physical activity in metabolic equivalent of tasks (quintiles), history of colorectal polyps (yes/no), family history of colon cancer (yes/no), smoking (pack-years), red meat intake (quintiles), multivitamin use (yes/no), calcium intake (quintiles) and folate intake (quintiles).