Statistical methods for the study of etiologic heterogeneity

Emily C. Zabor

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Public Health
in the Department of Biostatistics at the Mailman School of Public Health

COLUMBIA UNIVERSITY

2018

ABSTRACT

Statistical methods for the study of etiologic heterogeneity

Emily C. Zabor

Traditionally, cancer epidemiologists have investigated the causes of disease under the premise that patients with a certain site of disease can be treated as a single entity. Then risk factors associated with the disease are identified through case-control or cohort studies for the disease as a whole. However, with the rise of molecular and genomic profiling, in recent years biologic subtypes have increasingly been identified. Once subtypes are known, it is natural to ask the question of whether they share a common etiology, or in fact arise from distinct sets of risk factors, a concept known as etiologic heterogeneity. This dissertation seeks to evaluate methods for the study of etiologic heterogeneity in the context of cancer research and with a focus on methods for case-control studies. First, a number of existing regression-based methods for the study of etiologic heterogeneity in the context of pre-defined subtypes are compared using a data example and simulation studies. This work found that a standard polytomous logistic regression approach performs at least as well as more complex methods, and is easy to implement in standard software. Next, simulation studies investigate the statistical properties of an approach that combines the search for the most etiologically distinct subtype solution from high dimensional tumor marker data with estimation of risk factor effects. The method performs well when appropriate up-front selection of tumor markers is performed, even when there is confounding structure or high-dimensional noise. And finally, an application to a breast cancer case-control study demonstrates the usefulness of the novel clustering approach to identify a more risk heterogeneous class solution in breast cancer based on a panel of gene expression data and known risk factors.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I want to start by saying a sincere thank you to my dissertation advisors, Dr. Shuang Wang and Dr. Colin Begg, who have supported me and guided me through this research for the last several years. Their subject area expertise was invaluable throughout this process, and I would not have made such steady progress without learning from them how to continuously set small and achievable goals and then constantly work to achieve them.

I also want thank the rest of my dissertation committee members, including Dr. Min Qian, Dr. Mary Beth Terry, and the committee chair, Dr. Yuanjia Wang. Their feedback on this dissertation has led to meaningful improvements. I have been lucky to receive support from others outside of my committee throughout this process as well, including Dr. Venkatraman Seshan at Memorial Sloan Kettering Cancer Center and my collaborators at the University of North Carolina, Halei Benefield and Dr. Melissa Troester.

More personally I want to thank my mother, Kathryn Craig, for leading by example and getting her own PhD while working full time several years ago, and my father, Stephen Zabor, for always supporting and encouraging my academic endeavors. For stress relief, no one was better than Mike (the cat), who never failed to sit in my lap and let me pet him, and always made me feel more relaxed. And lastly, I thank my husband Rick Thayer for his love, understanding and support throughout this process.

Chapter 1

Introduction

1.1 Introduction to etiologic heterogeneity

The basic goal of most epidemiologic research is to investigate the prevalence and cause of disease. Traditionally, epidemiologists have organized this line of research under the premise that patients with a certain disease share an underlying etiology, or cause. In this framework, the disease is treated as a single entity, and investigators have sought to identify risk factors that are associated with the disease using case-control or cohort study designs. In the early 1990s epidemiologists began to focus attention on the possibility that risk factors, particularly occupational exposures and environmental carcinogens, may lead to biologically-distinct subtypes of disease with respect to individual somatic mutations (see, for example, Taylor *et al.* (1994)). More recently attention has increasingly focused on identifying subtypes of disease according to disease characteristics such as molecular markers or pathologic features. This has been especially true in cancer research because of the growing use of molecular and genomic profiling, which give researchers access to many more ways in which to classify a tumor. It is now widely accepted that many cancers, including but not limited to breast (Perou *et al.*, 2000; Sorlie *et al.*, 2001; Sotiriou *et al.*, 2003; Gaudet *et al.*, 2011), lung (Ahrendt *et al.*, 2001; Riely *et al.*, 2008; Marsit *et al.*, 2009), colorectal (Limsui *et al.*, 2010; Ogino *et al.*, 2011), ovarian (Merritt *et al.*, 2013), and endometrial (Brinton *et al.*, 2013; Schildkraut *et al.*, 2013) cancers, are comprised of specific molecular subtypes. As these subtypes are identified, it is natural to ask the question of whether they share a common etiology, or in fact arise from distinct sets of risk factors. The concept of differing risk factors across subtypes of disease is known as etiologic heterogeneity.

1.2 Introduction to statistical methods

There are many challenges related to the study of etiologic heterogeneity, and statistical methods are needed not only to detect the presence of heterogeneity, but also to quantify the extent of that heterogeneity. One challenge is the possibly high dimension of the data, which may include information from multiple molecular profiling platforms such as expression, copy number, mutation, and methylation data. Further, as more evidence accumulates for subtypes of cancers with distinct risk profiles according to known risk factors, it is natural to ask whether undiscovered risk factors will also exhibit differential associations across subtypes. An epidemiologic investigation of etiologic heterogeneity such as a case-control study would naturally be subject to the constraints of smaller subtype sample sizes as compared to the aggregate case group, as well as the prospect of false discovery due to the increasing number of statistical comparisons being made. However, such investigation also serves to benefit from a potentially larger effect size in at least one subtype and improved risk prediction accuracy for all patients.

Before undergoing the task of addressing these and many other statistical challenges, it is important to identify what advantage, if any, researchers stand to gain by considering subtypes of disease as opposed to an aggregate case group. Begg and Zabor (2012) investigated the statistical implications of these trade-offs in the choice between a traditional case-control design versus one that further classifies cases into subtypes using a simulation study to examine statistical power under various study design scenarios. This study found that over a range of risk factor prevalences and overall case-control odds ratios, only modest heterogeneity was needed before a study design that accounts for subtypes achieved equivalent power to a traditional case-control approach that considers all cases in aggregate. This result provides a practical motivation to pursue development of statistical methods for the study of etiologic heterogeneity.

1.2.1 Traditional approach

Early investigations of etiologic heterogeneity relied on standard statistical methods. Typically, an investigator would have data on cases and controls. The cases would then be divided into a small number of pre-determined subtypes. These subtypes could be based on a single disease characteristic, or on combinations thereof. Then, associations with risk factors could be examined using polytomous logistic regression (Dubin and Pasternack, 1986). Polytomous logistic regression allows for the simultaneous estimation of subtype-specific regression parameters, and differences in risk factor effects across subtypes can be tested. Data on subtypes from cohort studies can be similarly analyzed using competing risks regression, where those who have not yet developed the disease at the end of follow-up are censored, and the subtypes comprise each of the possible competing events.

Shortly thereafter, Begg and Zhang (1994) proposed that in fact all of the information needed to test for etiologic heterogeneity is contained in the cases, so it is not necessary to include data on controls. This is an important idea since often epidemiologists have access to case series data, with no data on a control population, especially in hospital research settings. Polytomous logistic regression can still be used, but now one of the subtypes is selected to serve as the reference group. Choice of reference subtype could be determined based on sample size, as selecting the subtype with the largest sample size as the reference group will produce the most stable estimates, or based on subject-area interest. Regression parameters are more difficult to interpret in the case-only setting due to the lack of a control group, as the control group allows for interpretations with respect to the non-diseased population, but the resulting regression parameters still allow for tests of differences in risk factor effects across subtypes.

These approaches have the advantage of being straightforward and easy to implement and interpret. However, they become inefficient or impossible to implement as the number of subgroups grows, and they do not provide direct information about the extent to which subtypes are etiologically distinct. Furthermore, the subtypes must be pre-specified, and subtype assignment occurs in isolation from the analysis of risk factors.

1.2.2 Recent advances

In recent years a number of new methods have been proposed, some of which are extensions of the traditional approach and others that are more novel.

1.2.2.1 Approaches that require pre-specified subtypes

Chatterjee (2004) proposed a two-stage regression model to address the analytic issue of having a potentially large number of subtypes that cannot be handled by standard polytomous logistic regression techniques while simultaneously allowing an investigator to determine which disease characteristics play a role in defining the etiologically distinct subtypes. Use of a second-stage model for the subtype-specific regression parameters of the first-stage polytomous logistic regression model reduces the dimensionality problem while also providing a testing strategy for etiologic heterogeneity. Interaction effects can be flexibly included in the second-stage model, though most of the time interest will be limited to the case of first-order interaction effects, which imply an additive model such that the effect of one disease characteristic does not depend on other disease characteristics. This model leads to a conditional interpretation, such that the degree of etiologic heterogeneity with respect to one disease characteristic is interpreted in the context of all of the other disease characteristics being held constant.

For estimation of the custom two-stage model, Chatterjee (2004) suggests a semiparametric approach that leaves the intercept parameters completely unspecified and limits the second-stage model to the regression parameters of interest. The reasons for this are twofold: 1) the intercept parameters themselves are not of scientific interest and 2) simulation studies revealed that mis-specification of the intercept parameters can lead to substantial bias in the regression parameters of interest. Estimation can be carried out using a proposed pseudo-conditional-likelihood approach, which Chatterjee (2004) shows to be asymptotically valid and computationally efficient in a simulation study; however, the estimation procedure requires customized programming.

While the original two-stage modeling approach proposed by Chatterjee (2004) is appropriate for case-control data, epidemiologic research is often conducted using data from prospective cohort studies. Chatterjee *et al.* (2010) extended the earlier work by proposing a two-stage modeling approach for use with data from cohort studies. Now the first-stage model is a competing risks regression model rather than a polytomous logistic regression model, and the second-stage model is the same as before. An additional extension of this work was to the case of missing data. It is common in epidemiologic studies for there to be missing disease characteristic data, and the proposed estimation procedure can handle missing data with a missing-at-random assumption through an extension of the estimating equation approach of Goetghebeur and Ryan (1995). Similarly to the unspecified intercept parameters in the case-control setting, the baseline hazard is unspecified. Chatterjee *et al.* (2010) demonstrate the asymptotic unbiasedness of the estimator and also show through a simulation study that the estimator performs well in most cases, even when the baseline hazard is mis-specified. These methods focus on testing for the association between a single risk factor and a single disease characteristic, when all other disease characteristics are held constant.

Rosner *et al.* (2013) proposed a single-stage approach to examining the effect of risk factors on disease characteristics while controlling for other disease characteristics. This has traditionally been accomplished by creating subtypes based on all possible combinations of the disease characteristics. However, as the number of disease characteristics available for study grows, this becomes increasingly infeasible. A regression approach that assesses interaction effects of risk factors with specific disease characteristics while controlling for levels of other disease characteristics for use in the setting of cohort studies was proposed, based on a variation of a cause-specific hazard model (Rosner *et al.*, 2013). This method is particularly appropriate when the assumption of independence of the effects of individual disease characteristics on the baseline hazard does not hold, as it allows for incorporation of interaction effects between individual disease characteristics, thus accounting for the common problem of correlation between disease characteristics. Additionally, it has the appeal of being implemented using standard software. This method contrasts with that of Chatterjee *et al.* (2010) in that it does not imply independence of the effects of disease characteristics on the baseline hazard and no custom estimation procedure is required.

Similarly to the problem with polytomous logistic regression, the approach of Rosner *et al.* (2013) can become computationally infeasible as the number of disease characteristics defining the subtypes grows. In response, Wang *et al.* (2015) introduced a two-stage version of the Rosner *et al.* (2013) method, with applications to cohort studies, nested case-control studies, and unmatched case-control studies. For a cohort study or nested case-control study, the first stage of the analysis uses a cause-specific hazards model. For an unmatched case-control study, the first stage is instead a polytomous logistic regression model. A second-stage analysis then allows one to test whether the association between the risk factor of interest and each subtype differs according to the levels of the individual disease characteristics that comprise the subtypes. This fixed effects two-stage method differs from that of Chatterjee *et al.* (2010) in the estimation approach rather than in the intrinsic model setup. To account for additional heterogeneity that may not be captured by the available disease characteristics, Wang *et al.* (2015) proposed an alternative approach that incorporates a random intercept in the second-stage model to account for unmeasured variance.

1.2.2.2 Approaches that search for the most heterogeneous subtypes

The methods described so far can be used when there are pre-defined disease characteristics of interest. While this approach can tell you whether individual disease characteristics demonstrate etiologic heterogeneity with respect to each individual risk factor, there is no intrinsic way to quantify the overall degree of etiologic heterogeneity. To provide epidemiologic investigators with a method that could be used to search for the set of subtypes that best explain etiologic heterogeneity, Begg *et al.* (2013) introduced a scalar heterogeneity measure that can be used to compare candidate subtyping schemes based on their overall degree of etiologic heterogeneity. The goal is to identify the set of subtypes that best explains the etiologic heterogeneity of the disease of interest by defining a measure of etiologic heterogeneity. The strategy is based on the use of the coefficient of variation, which is a measure of risk heterogeneity in the population. The proposed measure can be used to compare different subtyping options and determine which option demonstrates the highest degree of etiologic heterogeneity. Once the optimal subtype solution is identified, then traditional methods such as polytomous logistic regression can be used to investigate risk factor effects. This method not only integrates the classification of cases into subtypes with the examination of risk factor effects, but also provides a scalar measure of the extent to which the subtypes are etiologically distinct with respect to the entire set of risk factors simultaneously.

All of the preceding methods have used a regression framework to approach the study of etiologic heterogeneity. Yu *et al.* (2015) proposed an approach to the study of etiologic heterogeneity that alternatively uses binary recursive partitioning. The framework for this method is that disease characteristics can be combined to form subtypes of disease. Subjects are initially split into two groups. Then, at each split, one of the two terminal nodes is selected for further splitting based on a "goodness-of-split" criteria. After all splits are made, there is a set of candidate tree models, each of different size and complexity. The terminal nodes of each candidate tree can then be considered as candidate subtypes, and used as the outcome in polytomous logistic regression to accomplish effect estimation and heterogeneity testing. Because the candidate trees have been selected based on a search of the data for the best split, to avoid overoptimism Yu *et al.* (2015) propose a resampling-based procedure to estimate the p-value for each candidate tree, where the risk factor of interest is randomly permuted among subjects with the same observed covariates. Finally, the candidate tree with the smallest p-value is selected as the final definition for disease subtyping. This approach provides a way to subtype cases in order to get a result that maximizes etiologic heterogeneity by seeking to group cases into subtypes based on the association between the risk factor of interest and a binary grouping of cases at each step. This approach does not provide a way to quantify the extent of etiologic heterogeneity according to the various candidate trees.

Whereas the approaches discussed so far assume that each patient belongs to only a single subtype, Schildkraut *et al.* (2013) developed a method to examine the association between molecular subtypes of ovarian cancer and patient demographics and epidemiologic risk factors that allows for the possibility that membership in a subtype is not always rigid and the tumor of each patient could have characteristics of more than one subtype. This method is specifically motivated in cancer research by the increasing interest in intra-tumor heterogeneity. The method was applied to identify subtypes of ovarian cancer using consensus k-means clustering of gene expression data, and class prediction by k-nearest neighbors and diagonal linear discriminate analysis, with k, the number of subtypes, determined by the gap statistic (Tothill *et al.*, 2008). Once the subtypes are determined, each case is assigned a score for each subtype. The score is based on a weighted sum of the normalized expression values for overexpressed probes in the subtype minus a weighted sum of the normalized expression values for underexpressed probes in the subtype (Sfakianos *et al.*, 2013). High positive scores indicate that a case is likely to belong to that subtype whereas low scores indicate that the case is likely not to belong to that subtype. The moderate positive correlations observed between some pairs of the six subtypes provide evidence that the subtypes may not be mutually exclusive. The scores for each patient were then normalized, and Schildkraut *et al.* (2013) applied multivariate response multiple regression models with unstructured covariance to assess the relationship between cancer risk factors and the identified molecular subtypes. This approach accounts for the correlation between scores for the various subtypes from a single patient, and the unstructured covariance does not make any assumptions about the form that correlation takes. Note that this is a case-only analysis, and no controls or cancer-free participants are used. Testing for heterogeneity is conducted by incorporating an interaction term for the subtypes by the risk factor of interest. A significant interaction effect suggests that the risk factor is associated with score differently depending on subtype whereas a non-significant interaction effect suggests that the effect of the risk factor on score does not differ based on subtype. This approach is applicable to situations where tumors are being classified into subtypes based on a single molecular platform of information, rather than potentially multiple platforms or a combination of, for example, gene expression data and histo-pathologic features.

1.3 Introduction to breast cancer heterogeneity

Breast cancer incidence in the United States remained stable from 2002-2011, at an average age-standardized incidence of 122.8 cases per 100,000 women across all races, and breast cancer remains one of the three leading cancer causes of death among women, together with lung and colorectal cancers (Kohler *et al.*, 2015). Breast tumors are biologically diverse, and growing evidence over the past two decades supports the notion that breast cancer should not be considered a single disease, but rather a group of diseases with distinct etiologies, treatments, and prognoses. Numerous studies have used genomic data to classify breast cancers into subtypes (Perou *et al.*, 2000; Sorlie *et al.*, 2001; Sotiriou *et al.*, 2003), with most classification schemes relying on hierarchical clustering of microarray data. The most well-accepted subtyping scheme consists of four main subtypes of breast cancer, known as luminal A, luminal B, HER2-type, and basal-like subtypes, based on the PAM50 panel of gene expression data (Sorlie *et al.*, 2001). These subtypes are well-approximated by four subtypes of breast cancer based solely on immunohistochemical (IHC) markers for estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Luminal A tumors are defined as ER positive (+) and/or PR+, and HER2 negative (-), luminal B tumors are defined as ER+ and/or PR+ and HER2+, HER2-type tumors are defined as ER- and PR- and HER2+, and basal-like tumors are triple negative, defined as ER- and PR- and HER2-.

Numerous epidemiologic studies have investigated etiologic differences according to these subtypes, primarily in the context of case-control studies (Millikan *et al.*, 2008; Phipps *et al.*, 2008a,b; Yang *et al.*, 2007). Differential risk factor effects have consistently been identified, particularly with respect to body size, race, and hormonal risk factors such as menopausal status. To date such analyses have been conducted in two distinct phases, where identification of biologic subtypes of breast cancer occurs completely separately from epidemiologic investigation of differential risk factor effects. As a result, it is unclear if the most etiologically distinct subtypes of breast cancer have yet been identified.

1.4 Summary of introduction

This dissertation seeks to accomplish the following goals. Chapter 2 will explicate the similarities and differences among regression-based statistical approaches to the study of etiologic heterogeneity when there are pre-defined subtypes, including the standard polytomous logistic regression method introduced in Section 1.2, and the methods of Chatterjee (2004), Wang *et al.* (2015), and Rosner *et al.* (2013), introduced in Section 1.2.2.1. While these approaches utilize different modeling strategies, they all aim to test hypotheses about associations between risk factors and subtypes or individual disease characteristics. It is important to understand the intricacies of how these methods compare, and to enumerate the strengths and weaknesses of each. This will be accomplished through a data example and simulation studies. Next, Chapter 3 will explore the validity of the method for identifying optimally etiologicallly heterogeneous subtypes based on a scalar measure proposed by Begg *et al.* (2013), introduced in Section 1.2.2.2. This method is of particular interest as it is the only approach proposed to date the allows for quantification of the extent to which subtypes are etiologically distinct. While this method has been used in a number of applications, to date the statistical properties have not been rigorously studied. This will be accomplished through the use of simulation studies to explore the ability of the method to identify the truly etiologically heterogeneous subtypes, as quantified by the misclassification rate, under a variety of scenarios, and to examine the usefulness of upfront dimension reduction of the disease characteristic data. Finally, Chapter 4 will conduct a comprehensive application to data from the Carolina Breast Cancer Study, a breast cancer case-control study with available gene expression data on a subset of the cases, using the optimal D clustering method. Optimal subtype results will be compared to the traditional four classes of breast cancer as defined by IHC markers and the PAM50 gene expression panel, which were introduced in Section 1.3. This data application will elucidate the real-world methodologic challenges confronted by epidemiologists when seeking to study etiologic heterogeneity in the context of high dimensional disease characteristic data.

Chapter 2

Comparison of existing methods*

The results of this chapter show that when the number of tumor markers is small enough that the cross-classification of markers can be evaluated in the traditional polytomous logistic regression framework, then the statistical properties are at least as good as the more complex modeling approaches that have been proposed. The potential advantage of more complex methods is in the ability to accommodate multiple tumor markers in a model of reduced parametric dimension.

Epidemiologic questions of interest related to the study of etiologic heterogeneity may include 1) whether a risk factor of interest has the same effect across all subtypes of disease and 2) whether risk factor effects differ across levels of each individual disease characteristic by which the subtypes are defined. Early investigations of etiologic heterogeneity typically divided cases into a small number of pre-determined subtypes, based on a single molecular marker or pathologic feature, or on combinations thereof. Associations of specific subtypes with risk factors could be examined using polytomous logistic regression (Dubin and Pasternack, 1986). In recent years, however, a number of new statistical methods have been

*Note that the contents of this chapter have been published in Zabor and Begg, Statistics in Medicine 2017; 36:4050-60.

proposed for the study of etiologic heterogeneity.

In this chapter, four distinct available methods are compared: polytomous logistic regression; the two-stage meta-regression method proposed by Wang *et al.* (2015); the two-stage regression with simultaneous estimation approach proposed by Chatterjee (2004); and the stratified logistic regression approach of Rosner *et al.* (2013). These methods have very distinctive parametric structures and it is not immediately straightforward how results using the different methods align with each other. The goal is to reconcile the similarities among the methods, and to evaluate their statistical properties. To accomplish this, a simplified data example is employed to elucidate the interpretation of model parameters and available hypothesis tests, and a simulation study is performed to assess bias in effect size, type I error, and power.

2.1 Analytic framework

This chapter focuses solely on methods for the analysis of case-control data, though many of the approaches discussed can be applied in the context of other study designs. And because the simplified data example comes from breast cancer, throughout the disease characteristics that combine to form subtypes are referred to as "tumor markers," though notably these methods are generalizable to disease contexts besides cancer. Let i index study subjects, $i = 1, \ldots, N$, let k index tumor markers, $k = 1, \ldots, K$, let m index disease subtypes, $m = 0, \ldots, M$, where $m = 0$ denotes control subjects, and let p index risk factors, $p = 1, \ldots, P$. Initially, for simplicity, the focus is on a setting where there are two binary tumor markers, each of which can be either positive (+) or negative (-). These two tumor markers are cross-classified to form four disease subtypes (-/-, +/-, -/+, and +/+). Additionally, for conceptual simplicity in the primary exposition and simulations, the investigation is limited to the case of a single binary risk factor of interest. Therefore, the setting explored here has tumor markers $k = 1, 2$, disease subtypes $m = 1, 2, 3, 4$, and risk factor $p = 1$.

The first epidemiologic question of interest to be addressed is whether the risk factor of interest has the same effect across all subtypes of disease. This is frequently the primary question of interest in an investigation of etiologic heterogeneity and allows one to determine whether the risk factor of interest is associated with specific subtypes of disease. From each of the available methods, the parameters β_{pm} can be obtained, which represent the log odds ratio for a one-unit change in risk factor p for subtype m disease versus controls. In the case of four subtypes and one binary risk factor, there are four such log odds ratios $\beta_{11}, \beta_{12}, \beta_{13},$ and β_{14} (Table 2.1). Thus a test of the hypothesis $H_{0_\beta} : \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$ is of interest. A second epidemiologic question of specific interest is whether the risk factor effect differs across levels of each individual tumor marker. This question allows one to evaluate whether a specific tumor marker is in part responsible for observed differences in log odds ratios of the risk factor across the subtypes. To answer this question, estimates of parameters γ_{pk} are obtained, each of which represents the ratio of the log odds ratios for the risk factor defined by different levels of the kth tumor marker when each level of the other tumor markers is held constant. In the case of two binary tumor markers and a single binary risk factor, γ_{11} and γ_{12} are obtained (Table 2.1). Then this question can be addressed with tests of the hypotheses $H_{0_{\gamma_{11}}} : \gamma_{11} = 0$ and $H_{0_{\gamma_{12}}} : \gamma_{12} = 0$.

This chapter addresses how each of the four methods under consideration can be constructed to address these two epidemiologic questions, and compares the statistical properties of the methods. Throughout, it is important to keep in mind the original purpose of each of the four methods. Polytomous logistic regression is constructed in such a way

Table 2.1: Interpretation of model parameters

Does the risk factor effect differ with respect to subtypes?	
Parameter	Interpretation
β_{11}	log odds ratio for subtype $m = 1$ vs controls
β_{12}	log odds ratio for subtype $m = 2$ vs controls
β_{13}	log odds ratio for subtype $m = 3$ vs controls
β_{14}	log odds ratio for subtype $m = 4$ vs controls
Does the risk factor effect differ with respect to tumor markers?	
Parameter	Interpretation
γ_{11}	average of differences in log odds ratios when tumor marker $k = 1$ is + vs - and $k = 2$ is fixed
γ_{12}	average of differences in log odds ratios when tumor marker $k = 2$ is + vs - and $k = 1$ is fixed

as to naturally address the question of whether risk factor effects differ across subtypes of disease. The β_{pm} parameters are estimated directly in polytomous logistic regression. Section 2.2.1 shows that the γ_{pk} parameters can then be obtained indirectly as a linear combination of the estimated β_{pm} parameters. Conversely, the two-stage regression with simultaneous estimation approach of Chatterjee (2004) and the stratified logistic regression approach of Rosner *et al.* (2013) were originally proposed to address the question of whether risk factor effects differ across levels of each individual tumor marker. As such, the γ_{pk} parameters are estimated directly. Both methods also allow for inclusion of interaction effects between individual tumor markers. Sections 2.2.3 and 2.2.4 show that when all first-order interaction terms are included in the model, the β_{pm} parameters can be obtained indirectly as a linear combination of the estimated γ_{pk} parameters. The two-stage meta-regression approach of Wang *et al.* (2015) was specifically proposed to address both the question of whether risk factor effects differ across subtypes of disease and the question of whether risk factor effects differ across levels of each individual tumor marker. In this approach the β_{pm}

parameters are directly estimated in the first-stage model and then the γ_{pk} parameters are directly estimated in the second-stage model. Details of model specification and estimation for each of the four methods follow in Section 2.2.

2.2 Methods

This section presents details of the estimation of model parameters and the hypothesis testing procedure for each approach.

2.2.1 Polytomous logistic regression

Polytomous logistic regression allows for the simultaneous estimation of subtype-specific regression parameters. Let Y_i denote the disease status for subject i such that $Y_i = 0$ for a non-diseased control subject and $Y_i = m$ for a subject with disease subtype m. X_{1i} denotes the value of risk factor $p = 1$ for subject i. Then a polytomous logistic regression model is specified as

$$\Pr(Y_i = m|X_{1i}) = \frac{\exp(\beta_{0m} + \beta_{1m}X_{1i})}{1 + \sum_{m=1}^{4} \exp(\beta_{0m} + \beta_{1m}X_{1i})}, m = 1, 2, 3, 4 \tag{2.1}$$

where β_{0m} is the intercept parameter for the mth disease subtype. To evaluate whether the risk factor has the same effect across all subtypes of disease a Wald test of the hypothesis $H_{0_\beta} : \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$ is performed.

Defining w_{km} as the level of the kth tumor marker corresponding to the mth disease subtype, a linearly transformed set of parameters can be created using

$$\beta_{jm} = \gamma_{j0} + \gamma_{j1}w_{1m} + \gamma_{j2}w_{2m} + \gamma_{j12}w_{1m}w_{2m} \tag{2.2}$$

for $j = 0, 1$. It follows that estimates of the γ_{pk} parameters associated with the individual tumor marker effects in the case of $m = 4$ disease subtypes can be obtained as

$$\gamma_{11} = \frac{(\beta_{12} - \beta_{11}) + (\beta_{14} - \beta_{13})}{2} \text{ and } \gamma_{12} = \frac{(\beta_{13} - \beta_{11}) + (\beta_{14} - \beta_{12})}{2}.$$

Note that while here the case is limited to $m = 4$ disease subtypes formed by $k = 2$ tumor markers, this transformation is generalizable. Thus tests addressing the second set of questions, whether risk factor effects differ across levels of each individual tumor marker, can be accomplished using Wald tests of $H_{0_{\gamma_{11}}} : \beta_{12} - \beta_{11} + \beta_{14} - \beta_{13} = 0$ and $H_{0_{\gamma_{12}}} : \beta_{13} - \beta_{11} + \beta_{14} - \beta_{12} = 0$.

It is of interest to note that when data are not available on control subjects, the test for etiologic heterogeneity can be obtained using a case-only polytomous logistic regression model (Begg and Zhang, 1994). In the polytomous logistic regression model one of the four subtypes must be selected to serve as the reference group. Because there is no data on controls, the main effects of the individual tumor markers cannot be determined, so case-only polytomous logistic regression cannot test whether the effect of a risk factor differs across levels of each individual tumor marker. As this method produces almost identical results to those from polytomous logistic regression with regard to the question of whether a risk factor of interest has the same effect across all subtypes of disease, it will not be investigated in further detail.

2.2.2 Two-stage meta-regression

The method of Wang *et al.* (2015) is a two-stage approach. As noted in Section 2.1, this method was specifically proposed to address both the question of whether risk factor effects differ across disease subtypes and the question of whether risk factor effects differ across levels of each individual tumor marker. The first stage of the analysis uses the previously introduced polytomous logistic regression model (Equation 2.1). Thus the test of whether the risk factor has the same effect across all four subtypes of disease is identical to the one used in Section 2.2.1 above. A second-stage analysis is then employed to directly estimate the parameters γ_{10}, γ_{11}, and γ_{12} for risk factor $p = 1$ using a weighted linear regression model,

$$\hat{\beta}_{1m} = \gamma_{10} + \gamma_{11} w_{1m} + \gamma_{12} w_{2m} + e_{1m}, \tag{2.3}$$

where $\hat{\beta}_{1m}$ is the estimated log odds ratio of subtype m versus controls for risk factor $p = 1$ from the polytomous logistic regression model and e_{1m} is within study sampling error such that $Var(e_{1m}) = \widehat{Var}(\hat{\beta}_{1m})$. Wald tests of the hypotheses $H_{0_{\gamma_{11}}} : \gamma_{11} = 0$ and $H_{0_{\gamma_{12}}} : \gamma_{12} = 0$ are used to test whether the risk factor effect differs across levels of each individual tumor marker.

Wang *et al.* (2015) also propose that the second stage model in Equation 2.3 can be extended to include a random effect, which could capture variance between subtypes not explained by the included tumor markers. Alternatively, the second stage model in Equation 2.3 can incorporate interaction terms between the individual tumor markers in order to evaluate whether the effect of the risk factor associated with one tumor marker actually depends on the level of another tumor marker. These alternative second-stage model specifications are not examined in depth, but may in fact prove more appropriate in certain study settings.

2.2.3 Two-stage regression with simultaneous estimation

The method of Chatterjee (2004) is also a two-stage approach with a similar model structure. However, unlike the preceding two-stage meta-regression method (Wang *et al.*, 2015), this approach specifies a joint likelihood and uses a maximum likelihood estimation procedure to simultaneously estimate the first-stage and second-stage regression parameters. When the total number of disease subtypes is moderate, maximum likelihood estimation of the two-stage model is relatively straightforward, though a pseudo-conditional likelihood estimation method is also proposed for the case when the number of disease subtypes is large (Chatterjee, 2004). This method was proposed in order to address the question of whether risk factor effects differ across levels of each individual tumor marker. The first-stage model is the polytomous logistic regression model defined in Equation 2.1. The second-stage model to address the question of whether the effect of risk factor $p = 1$ differs across levels of each individual tumor marker can be constructed as,

$$\beta_{1m} = \gamma_{10} + \gamma_{11} w_{1m} + \gamma_{12} w_{2m}. \tag{2.4}$$

In this framework it is of interest to test the independent effect of each tumor marker when all other tumor markers are held constant. Score tests of the hypotheses $H_{0_{\gamma_{11}}} : \gamma_{11} = 0$ and $H_{0_{\gamma_{12}}} : \gamma_{12} = 0$ can test whether the risk factor effect differs across levels of each individual tumor marker.

However, this model also allows for inclusion of interaction effects between individual tumor markers. If all interaction effects are incorporated then

$$\beta_{1m} = \gamma_{10} + \gamma_{11} w_{1m} + \gamma_{12} w_{2m} + \gamma_{112} w_{1m} w_{2m}, \tag{2.5}$$

would be utilized, where $\gamma_{pk_1k_2}$ is a measure of the interaction effect between the k_1th and k_2th tumor markers with respect to the pth risk factor. Note that this is equivalent to Equation 2.2 for the case of $j = 1$. Analogous to Section 2.2.1, in this setting the β_{pm} parameter estimates can be obtained based on linear combinations of the γ_{pk} parameter estimates using the fact that $\beta_{11} = \gamma_{10}$, $\beta_{12} = \gamma_{10}+\gamma_{11}$, $\beta_{13} = \gamma_{10}+\gamma_{12}$, and $\beta_{14} = \gamma_{10}+\gamma_{11}+\gamma_{12}+\gamma_{112}$. Thus a test of whether the risk factor has the same effect across all four disease subtypes can be conducted by a Wald test of the hypothesis $H_{0_\beta} : \gamma_{11} = \gamma_{12} = \gamma_{112} = 0$. One could also test the hypothesis $H_{0_{\gamma_{pk_1k_2}}} : \gamma_{112} = 0$ in order to determine whether the effect of risk factor $p = 1$ associated with tumor marker $k = 1$ actually depends on the level of tumor marker $k = 2$.

2.2.4 Stratified logistic regression

As an alternative to a two-stage approach, Rosner *et al.* (2013) proposed a single-stage regression method. This method was originally designed to address the question of whether risk factor effects differ across levels of each individual tumor marker using a computational structure for which software is readily available. Let Z_{mi} indicate the disease status for subject i specific to subtype m disease such that

$$Z_{mi} = \begin{cases} 1 & \text{if } Y_i = m \\ 0 & \text{if } Y_i \neq m, \end{cases}$$

for $m = 1, \ldots, M$. In control subjects $Z_{mi} = 0$ for all m. In contrast to all previously discussed methods, here a data augmentation approach is used, such that each case contributes m correlated outcomes, one for each combination of tumor markers, i.e. each disease subtype m (Rosner *et al.*, 2013). This approach was originally designed for use in the setting of cohort studies and was implemented using a Cox regression model stratified by the disease subtype. However, using the fact that a stratified Cox regression model is equivalent to a stratified logistic regression model (Gail *et al.*, 1981), also known as a conditional logistic regression model, the method can easily be applied in the setting of a case-control study when time is constant for all included subjects and data are structured as described. The same data augmentation approach is used, and the logistic regression model is stratified by disease subtype.

To address the question of whether a risk factor of interest, X_{1i}, has the same effect across levels of each individual tumor marker, the stratified logistic regression model can be specified as

$$\Pr(Z_{mi} = 1 | X_{1i}, \mathbf{w}_m) = \frac{\exp\left(\alpha_m + \gamma_{10} X_{1i} + \gamma_{11} X_{1i} w_{1m} + \gamma_{12} X_{1i} w_{2m}\right)}{1 + \exp\left(\alpha_m + \gamma_{10} X_{1i} + \gamma_{11} X_{1i} w_{1m} + \gamma_{12} X_{1i} w_{2m}\right)}, \tag{2.6}$$

where $\mathbf{w}_m = \{w_{1m}, \ldots, w_{km}\}$ is the vector of tumor markers for the mth subtype and α_m is the stratum-specific intercept term, which cancels out in the conditional likelihood (Breslow and Day, 1980). This model can be used to test whether the risk factor effect differs across levels of each individual tumor marker using Wald tests of the hypotheses $H_{0_{\gamma_{11}}} : \gamma_{11} = 0$ and $H_{0_{\gamma_{12}}} : \gamma_{12} = 0$.

The stratified logistic regression approach of Rosner *et al.* (2013) also allows for inclusion of interaction effects between individual tumor markers. When all interaction effects are included, the model

$$\Pr(Z_{mi} = 1 | X_{1i}, \mathbf{w}_m) = \frac{\exp\left(\alpha_m + \gamma_{10} X_{1i} + \gamma_{11} X_{1i} w_{1m} + \gamma_{12} X_{1i} w_{2m} + \gamma_{112} X_{1i} w_{1m} w_{2m}\right)}{1 + \exp\left(\alpha_m + \gamma_{10} X_{1i} + \gamma_{11} X_{1i} w_{1m} + \gamma_{12} X_{1i} w_{2m} + \gamma_{112} X_{1i} w_{1m} w_{2m}\right)} \tag{2.7}$$

is obtained, and as in Section 2.2.3 the β_{pm} parameters can be obtained indirectly as a linear combination of the γ_{pk} parameters to test whether the risk factor has the same effect across all four disease subtypes.

2.2.5 Software

All statistical analyses were conducted using R software (R Core Team, 2018). For polytomous logistic regression, the `multinom` function from the `nnet` package (Venables and Ripley, 2002) was used for estimation and the `wald.test` function from the `aod` package (Lesnoff *et al.*, 2012) was used for significance testing. For the second-stage model in the two-stage meta-regression method of Wang *et al.* (2015), the `rma.mv` function from the `metafor` package (Viechtbauer, 2010) was used for estimation and significance testing. Estimation and significance testing for the two-stage regression with simultaneous estimation method of Chatterjee (2004) was conducted using an R function provided by the authors, except in the case of the test of H_{0_β}, which was conducted using the `wald.test` function from the `aod` package (Lesnoff *et al.*, 2012). Finally, following data augmentation, the `clogit` function from the `survival` package (Therneau, 2015; Terry M. Therneau and Patricia M. Grambsch, 2000) was used for estimation and testing for the stratified logistic regression method of Rosner *et al.* (2013). For all methods besides that of Chatterjee (2004), other standard software packages that support the underlying statistical models could be used. However all require transformation of the results from one parametric configuration to another and use the parameter estimates and variance-covariance matrices for hypothesis testing.

2.3 Data example

Data from a previous study that combined two large breast cancer case-control studies, the Cancer and Steroid Hormone (CASH) study and the Womens' Contraceptive and Reproductive Experiences (CARE) study, leading to a total of 984 cases and a corresponding 1592 controls, are used to illustrate the methods (Begg *et al.*, 2013). The goal in this section is not to conduct a detailed analysis of etiologic heterogeneity in breast cancer. Instead, the focus is on a simplified strategy that addresses the etiologic heterogeneity of breast cancer classified into subtypes described by estrogen receptor (ER) and progesterone receptor (PR) status from the perspective of a single risk factor, oral contraceptive (OC) use. The purpose is simply to contrast the various modeling strategies.

The primary results from the four methods are presented in Table 2.2. The top portion of the table contains results relevant to the question of whether OC use has the same effect across the four disease subtypes. This question is addressed with Equation 2.1 for polytomous logistic regression and the method of Wang *et al.* (2015), Equation 2.5 for the method of Chatterjee (2004), and Equation 2.7 for the method of Rosner *et al.* (2013). All methods lead to rejection of the null hypothesis (all p-values < 0.05), so regardless of the method the conclusion is that the effect of OC use differs across the four disease subtypes. Of note the parameter estimates for polytomous logistic regression, the method of Wang *et al.* (2015), and the method of Chatterjee (2004) are practically identical. This is expected as the first-stage model for the method of Wang *et al.* (2015) is simply the polytomous logistic regression model, and when all first order interaction effects are included in the method of Chatterjee (2004) and maximum likelihood estimation is used, this model should produce results that are nearly identical to those from the polytmous logistic regression model. Finally, note that there are some small differences between the parameter estimates from the method of Rosner *et al.* (2013) as compared to the other three methods, in that the parameter estimates are all less positive in magnitude.

The lower portion of Table 2.2 displays results related to the questions of whether

Table 2.2: Results of data example comparing existing methods

Does the risk factor effect differ with respect to subtypes?

Method	Subtype	Parameter	Estimate	p-value
Polytomous[1]	ER-/PR-	β_{11}	0.31	0.042
	ER+/PR-	β_{12}	-0.11	
	ER-/PR+	β_{13}	0.22	
	ER+/PR+	β_{14}	0.03	
Wang[2]	ER-/PR-	β_{11}	0.31	0.042
	ER+/PR-	β_{12}	-0.11	
	ER-/PR+	β_{13}	0.22	
	ER+/PR+	β_{14}	0.03	
Chatterjee[3]	ER-/PR-	β_{11}	0.31	0.042
	ER+/PR-	β_{12}	-0.11	
	ER-/PR+	β_{13}	0.21	
	ER+/PR+	β_{14}	0.03	
Rosner[4]	ER-/PR-	β_{11}	0.29	0.029
	ER+/PR-	β_{12}	-0.15	
	ER-/PR+	β_{13}	0.18	
	ER+/PR+	β_{14}	0.00	

Does the risk factor effect differ with respect to tumor markers?

Method	Tumor marker	Parameter	Estimate	p-value
Polytomous[1]	ER	γ_{11}	-0.30	0.046
	PR	γ_{12}	0.02	0.887
Wang[2]	ER	γ_{11}	-0.33	0.031
	PR	γ_{12}	0.05	0.731
Chatterjee[3]	ER	γ_{11}	-0.33	0.028
	PR	γ_{12}	0.05	0.719
Rosner[4]	ER	γ_{11}	-0.34	0.024
	PR	γ_{12}	0.05	0.739

[1] Polytomous logistic regression

[2] Two-stage meta-regression (Wang *et al.*, 2015)

[3] Two-stage regression with simultaneous estimation (Chatterjee, 2004)

[4] Stratified logistic regression (Rosner *et al.*, 2013)

the effect of OC use differs across levels of ER status when PR status is held constant, and whether the effect of OC use differs across levels of PR status when ER status is held constant. These questions are addressed with Equation 2.1 for polytomous logistic regression, Equation 2.3 for the method of Wang *et al.* (2015), Equation 2.4 for the method of Chatterjee (2004), and Equation 2.6 for the method of Rosner *et al.* (2013). Again, the parameter estimates and p-values are similar. Regardless of the method, the conclusion is that the effect of OC use on breast cancer risk differs by ER status, but the effect of OC use on breast cancer risk does not differ by PR status.

2.4 Simulation study

The simulation study is conducted using a similar framework to the data example, with four disease subtypes formed by cross-classification of two tumor markers as described in Section 2.1. There is a single binary risk factor of interest, with a prevalence among control subjects of $q = 0.3$. Each simulation uses 1000 controls and 1000 cases, with the cases divided equally among the four disease subtypes. The true regression coefficients are fixed at β_{1m} for subtype m disease, $m = 1, 2, 3, 4$. Risk factor data are randomly generated for each subject with disease subtype m from a binomial distribution with probability $\exp(q\beta_{1m})/[1+\exp(q\beta_{1m})]$ and for each control subject with probability $\exp(q)/[1+\exp(q)]$. For each simulation setting, 1000 simulated data sets are generated.

To address the question of whether the risk factor effect differs across the disease subtypes, the simulation study employs Equation 2.1 for polytomous logistic regression and the method of Wang *et al.* (2015), Equation 2.5 for the method of Chatterjee (2004), and Equation 2.7 for the method of Rosner *et al.* (2013). To address the question of whether the risk factor effect differs across levels of each individual tumor marker, the simulation study uses Equation 2.1 for polytomous logistic regression, Equation 2.3 for the method of Wang *et al.* (2015), Equation 2.4 for the method of Chatterjee (2004), and Equation 2.6 for the method of Rosner *et al.* (2013).

It is important to note that some of the simulation settings imply an interaction effect between the individual tumor markers whereas some of the simulation settings imply a main effects model with no interaction effect. When there is no interaction between the individual tumor markers, i.e. when $\gamma_{112} = 0$, then $\beta_{14} = \beta_{12} + \beta_{13} - \beta_{11}$ and a test of whether the risk factor effect differs across the disease subtypes can be conducted with a test of $H_{0_\beta} : \gamma_{11} = \gamma_{12} = 0$. In settings where there is truly no interaction effect, the method of Chatterjee (2004) is explored using both Equation 2.5 and Equation 2.4 to determine whether there is an efficiency gain from using a model that does not incorporate an interaction effect as compared to a model that does.

2.4.1 Data simulated under the null hypothesis

The first set of simulations is conducted under the null hypothesis for the question of whether the risk factor effect differs across the four disease subtypes, and under the null hypothesis for the question of whether the risk factor effect differs across levels of each individual tumor marker. Set $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0.1$ and therefore $\gamma_{11} = \gamma_{12} = 0$. Equivalent disease subtype effects such as this imply no interaction effect between the individual tumor markers. For the question of whether the risk factor effect differs across the four disease subtypes, the size of the test is 0.051 for polytomous logistic regression, the method of Wang *et al.* (2015), and the method of Chatterjee (2004) whereas the method of Rosner *et al.* (2013) has an inflated type I error of 0.089 (Table 2.3, upper portion). The biases in parameter estimates are small for all methods except that of Rosner *et al.* (2013). When the main effects model of Chatterjee (2004) is applied using Equation 2.4, similarly small biases of -0.000, -0.004, -0.001, and -0.006 are found for β_{11}, β_{12}, β_{13}, and β_{14}, respectively, but there is a slightly inflated type I error of 0.068.

For the question of whether the risk factor effect differs across levels of each individual tumor marker, polytomous logistic regression and the method of Chatterjee (2004) have very similar type I errors for γ_{11} of 0.063 and for γ_{12} of 0.051 and 0.050, respectively (Table 2.3, lower portion). The method of Wang *et al.* (2015) has lower type I errors, 0.037 and 0.031 for testing γ_{11} and γ_{12}, respectively; conversely, the method of Rosner *et al.* (2013) has inflated type I errors of 0.077 and 0.073. In this setting all methods produce parameter estimates with comparably small biases.

2.4.2 Data simulated under the alternative hypothesis

The second set of simulations is conducted under the alternative hypothesis for the question of whether the risk factor effect differs across the four disease subtypes, and under the alternative hypothesis for the question of whether the risk factor effect differs across levels of each individual tumor marker. Here let $\beta_{11} = 0.2$, $\beta_{12} = \beta_{13} = 0.3$, and $\beta_{14} = 0.8$ so that $\gamma_{11} = \gamma_{12} = 0.3$. For the question of whether the effect of the risk factor differs across the four subtypes, polytomous logistic regression, the method of Wang *et al.* (2015) and the method of Chatterjee (2004) all have power of 0.822 whereas the method of Rosner *et al.* (2013) has higher power of 0.862 (Table 2.3, upper portion). However, recall that the method of Rosner *et al.* (2013) had higher type I error than the other methods, and so calibration is needed to truly compare the signal detection srengths of the methods. While biases are generally very small for most methods, there is substantial bias in parameter estimates for the method of Rosner *et al.* (2013).

Table 2.3: Results of simulation study comparing existing methods

Does the risk factor effect differ with respect to subtypes?

		Null hypothesis			Alternative hypothesis		
Method	Parameter	Truth	Bias	Type I error	Truth	Bias	Power
Polytomous[1]	β_{11}	0.1	-0.000	0.051	0.2	-0.001	0.822
	β_{12}	0.1	-0.004		0.3	-0.006	
	β_{13}	0.1	-0.001		0.3	-0.002	
	β_{14}	0.1	-0.007		0.8	-0.007	
Wang[2]	β_{11}	0.1	-0.000	0.051	0.2	-0.001	0.822
	β_{12}	0.1	-0.004		0.3	-0.006	
	β_{13}	0.1	-0.001		0.3	-0.002	
	β_{14}	0.1	-0.007		0.8	-0.007	
Chatterjee[3]	β_{11}	0.1	-0.000	0.051	0.2	-0.001	0.822
	β_{12}	0.1	-0.004		0.3	-0.006	
	β_{13}	0.1	-0.001		0.3	-0.002	
	β_{14}	0.1	-0.007		0.8	-0.007	
Rosner[4]	β_{11}	0.1	0.047	0.089	0.2	0.190	0.862
	β_{12}	0.1	0.043		0.3	0.178	
	β_{13}	0.1	0.046		0.3	0.182	
	β_{14}	0.1	0.040		0.8	0.146	

Does the risk factor effect differ with respect to tumor markers?

		Null hypothesis			Alternative hypothesis		
Method	Parameter	Truth	Bias	Type I error	Truth	Bias	Power
Polytomous[1]	γ_{11}	0.0	-0.005	0.063	0.3	-0.005	0.589
	γ_{12}	0.0	-0.001	0.051	0.3	-0.001	0.599
Wang[2]	γ_{11}	0.0	-0.005	0.037	0.3	0.005	0.483
	γ_{12}	0.0	-0.001	0.031	0.3	0.010	0.475
Chatterjee[3]	γ_{11}	0.0	-0.005	0.063	0.3	0.006	0.560
	γ_{12}	0.0	-0.001	0.050	0.3	0.010	0.572
Rosner[4]	γ_{11}	0.0	-0.005	0.077	0.3	-0.012	0.605
	γ_{12}	0.0	-0.001	0.073	0.3	-0.008	0.624

[1] Polytomous logistic regression

[2] Two-stage meta-regression (Wang *et al.*, 2015)

[3] Two-stage regression with simultaneous estimation (Chatterjee, 2004)

[4] Stratified logistic regression (Rosner *et al.*, 2013)

For the question of whether the risk factor effect differs across levels of each individual tumor marker, polytomous logistic regression and the method of Chatterjee (2004) again have similar power (Table 2.3, lower portion). The method of Wang *et al.* (2015) has lower power whereas the method of Rosner *et al.* (2013) has slightly higher power. However, again recall that the method of Rosner *et al.* (2013) had inflated type I error. All methods produce parameter estimates with small biases.

The following was done in order to compare the power of the methods in a calibrated manner. First the effect size was varied by fixing $\beta_{11} = 0.2$ and $\beta_{12} = \beta_{13} = 0.3$, and incrementally increasing β_{14} from 0.3 to 0.9. This allowed for determination of how large the subtype four effect size, β_{14}, needs to be in order to achieve various levels of power to address whether the risk factor effect differs across the four disease subtypes. The comparison was calibrated by ranking the simulated p-values under the null hypothesis and choosing the critical value that ensured the test size was exactly 0.05, then this critical value was used to determine power. Figure 2.1A shows the resulting power curves for the different methods. After calibration of type I error, the four methods have indistinguishable power. Note that one of these cases, when $\beta_{14} = 0.4$, implies no interaction effect between the individual tumor markers. Whereas the calibrated power using Chatterjee's Equation 2.5 results in a power of 0.122 in this setting, Equation 2.4 results in slightly lower calibrated power of 0.119.

The power to address whether the risk factor effect differs across levels of each individual tumor marker is similarly compared. Figure 2.1B shows the power to detect an effect for γ_{1k}. The results are similar across the four methods.

Figure 2.1: Log odds ratio required to achieve various levels of power when type I error is calibrated to $\alpha = 0.05$ for (A) β_{14} to address whether risk factor effects differ across subtypes, and (B) γ_{1k} to address whether risk factor effects differ across levels of each individual tumor marker.

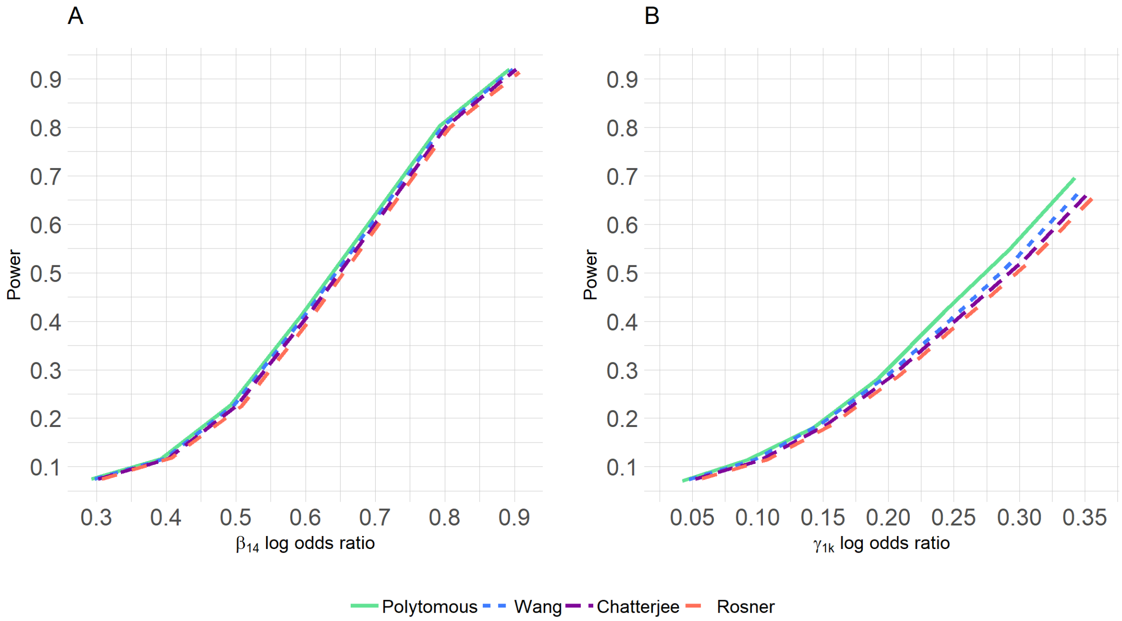


Table 2.4: Type I error for different risk factor prevalences q and true effect sizes β_{1m} with $M = 4$ disease subtypes formed by $K = 2$ individual tumor markers

q	0.3					0.6				
β_{1m}	0	0.1	0.2	0.3	0.4	0	0.1	0.2	0.3	0.4
Does the risk factor effect differ with respect to subtypes?										
Polytomous[1]	0.052	0.051	0.060	0.055	0.054	0.055	0.054	0.047	0.051	0.046
Wang[2]	0.052	0.051	0.060	0.055	0.054	0.055	0.054	0.047	0.051	0.046
Chatterjee[3]	0.052	0.051	0.060	0.055	0.054	0.055	0.054	0.047	0.051	0.046
Rosner[4]	0.093	0.089	0.077	0.086	0.082	0.088	0.084	0.076	0.072	0.076
Does the risk factor effect differ with respect to tumor marker 1 (γ_{11})?										
Polytomous[1]	0.064	0.063	0.056	0.057	0.058	0.057	0.058	0.051	0.042	0.045
Wang[2]	0.035	0.037	0.037	0.033	0.033	0.033	0.032	0.025	0.022	0.028
Chatterjee[3]	0.064	0.063	0.056	0.057	0.058	0.057	0.058	0.051	0.043	0.047
Rosner[4]	0.079	0.077	0.068	0.079	0.074	0.079	0.074	0.069	0.063	0.060
Does the risk factor effect differ with respect to tumor marker 2 (γ_{12})?										
Polytomous[1]	0.057	0.051	0.055	0.058	0.049	0.059	0.049	0.048	0.044	0.052
Wang[2]	0.031	0.031	0.030	0.030	0.032	0.030	0.030	0.027	0.031	0.034
Chatterjee[3]	0.057	0.050	0.055	0.058	0.046	0.058	0.046	0.047	0.046	0.052
Rosner[4]	0.073	0.073	0.067	0.070	0.062	0.071	0.063	0.057	0.062	0.063

[1] Polytomous logistic regression

[2] Two-stage meta-regression (Wang *et al.*, 2015)

[3] Two-stage regression with simultaneous estimation (Chatterjee, 2004)

[4] Stratified logistic regression (Rosner *et al.*, 2013)

2.4.3 Data simulated under different configurations

Sensitivity analyses are conducted in order to further elucidate the statistical properties of the four methods for the study of etiologic heterogeneity.

First, the sensitivity of the results in Sections 2.4.1 and 2.4.2 to the prevalence of the risk factor are explored. Additional simulations were conducted for the case of four disease subtypes and a single binary risk factor, with data generated as described at the beginning of Section 2.4 using 1000 controls and 1000 cases. For each setting, 1000 simulated data sets are generated. Here settings where the risk factor prevalence is $q = 0.3$ or $q = 0.6$ are separately investigated. Data are first generated under the null hypothesis, and the true common regression coefficients ($\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$) are each fixed at $0, 0.1, 0.2, 0.3$, and 0.4. Results are presented in Table 2.4, with a similar pattern of results to the null case presented in Table 2.3, which corresponds to risk factor prevalence $q = 0.3$ and true common regression coefficients fixed at 0.1. Polytomous logistic regression and the methods of Chatterjee (2004) and Wang *et al.* (2015) perform similarly with respect to type I error for the test of H_{0_β} whereas the method of Rosner *et al.* (2013) is anti-conservative. Polytomous logistic regression and the method of Chatterjee (2004) perform similarly with respect to type I error for the tests of $H_{0_{\gamma_{11}}}$ and $H_{0_{\gamma_{12}}}$ whereas the method of Wang *et al.* (2015) is conservative and the method of Rosner *et al.* (2013) is again anti-conservative. Data are next generated under the alternative hypothesis. For each of the two risk factor prevalences, three alternative scenarios are investigated, with true values for $\{\beta_{11}, \beta_{12}, \beta_{13}\}$ fixed at $\{0.2, 0.25, 0.25\}$, $\{0.2, 0.3, 0.3\}$, and $\{0.2, 0.4, 0.4\}$ and values of β_{14} ranging from 0.25 to 0.85, 0.3 to 0.9 and 0.4 to 1.0, respectively. Power was calibrated for all results as described in Section 2.4.2. Results are presented in Figure 2.2 for β_{14} and Figure 2.3 for γ_{1k}. In all configurations of parametric values and risk factor prevalences, the pattern of results is in line with those presented in Figure 1, such that all methods have similar power after calibration for differences in type I error.

Next, the methods other than polymotomous logistic regression were created to accommodate multiple tumor factors, and thus have the capacity to take advantage of dimension reduction. A limited exploration of the expansion of the number of tumor markers to $K = 4$ was conducted, whereby there are $M = 16$ subtypes that must be evaluated separately in the logistic regression model. Data are generated for sixteen disease subtypes formed by cross-classification of four binary tumor markers as described at the start of Section 2.4. Again there is a single binary risk factor, and the settings where the risk factor prevalence

Figure 2.2: Log odds ratio required to achieve various levels of power when type I error is calibrated to $\alpha = 0.05$ to address whether risk factor effects differ across $M = 4$ disease subtypes

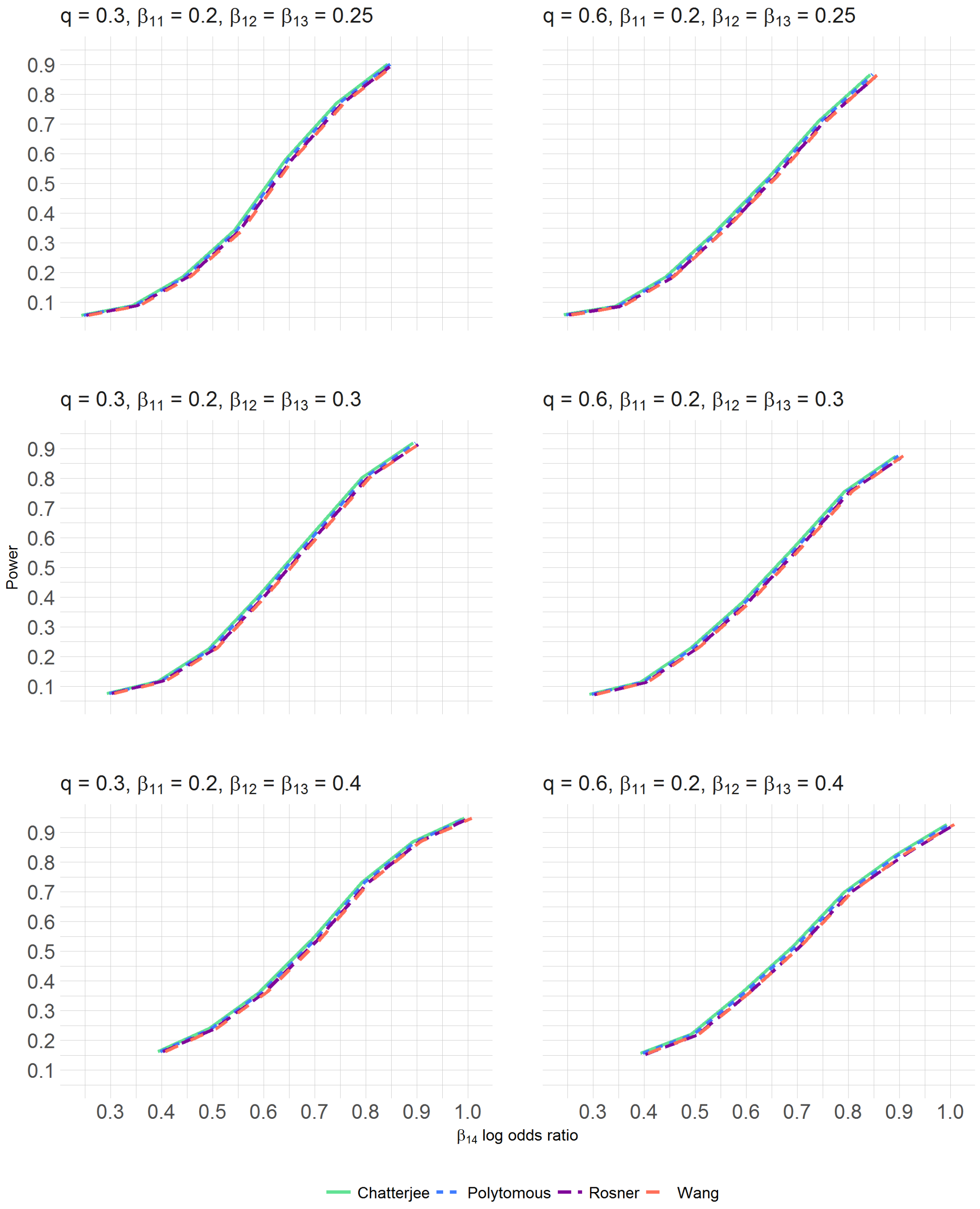
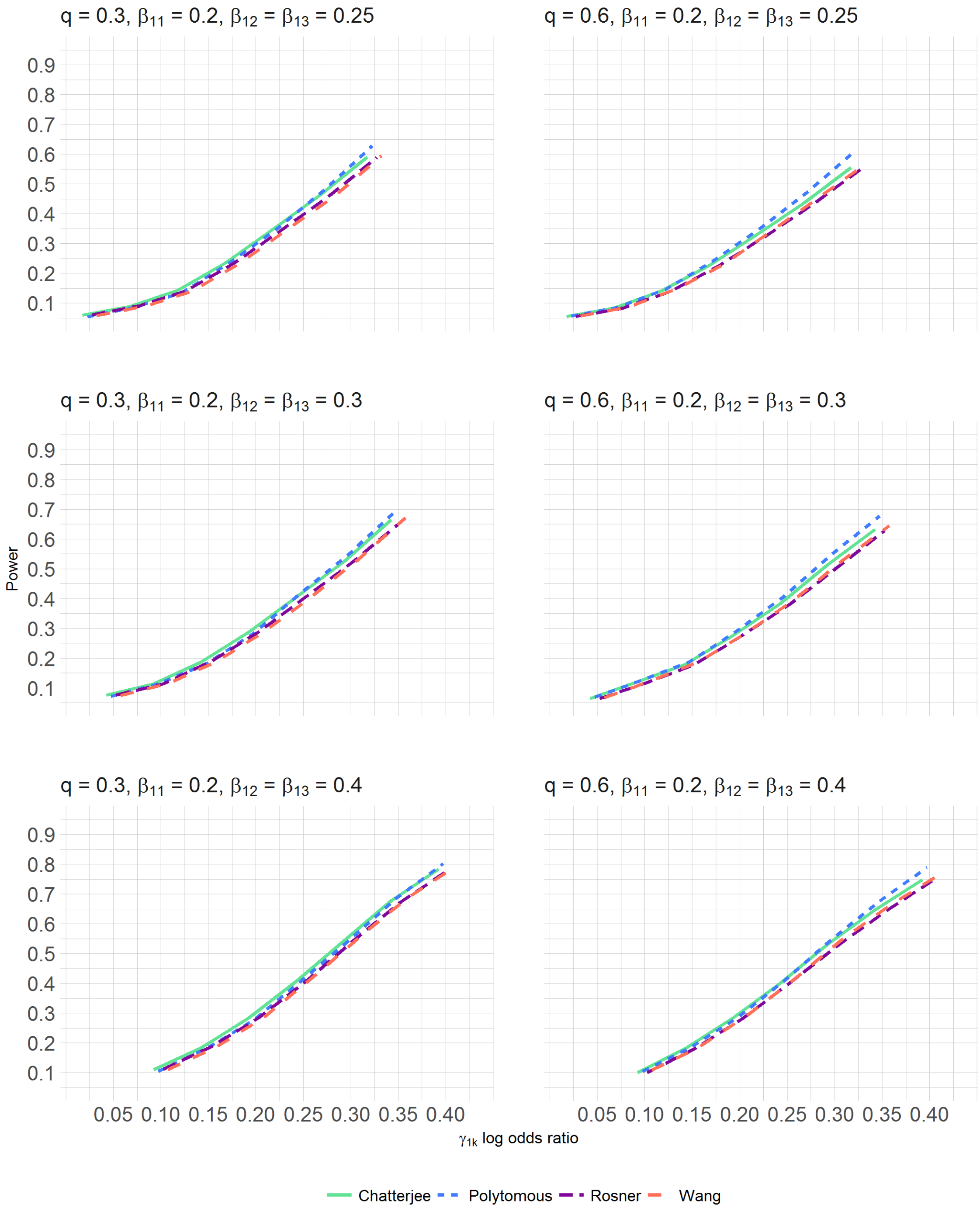


Figure 2.3: Log odds ratio required to achieve various levels of power when type I error is calibrated to $\alpha = 0.05$ to address whether risk factor effects differ across each of the $K = 2$ individual tumor markers that form $M = 4$ disease subtypes



is $q = 0.3$ or $q = 0.6$ are separately investigated. For each simulation setting, 500 simulated data sets were generated using 1008 controls and 1008 cases to allow for equal subdivision of cases into $M = 16$ subtypes. Data are first generated under the null hypothesis, and the true common regression coefficients ($\beta_{11} = \beta_{12} = \cdots = \beta_{1(16)}$) are each fixed at $0.05, 0.1, 0.15$, and 0.2. Results are presented in Table 2.5. A similar pattern of results as in the case of $M = 4$ subtypes was seen. Data are next generated under the alternative hypothesis. For each of the two risk factor prevalences, three alternative scenarios are investigated, with true values $\beta_{1m} = \{0.2, 0.4, 0.4, 0.4, 0.4, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.8, 0.8, 0.8, 0.8, 1.0\}$, $\beta_{1m} = \{0.2, 0.4, 0.4, 0.4, 0.4, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 1.2, 1.2, 1.2, 1.2, 1.2\}$, and $\beta_{1m} = \{0.2, 0.4, 0.4, 0.4, 0.4, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 1.2, 1.2, 1.2, 1.2, 1.4\}$. Results are presented in Table 2.6. Even as the number of subtypes increases to sixteen, a similar pattern of results is seen as in the setting of four disease subtypes.

Finally, to investigate the setting where more than one risk factor is included, a data example is conducted using the same data and subtypes described in Section 2.3 and incorporating a variety of continuous and binary risk factors of relevance to breast cancer risk (Begg *et al.*, 2013). Results are presented in Tables 2.7 and 2.8. The interpretation of each risk factor must now be made in the context of adjustment for all other risk factors. Across all risk factors, for the question of whether risk factor effects differ across disease subtypes, polytomous logistic regression and the methods of Chatterjee (2004) and Wang *et al.* (2015) result in similar parameter estimates and p-values whereas results from the method of Rosner *et al.* (2013) differ slightly from the other methods. It is of interest to note that in the context of a multivariable data analysis, the effect of oral contraceptive use is no longer significantly different across disease subtypes (Table 2.7) whereas in the simplified data example in Table 2.2, where only oral contraceptive use was included in the model,

Table 2.5: Type I error for different risk factor prevalences q and true effect sizes β_{1m} with $M = 16$ disease subtypes formed by $K = 4$ individual tumor markers

q	0.3				0.6			
β_{1m}	0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2
Does the risk factor effect differ with respect to subtypes?								
Polytomous[1]	0.052	0.050	0.038	0.034	0.036	0.028	0.030	0.044
Wang[2]	0.052	0.050	0.038	0.034	0.036	0.028	0.030	0.044
Chatterjee[2]	0.052	0.050	0.038	0.034	0.036	0.028	0.030	0.044
Rosner[4]	0.060	0.058	0.050	0.046	0.048	0.036	0.046	0.052
Does the risk factor effect differ with respect to tumor marker 1 (γ_{11})?								
Polytomous[1]	0.060	0.054	0.040	0.032	0.050	0.054	0.040	0.042
Wang[2]	0.050	0.042	0.032	0.032	0.042	0.032	0.028	0.034
Chatterjee[3]	0.058	0.052	0.040	0.032	0.048	0.050	0.042	0.040
Rosner[4]	0.060	0.058	0.046	0.034	0.054	0.056	0.042	0.040
Does the risk factor effect differ with respect to tumor marker 2 (γ_{12})?								
Polytomous[1]	0.044	0.052	0.050	0.042	0.050	0.058	0.050	0.058
Wang[2]	0.036	0.042	0.040	0.032	0.036	0.032	0.038	0.038
Chatterjee[3]	0.044	0.050	0.048	0.040	0.046	0.054	0.048	0.054
Rosner[4]	0.044	0.052	0.050	0.040	0.052	0.060	0.048	0.056
Does the risk factor effect differ with respect to tumor marker 3 (γ_{13})?								
Polytomous[1]	0.058	0.042	0.058	0.048	0.046	0.048	0.044	0.042
Wang[2]	0.044	0.040	0.042	0.038	0.038	0.036	0.032	0.030
Chatterjee[3]	0.058	0.042	0.056	0.044	0.042	0.048	0.042	0.038
Rosner[4]	0.058	0.044	0.060	0.048	0.048	0.054	0.046	0.044
Does the risk factor effect differ with respect to tumor marker 4 (γ_{14})?								
Polytomous[1]	0.058	0.042	0.058	0.048	0.046	0.048	0.044	0.042
Wang[2]	0.044	0.040	0.042	0.038	0.038	0.036	0.032	0.030
Chatterjee[3]	0.058	0.042	0.056	0.044	0.042	0.048	0.042	0.038
Rosner[4]	0.058	0.044	0.060	0.048	0.048	0.054	0.046	0.044

[1] Polytomous logistic regression

[2] Two-stage meta-regression (Wang *et al.*, 2015)

[3] Two-stage regression with simultaneous estimation (Chatterjee, 2004)

[4] Stratified logistic regression (Rosner *et al.*, 2013)

Table 2.6: Power for different risk factor prevalences q and different alternative hypothesis scenarios with $M = 16$ disease subtypes formed by $K = 4$ individual tumor markers

q	0.3			0.6		
Alternative scenario*	1	2	3	1	2	3
Does the risk factor effect differ with respect to subtypes?						
Polytomous[1]	0.332	0.834	0.872	0.260	0.750	0.800
Wang[1]	0.332	0.834	0.872	0.260	0.750	0.800
Chatterjee[1]	0.332	0.834	0.872	0.262	0.748	0.798
Rosner[1]	0.356	0.854	0.880	0.278	0.780	0.816
Does the risk factor effect differ with respect to tumor marker 1 (γ_{11})?						
Polytomous[1]	0.308	0.596	0.644	0.284	0.538	0.590
Wang[2]	0.270	0.568	0.638	0.246	0.500	0.558
Chatterjee[3]	0.296	0.614	0.674	0.274	0.564	0.608
Rosner[4]	0.308	0.620	0.682	0.280	0.574	0.622
Does the risk factor effect differ with respect to tumor marker 2 (γ_{12})?						
Polytomous[1]	0.312	0.592	0.646	0.270	0.552	0.602
Wang[2]	0.280	0.564	0.636	0.232	0.530	0.576
Chatterjee[3]	0.312	0.614	0.678	0.266	0.564	0.610
Rosner[4]	0.326	0.630	0.682	0.270	0.574	0.620
Does the risk factor effect differ with respect to tumor marker 3 (γ_{13})?						
Polytomous[1]	0.298	0.602	0.648	0.300	0.552	0.606
Wang[2]	0.262	0.562	0.620	0.242	0.502	0.554
Chatterjee[3]	0.294	0.606	0.654	0.276	0.568	0.606
Rosner[4]	0.300	0.616	0.666	0.286	0.576	0.618
Does the risk factor effect differ with respect to tumor marker 4 (γ_{14})?						
Polytomous[1]	0.290	0.610	0.660	0.242	0.540	0.590
Wang[2]	0.264	0.574	0.634	0.210	0.506	0.560
Chatterjee[3]	0.288	0.632	0.672	0.238	0.570	0.622
Rosner[4]	0.296	0.638	0.676	0.250	0.586	0.630

*Alternative scenarios:

1: $\beta_{1m} = \{0.2, 0.4, 0.4, 0.4, 0.4, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.8, 0.8, 0.8, 0.8, 1.0\}$

2: $\beta_{1m} = \{0.2, 0.4, 0.4, 0.4, 0.4, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 1.2, 1.2, 1.2, 1.2, 1.2\}$

3: $\beta_{1m} = \{0.2, 0.4, 0.4, 0.4, 0.4, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 1.2, 1.2, 1.2, 1.2, 1.4\}$

[1] Polytomous logistic regression

[2] Two-stage meta-regression (Wang *et al.*, 2015)

[3] Two-stage regression with simultaneous estimation (Chatterjee, 2004)

[4] Stratified logistic regression (Rosner *et al.*, 2013)

the effect was significantly different across disease subtypes according to all subtypes.

2.5 Discussion

This chapter defined two key questions that epidemiologists seek to answer in studies of etiologic heterogeneity and then showed how to address these questions using each of the methods that have been proposed. It demonstrated the distinctions of the methods by creating a unified notation. The simulations show that the stratified logistic regression method of Rosner *et al.* (2013) results in substantial biases in parameter estimation for addressing whether risk factor effects differ across levels of the disease subtype, although it is acknowledged that this was not a stated goal of the method by the authors. Additionally, the method is anti-conservative. All other methods have type I error close to the nominal level. In the simplified setting examined here, whereas the other methods all estimate eight parameters, the method of Rosner *et al.* (2013) conditions out the constant terms and only involves estimation of four parameters. The conditional nature of this model clearly has implications for the validity of parameter estimates and hypothesis tests related to the question of heterogeneity across disease subtypes. For addressing whether risk factor effects differ across levels of each individual tumor marker, polytomous logistic regression and the two-stage regression with simultaneous estimation method of Chatterjee (2004) perform similarly with respect to type I error whereas the two-stage meta-regression method of Wang *et al.* (2015) is overly conservative and the stratified logistic regression method of Rosner *et al.* (2013) is anti-conservative. When differences in type I error are calibrated, all methods achieve similar power.

In this chapter the focus was on subtypes formed by cross-classification of tumor markers, and on the distinct influences of the individual tumor markers. In breast cancer research,

Table 2.7: Full data application to address the question of whether each risk factor differs across levels of subtypes formed by ER and PR status. The model is additionally adjusted for study center.

Risk factor	Method	ER-/	ER+/	ER-/	ER+/	p-value
		PR-	PR-	PR+	PR+	
Age at diagnosis (per 10 years)	Polytomous[1]	-0.03	0.57	0.16	0.53	<.001
	Wang[2]	-0.03	0.57	0.16	0.53	<.001
	Chatterjee[3]	-0.03	0.57	0.16	0.53	<.001
	Rosner[4]	-0.13	0.52	0.13	0.75	<.001
Age at menarche (per 2 years)	Polytomous[1]	0.04	-0.04	-0.00	-0.14	0.065
	Wang[2]	0.04	-0.04	-0.00	-0.14	0.065
	Chatterjee[3]	0.04	-0.04	-0.00	-0.14	0.065
	Rosner[4]	0.06	-0.01	0.04	-0.03	0.041
Nulliparous	Polytomous[1]	-0.08	0.65	0.21	0.33	0.004
	Wang[2]	-0.08	0.65	0.21	0.33	0.004
	Chatterjee[3]	-0.08	0.65	0.21	0.33	0.004
	Rosner[4]	-0.16	0.59	0.16	0.29	0.002
Age at first birth (per 5 years)	Polytomous[1]	0.07	0.01	0.13	0.15	0.343
	Wang[2]	0.07	0.01	0.13	0.15	0.343
	Chatterjee[3]	0.07	0.01	0.13	0.15	0.343
	Rosner[4]	0.02	0.02	0.15	0.14	0.288
Months of breastfeeding (per 6)	Polytomous[1]	-0.11	-0.09	-0.22	-0.08	0.566
	Wang[2]	-0.11	-0.09	-0.22	-0.08	0.566
	Chatterjee[3]	-0.11	-0.09	-0.22	-0.08	0.567
	Rosner[4]	-0.09	-0.04	-0.14	-0.09	0.380
Post-menopausal	Polytomous[1]	-0.23	-0.12	-1.29	-0.75	<.001
	Wang[2]	-0.23	-0.12	-1.29	-0.75	<.001
	Chatterjee[3]	-0.23	-0.12	-1.29	-0.75	<.001
	Rosner[4]	-0.08	0.02	-1.19	-0.64	<.001
Pre-menopausal BMI (per 20)	Polytomous[1]	0.34	-0.06	0.98	-0.34	0.010
	Wang[2]	0.34	-0.06	0.98	-0.34	0.010
	Chatterjee[3]	0.34	-0.06	0.98	-0.34	0.010
	Rosner[4]	0.32	-0.20	0.94	0.12	0.005
Post-menopausal BMI (per 20)	Polytomous[1]	-0.17	-0.79	-0.19	-0.04	0.490
	Wang[2]	-0.17	-0.79	-0.19	-0.04	0.490
	Chatterjee[3]	-0.17	-0.79	-0.19	-0.04	0.489
	Rosner[4]	-0.14	-0.68	0.09	-0.29	0.456
Oral contraceptive use	Polytomous[1]	0.07	0.04	-0.25	-0.08	0.497
	Wang[2]	0.07	0.04	-0.25	-0.08	0.497
	Chatterjee[3]	0.07	0.04	-0.25	-0.08	0.497
	Rosner[4]	0.08	0.06	-0.23	-0.08	0.477
Family history of breast cancer	Polytomous[1]	0.64	0.88	-0.02	0.73	0.226
	Wang[2]	0.64	0.88	-0.02	0.73	0.226
	Chatterjee[3]	0.64	0.88	-0.02	0.73	0.226
	Rosner[4]	0.42	0.62	-0.29	0.53	0.186

[1] Polytomous logistic regression

[2] Two-stage meta-regression (Wang *et al.*, 2015)

[3] Two-stage regression with simultaneous estimation (Chatterjee, 2004)

[4] Stratified logistic regression (Rosner *et al.*, 2013)

Table 2.8: Full data application to address the question of whether each risk factor differs across levels of ER and PR status. The model is additionally adjusted for study center.

		ER		PR	
Risk factor	Method	Estimate	*p*-value	Estimate	*p*-value
Age at diagnosis (per 10 years)	Polytomous[1]	0.49	<.001	0.07	0.567
	Wang[2]	0.50	<.001	0.06	0.642
	Chatterjee[3]	0.50	<.001	0.06	0.656
	Rosner[4]	0.51	<.001	0.08	0.542
Age at menarche (per 2 years)	Polytomous[1]	-0.11	0.225	-0.07	0.439
	Wang[2]	-0.11	0.252	-0.08	0.402
	Chatterjee[3]	-0.11	0.238	-0.08	0.395
	Rosner[4]	-0.11	0.222	-0.08	0.350
Nulliparous	Polytomous[1]	0.43	0.019	-0.01	0.957
	Wang[2]	0.48	0.009	-0.08	0.648
	Chatterjee[3]	0.49	0.008	-0.09	0.604
	Rosner[4]	0.50	0.006	-0.08	0.653
Age at first birth (per 5 years)	Polytomous[1]	-0.01	0.865	0.10	0.206
	Wang[2]	-0.02	0.830	0.11	0.194
	Chatterjee[3]	-0.01	0.910	0.10	0.199
	Rosner[4]	-0.01	0.947	0.11	0.168
Months of breastfeeding (per 6)	Polytomous[1]	0.08	0.218	-0.05	0.400
	Wang[2]	0.05	0.367	-0.03	0.648
	Chatterjee[3]	0.06	0.273	-0.03	0.564
	Rosner[4]	0.06	0.267	-0.03	0.558
Post-menopausal	Polytomous[1]	0.33	0.132	-0.84	<.001
	Wang[2]	0.29	0.185	-0.80	<.001
	Chatterjee[3]	0.30	0.158	-0.81	<.001
	Rosner[4]	0.31	0.147	-0.84	<.001
Pre-menopausal BMI (per 20)	Polytomous[1]	-0.86	0.015	0.18	0.602
	Wang[2]	-0.96	0.007	0.29	0.418
	Chatterjee[3]	-1.00	0.006	0.33	0.359
	Rosner[4]	-1.02	0.005	0.31	0.383
Post-menopausal BMI (per 20)	Polytomous[1]	-0.24	0.623	0.36	0.445
	Wang[2]	-0.40	0.359	0.55	0.204
	Chatterjee[3]	-0.41	0.311	0.56	0.155
	Rosner[4]	-0.42	0.293	0.59	0.130
Oral contraceptive use	Polytomous[1]	0.07	0.692	-0.22	0.190
	Wang[2]	0.05	0.785	-0.19	0.241
	Chatterjee[3]	0.04	0.789	-0.19	0.240
	Rosner[4]	0.05	0.778	-0.20	0.227
Family history of breast cancer	Polytomous[1]	0.49	0.037	-0.40	0.093
	Wang[2]	0.38	0.082	-0.28	0.198
	Chatterjee[3]	0.38	0.054	-0.28	0.155
	Rosner[4]	0.38	0.051	-0.26	0.173

[1] Polytomous logistic regression

[2] Two-stage meta-regression (Wang *et al.*, 2015)

[3] Two-stage regression with simultaneous estimation (Chatterjee, 2004)

[4] Stratified logistic regression (Rosner *et al.*, 2013)

subtypes based on immunohistochemical staining of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are commonly formed. Each of these tumor markers can be either positive (+) or negative (-) and the disease subtypes are defined as luminal A (ER+ or PR+, HER2-), luminal B (ER+ or PR+, HER2+), HER2-type (ER-, PR-, HER2+), and triple negative (ER-, PR-, HER2-). This configuration is not congruent with the second stage models described in Equations 2.3 and 2.4 on which the methods proposed by Wang *et al.* (2015), Chatterjee (2004) and Rosner *et al.* (2013) are based. Of the methods compared here, only polytomous logistic regression can address whether a risk factor effect differs across subtypes that are not formed by cross-classification of the individual tumor markers. This is important for epidemiologic researchers, who must carefully consider whether the individual tumor markers are of interest, or if it is truly a more complex aggregation of those tumor markers that is expected to demonstrate a differential association with risk factors.

The methods of Chatterjee (2004) and Rosner *et al.* (2013) were clearly designed with the goal of studying multiple tumor markers in a flexible modeling framework. Thus one can envision a study with a number of tumor markers where the dimension is reduced by eliminating selected, or all, interactions, and thereby permitting an analysis that would not be possible in the context of polytomous logistic regression. Further exploration is needed into the performance of each method under an increasing number of subtypes and risk factors.

This investigation was limited to methods that require pre-specification of subtypes. With increasing use of genomic profiling, often it will be of interest to first identify disease subtypes based on a large number of either binary or continuous tumor markers. Begg *et al.* (2013) proposed an approach to address this challenge by introducing a scalar measure of heterogeneity that allows an investigator to compare different subtyping configurations based on, for example, gene expression data. The ultimate investigation of risk factor associations with the resulting subtypes in this approach relies on polytomous logistic regression. The scalar measure additionally provides a quantification of the extent of heterogeneity for a given subtype solution, which the methods discussed in this chapter cannot accommodate. This approach will be investigated in more detail in the next chapter. Another consideration is the fact that all methods investigated in this chapter use a relative risk structure for defining and evaluating etiologic heterogeneity. An investigation of how the methods might be adapted to formulate the issues in the context of additive models is an area of future work.

In conclusion, the study of etiologic heterogeneity will become increasingly common in the age of genomic profiling and personalized medicine, and statistical methods are needed to reliably address these questions. The results of this investigation can serve to guide selection of a method that will favorably balance statistical and practical considerations.

Chapter 3

Validity of optimal D clustering[*]

The results of this chapter show when the strenth of structure in markers that truly represents etiologic heterogeneity exceeds the strength of structure in tumor marker data that is unrelated to disease risk, a novel method to cluster tumor markers and identify disease subtypes that differ maximally works well. However when this condition is not met, or when there are many tumor markers that simply represent noise, the truly etiologically heterogeneous subtype solution can still be identified by first performing variable selection to identify the disease markers most strongly related to risk factors.

In the previous chapter, a data example and simulation study were used to compare the statistical properties of methods for investigating etiologic heterogeneity by examining the differential effects of individual risk factors on pre-defined disease subtypes, or with respect to individual tumor markers. That study found that when the number of disease subtypes is small, a simple polytomous logistic regression model performs comparably to the more complex methods that have since been proposed (Zabor and Begg, 2017). Using polytomous

[*] Note that the contents of this chapter were submitted for publication in the Annals of Applied Statistics in July 2018, and are currently under review.

logistic regression, one can test for differences in relative risks of individual risk factors across disease subtypes. Polytomous logistic regression relies on there being a small number of disease subtypes in the context of a case-control study, but it is increasingly common for epidemiologic studies of cancer to obtain high-dimensional tumor marker information, such as gene expression, mutation, or copy number data. In such a setting, one must first employ substantive dimension reduction of the tumor marker data in order to establish a meaningful framework for examining the effects of the risk factors using a model such as polytomous logistic regression. To address this problem, earlier work sought to develop a method to identify the most etiologically distinct subtypes in the context of high dimensional tumor marker data (Begg *et al.*, 2013). This method involved two critical concepts. First, a scalar measure that captures the extent of etiologic heterogeneity of any succinct set of mutually exclusive subtypes was defined. Second, dimension reduction was accomplished through the use of unsupervised k-means clustering of the tumor marker data. Finally, the scalar measure of etiologic heterogeneity was calculated for each candidate subtype solution that resulted from the unsupervised clustering, and the best solution was chosen as the one that maximized the scalar measure of etiologic heterogeneity. In empirical studies using this method in breast cancer, melanoma and kidney cancer, the method led to solutions that were in line with relationships between risk factors and tumor markers that are already well known to cancer epidemiologists (Begg *et al.*, 2014, 2015; Mauguen *et al.*, 2017).

While these results are encouraging, they do not provide definitive evidence that the method can accomplish what it sets out to do, which is to identify the subtypes that are truly the most etiologically heterogeneous. There are reasons to be skeptical. Unsupervised k-means clustering is designed to identify subtypes that are distinctive with respect to the Euclidean distances of the markers of cases in a cluster compared to the markers of cases in other clusters. But there may exist clusters of cases that are separated on this basis but which have no relationship with etiology. Such "counterfeit" clusters could confound the ability of unsupervised clustering to find the clusters of cases that are truly etiologically distinctive. In this chapter a simulation framework is constructed to address the question of whether or not the method can be confounded by counterfeit clusters of this nature. The modeling framework involves creating datasets with the kind of high dimensional structure that is identifiable by clustering. Structure is created in the data, on the basis of specified tumor markers, that defines subtypes that are related to the risk factors, and counterfeit structure is created, on the basis of additional tumor markers, that is unrelated to the risk factors. Much larger numbers of tumor markers are also generated that neither possess structure nor are related to risk factors so that they simply introduce noise. The goals of this chapter are to understand the influences of these two sources of information that have the potential to prevent the method from identifying the truly etiologically distinct subtypes. Finally the influence of pre-clustering variable selection is explored as a strategy for improving the sensitivity of the method.

3.1 Methodologic details

This chapter focuses on a method for the analysis of case-control data, though the approach could be applied broadly by replacing polytomous logistic regression with an alternative regression approach appropriate to the study design under consideration. Also, because the data example comes from breast cancer, the term "tumor marker" will be used throughout, though all methods could be applied in other disease areas.

The method involves first performing unsupervised clustering of the tumor marker data. The goal of the unsupervised clustering is to obtain a variety of candidate sets of subtypes from which to choose the solution that optimizes the degree of etiologic heterogeneity observed, defined by a measure of etiologic heterogeneity denoted D, which is described in detail in the next paragraph. K-means clustering with many random starts is used to obtain candidate sets of subtype solutions. K-means clustering seeks clusters that exhibit high inter-cluster versus intra-cluster Euclidean distance. It is useful for this purpose because it does not typically reach a global maximum, and therefore when the process is repeated with different random starts many candidate solutions can be obtained, each at a different local maximum. In a traditional clustering analysis, one would then select the solution that maximizes the inter-cluster distance. However, interest is in identifying the class solution that maximizes etiologic heterogeneity rather than Euclidean distance, and so instead D is calculated for each of the candidate solutions that result from the different random starts of k-means clustering, and the solution that maximizes D is chosen as optimal. While alternative clustering algorithms to k-means clustering are not explored in detail, most other clustering methods are constrained to reach the same solution on every random start and so would not produce a variety of solutions that could be used to maximize the measure of etiologic heterogeneity. The fact that k-means clustering produces many potential clustering solutions is the feature that makes it especially useful for this purpose. Section 3.5 includes a cursory exploration of the performance of alternative clustering algorithms in the context of an analysis of this type.

The methodologic details of the approach have been outlined previously (Begg *et al.*, 2013). The method involves identifying different clustering solutions, each involving a set of M disease subtypes, and calculating a measure of etiologic heterogeneity for each solution. To calculate the measure, denoted D, one must first perform polytomous logistic regression of the risk factors on the subtypes and obtain estimated risk predictions from this model for each of the subtypes for each subject. Since the measure is population-based it is calculated solely using the study controls. Let i denote these control subjects $i = 1, \ldots, N_H$, where N_H denotes the total number of non-diseased control subjects, and let m index the set of disease subtypes, $m = 1, \ldots, M$. The risk predictions obtained from the polytomous logistic regression model for the ith individual are denoted r_{mi} such that the total risk of disease for that individual is $r_i = \sum_{m=1}^{M} r_{mi}$. Let the coefficients of variation of the subtype risks in the population be denoted $C_m^2 = v_m/\mu_m^2$ where $v_m = N_H^{-1} \sum_{i=1}^{N_H} r_{mi}^2 - \mu_m^2$ and $\mu_m = N_H^{-1} \sum_{i=1}^{N_H} r_{mi}$. Let the corresponding total coefficient of variation be denoted $C^2 = v/\mu^2$, where μ and v are the overall disease risk mean and variance. Then the measure of etiologic heterogeneity is defined as

$$D = \sum_{m=1}^{M} \pi_m C_m^2 - C^2, \tag{3.1}$$

where π_m represents the prevalence of the mth disease subtype. Further details of the rationale for this measure are provided in Begg *et al.* (2013). Even though absolute risks cannot be obtained from a case-control study, the relative risks obtainable from the polytomous logistic regression model can be used instead since all the terms in D are scale-adjusted.

3.2 Simulation methods

All statistical analyses were conducted using R software (R Core Team, 2018). An R package containing functions to perform the various calculations included in this analysis is available on GitHub at `https://github.com/zabore/riskclustr`. Additional code related to the specific simulations conducted can be found at `https://github.com/zabore/manuscript-code-repository`.

3.2.1 Risk factor generation

Individual risk factors are denoted X_p, $p = 1, \ldots, P$, and therefore $X = (X_1, \ldots, X_P)$. In order to most clearly highlight the concepts, a simplified setting is used where there are only $P = 2$ risk factors, so that $X = (X_1, X_2)$, and there are $M = 3$ disease subtypes that are heterogeneous with respect to the risk factors, as defined below. The density of the risk factors in the non-diseased, or control, subjects is assumed to follow $f(X) \sim N(\epsilon_0, \Sigma)$ and the density of the risk factors in diseased subjects is $f_m(X) \sim N(\epsilon_m, \Sigma)$, where $\epsilon_m = (\epsilon_{m1}, \epsilon_{m2})$ represents the mean vector of the two risk factors for disease subtype m. Equal covariance matrices, Σ, are assumed for diseased and non-diseased subtype risk factor distributions, since this is congruent with using polytomous logistic regression to model the conditional probabilities of the disease subtypes given the risk factors (Anderson, 1972). For convenience in the simulation studies, and without loss of generality, let Σ be the identity matrix, I, allowing mean values to represent standardized effect sizes. In all simulations ϵ_0, the mean vector for the two risk factors in non-diseased subjects, will be fixed at $\epsilon_0 = (0, 0)$ without loss of generality. Then set $\epsilon_1 = (e, 0)$, $\epsilon_2 = (e/2, e/2)$, and $\epsilon_3 = (0, e)$ for the three disease subtypes. In this way the strength of the differential risk factor associations with subtypes is represented by a scalar quantity, e.

3.2.2 Tumor marker generation

Risk factor data X are randomly generated from normal distributions with distinct mean vectors for the controls and for each disease subtype, as described above in Section 3.2.1. Etiologic heterogeneity is induced in the disease subtypes by generating the tumor marker data in such a way that certain tumor markers are correlated with the risk factors. Tumor marker data are simulated for case subjects only, and consist of tumor markers that are correlated with the risk factors, meaning that these markers possess etiologic heterogeneity, as well as tumor markers that are unrelated to the risk factors. Let k index tumor markers, $k = 1, \ldots, K$. Tumor markers are denoted $T = (T_A, T_B, T_C)$ where $T_A = (T_1, \ldots, T_{K_A})$ denotes the set of K_A tumor markers related to the risk factors and $(T_B, T_C) = (T_{K_A+1}, \ldots, T_K)$ denotes the set of tumor markers that are unrelated to the risk factors. The tumor markers that are related to the risk factors are distributed as $T_A \sim N(\lambda_{Am}, V_{Am})$ where $\lambda_{Am} = (\lambda_{Am1}, \ldots, \lambda_{AmK_A})$ represents the mean vector of tumor markers for subtype m disease, where, as indicated above $m = 1, 2, 3$. Through the mean vectors λ_{Am}, a relationship between a specific tumor marker and a specific disease subtype is induced. In this way correlations are also induced between the risk factors X and the individual tumor markers in T_A. The covariance matrix of these markers, V_{Am}, is set to be the identity matrix, I, for conceptual and interpretive simplicity.

The tumor markers that are unrelated to the risk factors include K_B tumor markers that have the kind of structure that is identifiable by clustering but that are generated independently of the risk factors. These are denoted $T_B = (T_{K_A+1}, \ldots, T_{K_A+K_B})$. Let l index the "counterfeit" subtypes defined by the markers in T_B, taking the values $l = 1, \ldots, L$. In all included simulation studies let $L = 3$. Assignment of each case to one of these L non-etiologically distinct subtypes is randomly generated from a multinomial distribution with L equal event probabilities, independent of the etiologically distinct class label m. The tumor markers that characterize these subtypes, T_B, are distributed as $T_B \sim N(\lambda_{Bl}, V_{Bl})$ where $\lambda_{Bl} = (\lambda_{Bl1}, \ldots, \lambda_{BlK_B})$ represents the mean vector of tumor markers for subtype l. Again, for simplicity, let the variance matrix $V_{Bl} = I$ throughout. Through the mean vectors λ_{Bl}, a relationship between a specific tumor marker and a specific counterfeit subtype l is induced, but since these subtypes are assigned to each case randomly, in contrast to the subtypes defined by T_A for which there is an induced relationship to the etiologically heterogeneous disease subtypes, there is no relationship induced between the tumor markers T_B and the risk factors X. K_C tumor markers, denoted $T_C = (T_{K_A+K_B+1}, \ldots, T_K)$, that have no defined structure are also generated . These markers are distributed as $T_C \sim N(0, I)$ and simply represent noise in the data.

3.2.3 Simulation parameters

For all included simulation studies set $N = 2000$ subjects, set $\pi_0 = 0.4$ to be the sampling proportion of non-diseased subjects and $\pi_m = 0.2$, $m = 1, 2, 3$, to be the sampling proportions of cases in the disease subtypes. Set $e = 1.5$ so that the mean vectors for the risk factors are $\epsilon_1 = (1.5, 0)$, $\epsilon_2 = (0.75, 0.75)$, and $\epsilon_3 = (0, 1.5)$ for the three subtypes. Generate 1000 simulated datasets.

The strength of the structure in T_A and T_B is quantified by the mean vectors λ_{Am} and λ_{Bj}, respectively, using 15 markers in each group so that $K_A = K_B = 15$. Let $\lambda_{A1} = (a, a, a, a, a, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, $\lambda_{A2} = (0, 0, 0, 0, 0, a, a, a, a, a, 0, 0, 0, 0, 0)$, and $\lambda_{A3} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, a, a, a, a, a)$ be the mean vectors for T_A, where $a = 1.3$ for weak structure, $a = 1.7$ for moderate structure, and $a = 2.1$ for strong structure. These mean values were selected to achieve separation in clusters, as measured by the inter-cluster dissimilarity, that is comparable to cluster separation previously seen in real data analyses (Begg *et al.*, 2014, 2015; Mauguen *et al.*, 2017). Let $\lambda_{B1} = (b, b, b, b, b, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, $\lambda_{B2} = (0, 0, 0, 0, 0, b, b, b, b, b, 0, 0, 0, 0, 0)$, and $\lambda_{B3} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, b, b, b, b, b)$ be the mean vectors for T_B, where b will be varied from 1.275 to 2.3 by small increments, for a continuum of weaker to stronger structure. In this way the strength of the structure that truly represents etiologic heterogeneity, and the strength of the unrelated counterfeit structure, are characterized by scalars a and b, respectively

3.2.4 Clustering methods

After simulating the tumor marker data, unsupervised k-means clustering with 1000 random starts is performed on the combined tumor marker data T, or subsets thereof, to obtain a set of candidate solutions. K-means clustering requires up-front specification of the number of subtypes of interest, and all included simulation studies specified that 3 subtypes be identified through clustering. In fact there is the possibility of 9 subtypes, defined by $\lambda_{Am}, m = 1, 2, 3$, and $\lambda_{Bl}, l = 1, 2, 3$. However, the goal is to identify the 3 subtypes defined by λ_{Am} that are etiologically heterogeneous. D is calculated based on the predicted risks from a polytomous logistic regression model that includes all risk factors X using Equation 3.1 for each of the candidate solutions that result from k-means clustering, and the clustering solution that maximizes D is identified. To assess the overall ability of k-means clustering to identify a reliable solution, the number of unique local solutions that occur among the 1000 random starts of k-means clustering is also recordeed in each simulated data set. The method relies on being able to determine the optimal solution by selecting the largest D from a variety of clustering solutions, so accuracy could be compromised in settings where too few unique cluster solutions are identified.

The misclassification rate is calculated as a measure of how closely aligned the class solution identified by k-means clustering is to the true class solution. While in a real data analysis the true class labels would not be known, and therefore it would not be possible to calculate misclassification rates, in the context of this simulation study misclassification rates are used to evaluate whether the approach is able to identify the truly etiologically heterogeneous class solution from which the data are generated. To accomplish this, the class labels for the subtype solution that optimizes D are cross-tabulated with the class labels for the truly etiologically heterogeneous subtype solution. Then the misclassification rate is calculated as $1 - \sum d/N_U$, where d indicates the diagonal of the cross-tabulation and N_U indicates the total number of cases. Since the labels that result from clustering are arbitrary, the class labels for the optimal solution must first be aligned with the true class labels by identifying the configuration that minimizes misclassification as defined above. An alternative measure for misclassification that is sometimes used would involve identifying whether a pair of cases who are classified similarly according to the truly etiologically distinct subtype solution are also classified similarly according to the subtype solution that optimizes D. The proportion of pairs of cases classified differently to the total number of pairs of cases would represent the misclassification. In a single simulation scenario, these two measures of misclassification were found to be highly comparable, with the measure based on pairs of cases consistently resulting in slightly lower levels of misclassification as compared to the measure based on each case's individual class membership, indicating that this alternative measure would not lead to meaningful differences in the pattern of results.

The influence of pre-clustering variable selection is then explored. To this end $K_C = 70$ additional tumor markers T_C that have no structure and are not related to the risk factors are included as a way to add noise to all simulation settings that were previously described, resulting in a total of 100 tumor markers. To achieve dimension reduction, the tumor markers are first ranked according to the D values that characterize the individual contributions of each marker to etiologic heterogeneity. These are obtained by creating two classes defined by high versus low values of each marker classified at the median, and using polytomous logistic regression and Equation 3.1 to obtain D. In this way an ordering of the tumor markers from most heterogeneous with respect to the risk factors to least heterogeneous with respect to the risk factors can be obtained. The analysis is then restricted to the markers with higher heterogeneity by sequentially reducing the ordered tumor marker set from $K = 100$ through $K = 5$ by increments of 5. The continuous versions of all selected tumor markers are then used in the clustering and the optimal D is identified, for each number of selected markers. This allows for exploration of the extent of misclassification as a function of the degree of dimension reduction.

Finally the influence of including an increasingly large set of tumor markers T_C that have no structure and are not related to the risk factors is examined. To simplify the interpretation, the K_B markers with counterfeit structure are eliminated. Set $a = 1.7$, representing moderate strength of structure in the K_A tumor markers that are related to the risk factors. Then increase the number of unstructured tumor markers that are included from $T_C = 50$ to $T_C = 15000$ to see how many such noisy tumor markers must be present in the data before the approach can no longer reliably identify the true class solution.

There are some additional considerations when using k-means clustering to obtain candidate sets of subtype solutions. It is sometimes possible for k-means clustering to result in a local maximum that has low inter-cluster dissimilarity. These scenarios were arbitrarily avoided by selecting the optimal class solution as the one that has maximal D from the subset of class solutions with sufficiently high dissimilarity, defined as a dissimilarity at least greater than the average dissimilarity across all k-means solutions. Additionally, on rare occasions k-means clustering will result in a solution with one or more very small classes. A class solution of this type would not be suitable for use in a polytomous logistic regression model, therefore calculation of D was restricted to solutions where each class consisted of at least 20 cases.

3.3 Results

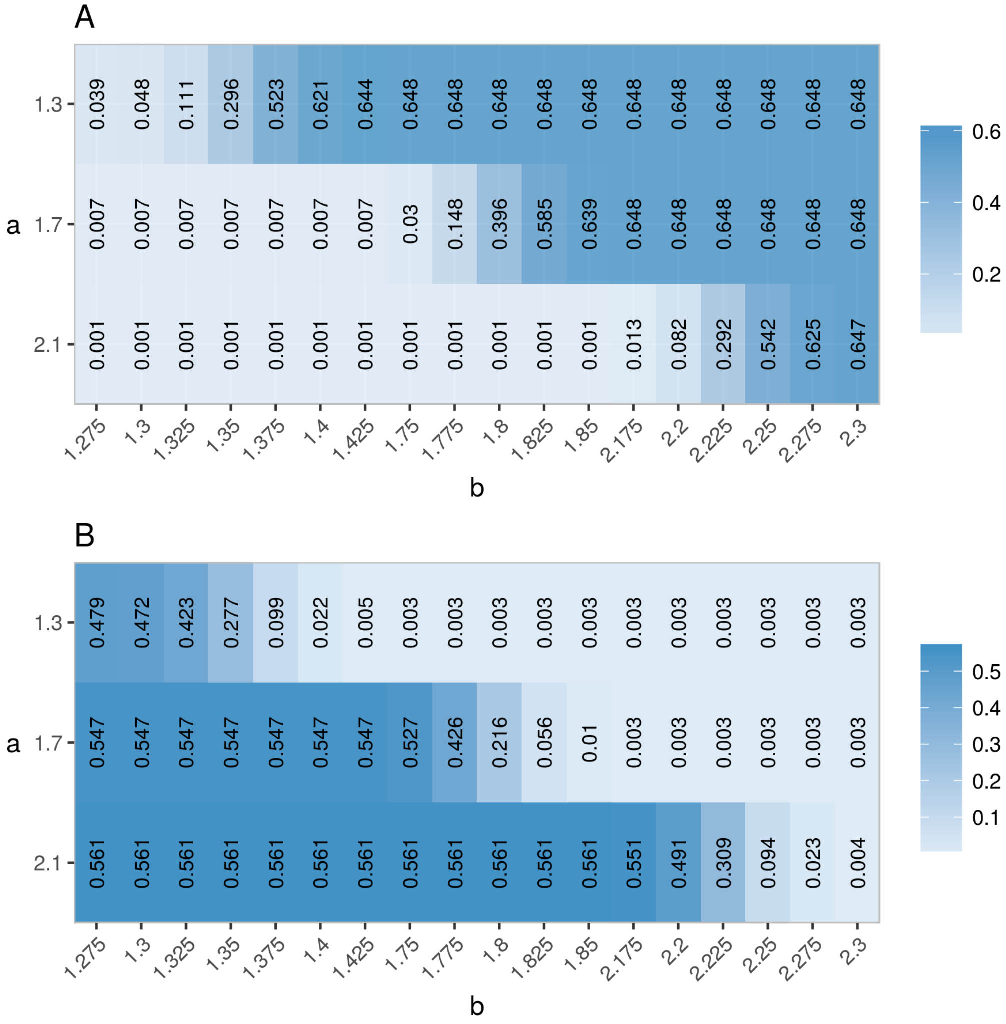
The primary measure of success is the accuracy by which the data are classified into the three truly etiologically distinct clusters $m = 1, 2, 3$. This success is represented by a low misclassification rate. Also, when the prevalence of each subtype in the source population is known, with known risk factor distributions, the true population value of D can be established. Given the mean vectors for the risk factors are $\epsilon_1 = (1.5, 0)$, $\epsilon_2 = (0.75, 0.75)$, and $\epsilon_3 = (0, 1.5)$ for each of three subtypes, and the prevalence of each of the three subtypes in the population is 0.2, the true population value of D is 0.506. The simulations seek to evaluate whether the method can achieve estimates of D that approach this true population value.

3.3.1 Impact of counterfeit structure

First, the influence of including tumor markers with counterfeit structure, unrelated to the risk factors, is addressed. Here the analysis is restricted to (T_A, T_B), the $K_A = 15$ tumor markers that have structure related to the risk factors and the $K_B = 15$ tumor markers that have structure unrelated to the risk factors, for a total of 30 tumor markers. On average across the various simulation settings, k-means clustering results in a variety of unique clustering solutions. The average number of unique solutions across all simulation settings is approximately 20 unique solutions from the 1000 random starts of k-means clustering, with a minimum number of 12, when $a = 1.7$ and $b = 1.275$, and a maximum of 48, when $a = 1.3$ and $b = 1.275$. Recall that a variety of solutions is needed in order for the method to produce candidate solutions with a range of etiologic heterogeneity, since identification of the various clustering solutions is not influenced directly by the risk factors.

The average misclassification rates are shown in Figure 3.1A. Note that the scale of the

Figure 3.1: Average minimum misclassification (A) and average maximum D (B) across varying strengths of structure in the tumor marker data based on $K_A = 15$ tumor markers with structure related to the risk factors and $K_B = 15$ tumor markers with structure unrelated to the risk factors. a = strength of structure in tumor markers related to risk factors, b = strength of structure in tumor markers unrelated to risk factors. A darker color indicates a higher value.



axis according to b is not evenly spaced but rather contains informative values. Darker colors denote higher values, which are undesirable since minimization of the misclassification rate is sought. As the strength of structure in the tumor markers that are unrelated to the risk factors, denoted by b, increases, the estimated misclassification rates increase. The smallest estimated misclassification rate of 0.001 occurs when $a = 2.1$, that is, when the structure in the tumor markers that are related to the risk factors is strong. This value indicates that on average only 1 case is being misclassified in this setting. When the strength of structure in the tumor markers that are related to the risk factors is weak and moderate, minimum misclassification rates of 0.039 and 0.007, respectively, are achieved. As b equals and then surpasses a, the misclassification rates increase rapidly. For example when $a = 2.1$ and $b = 2.3$ the misclassification rate is 0.647, which means that the chance of a misclassification is essentially random since there are three subtypes of which only one is the correct subtype. Clearly the method is preferentially selecting the counterfeit subtypes that are defined by the markers in T_B.

The average maximum values of D are shown in Figure 3.1B. Darker colors denote higher values, which are desirable since maximization of D is sought. As b increases, estimates of D decrease. The largest estimated D of 0.561 occurs when $a = 2.1$, where the structure in the tumor markers that are related to the risk factors is strong. This value exceeds the true population value of D of 0.506. This overoptimism is presumably due to the effect of picking the largest value of D in a setting where these are estimated and thus subject to statistical variation. Also, when $a = 2.1$ the estimated D drops quickly as b approaches and then exceeds a, similar to the trends seen in Figure 3.1A. When the strength of structure in the tumor markers that are related to the risk factors is weak and moderate, maximum D estimates of 0.479 and 0.547, respectively, are achieved and similar patterns to those seen for the misclassification rates are seen with respect to the rapid drop in estimated D as the strength of structure in the tumor markers that are unrelated to the risk factors equals and then surpasses the strength of structure in the tumor markers that are related to the risk factors. Overall these results indicate that the clustering will identify with high probability the class solution with the strongest signal, regardless of whether the solution represents clusters that are related to risk factors or not.

3.3.2 Pre-clustering variable selection

To try to improve these properties, the influence of pre-clustering variable selection is addressed. In these simulations $K_C = 70$ tumor markers in T_C that have no structure and are not related to the risk factors are included in addition to the tumor markers with structure in T_A and T_B, for a total of 100 tumor markers. The first observation is that when the full data $T = (T_A, T_B, T_C)$ are included in k-means clustering, there is no substantial impact on the results described in Section 3.3.1 (Figure 3.2). Next it is examined whether variable selection of tumor markers prior to clustering, based on their individual relationships with the risk factors as measured by their individual values of D, can improve the properties of the method. After rank-ordering the individual tumor markers based on their individual contributions to heterogeneity and reducing the set sequentially from 100 to 5 by increments of 5 tumor markers, the resulting misclassification rates are found to be uniformly low when a relatively small number of tumor markers are included, since most of the included tumor markers in this setting are selected to have structure related to the risk factors (Figure 3.3). Additionally, when the strength of structure in the tumor markers in the counterfeit clusters (represented by b) is less than the strength of structure in the tumor markers that are related to the risk factors (represented by a), all 100 tumor markers can be included

Figure 3.2: Average minimum misclassification (A) and average maximum D (B) across varying strengths of structure in the tumor marker data based on $K_A = 15$ tumor markers with structure related to the risk factors and $K_B = 15$ tumor markers with structure unrelated to the risk factors and $K_C = 70$ tumor markers that represent noise. a = strength of structure in tumor markers related to risk factors, b = strength of structure in tumor markers unrelated to risk factors. A darker color indicates a higher value.

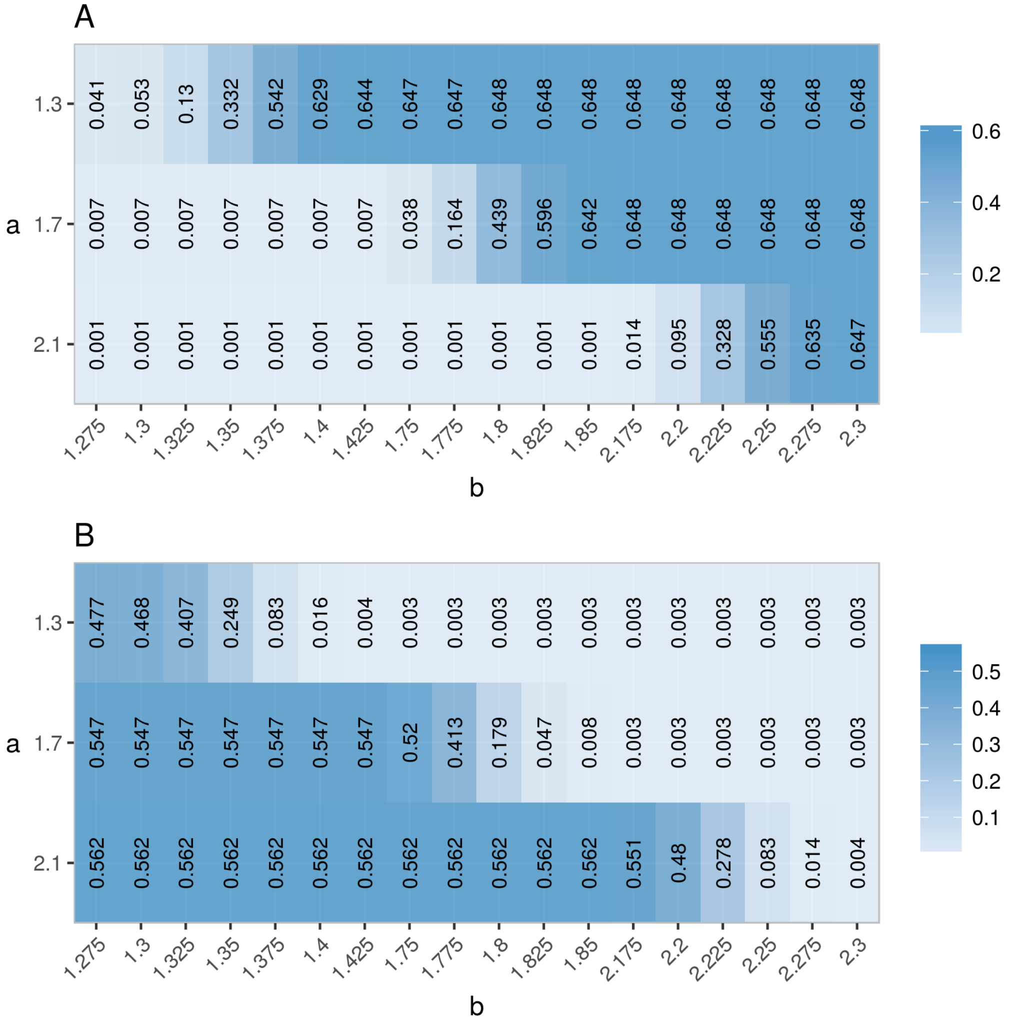
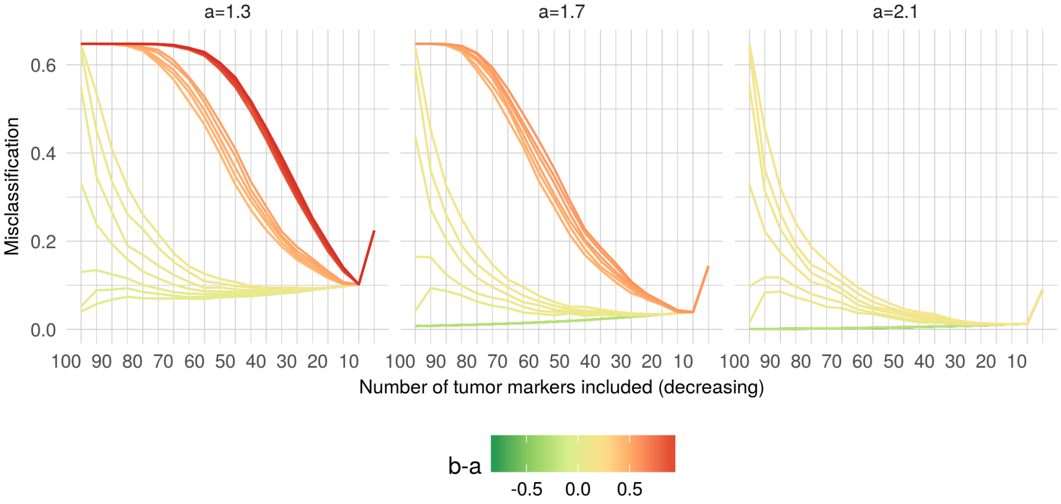


Figure 3.3: Average minimum misclassification for varying strengths of structure in the tumor markers with structure unrelated to the risk factors (denoted b), across different numbers of tumor markers included in k-means clustering based on univariate D ranking, according to the strength of structure in the tumor markers with structure related to the risk factors (denoted a). The color represents the difference in strength for the two types of markers, $b - a$.



with little to no impact on the results (green lines in Figure 3.3). When the strength of structure in the tumor markers in the counterfeit clusters is approximately equal to the strength of structure in the tumor markers that are related to the risk factors, reasonable misclassification rates can be achieved by reducing the dimension of the tumor marker set by about half (yellow lines in Figure 3.3). However, when the strength of structure in the tumor markers in the counterfeit clusters surpasses the strength of structure in the tumor markers that are related to the risk factors, much more stringent dimension reduction is required before reasonable misclassification rates can be achieved (red lines in Figure 3.3).

Since neither noisy markers in T_C nor the undesirable structured markers in T_B are related to disease risk, these markers compete with each other on an equal footing in the variable selection strategy, while the markers related to the risk factors in T_A are selected preferentially, as desired. This suggests a somewhat paradoxical result, that a larger number of noisy markers is beneficial by making it increasingly difficult for the confounding markers to be selected, provided that the variable selection is sufficiently strict (Figure 3.4). However, logic suggests that if there are too many noisy markers it will be increasingly difficult for the structure defined by markers with the true signal to be identified. To examine this, the number of noisy tumor markers, T_C, is increased incrementally from 50 to 15000 to see when the truly etiologically heterogeneous structure defined by the 15 markers in T_A can no longer be successfully identified. As the number of unstructured tumor markers increases, average minimum misclassification increases, with large changes in misclassification between 1500 and 5000 tumor markers (Figure 3.5). The impact of additional tumor markers is more pronounced at the smaller numbers when the strength of structure in the tumor markers related to the risk factors is weak.

3.4 Data application

The goal in the data application is to cluster gene expression data to identify breast cancer subtypes that demonstrate the highest degree of etiologic heterogeneity. Data from the Cancer and Steroid Hormone (CASH) breast cancer case-control study are analyzed. This study includes data from 2990 population controls and 551 breast cancer cases with a panel of gene expression data related to estrogen receptor status, 202 genes in total. The data also include standard breast cancer risk factors. In line with previous research (Gaudet *et al.*, 2011) age at diagnosis, race, premenopausal body mass index (BMI), postmenopausal BMI, family history of breast cancer, prior benign breast disease, age at menarche, nulliparity, number of live births (parity), age at first birth, months of breastfeeding, and menopausal

Figure 3.4: Number of each type of tumor marker (T_A, T_B, T_C) selected as the size of the selected tumor marker set increases, averaged across all simulation settings.

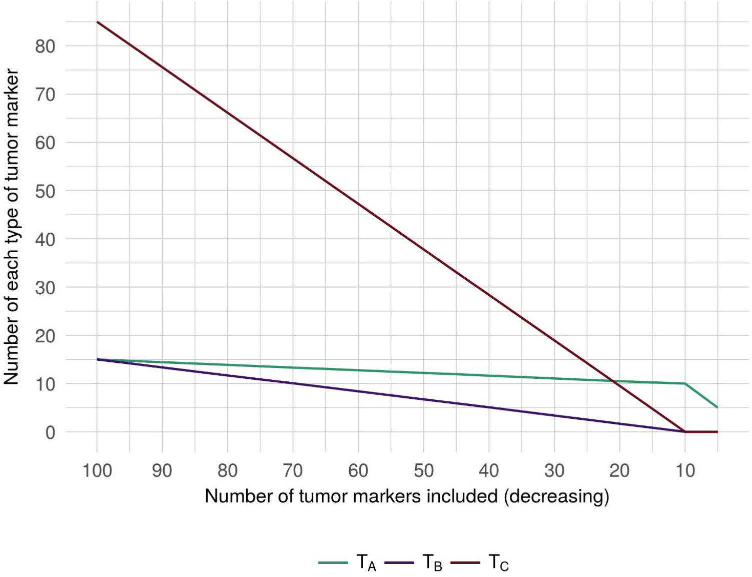
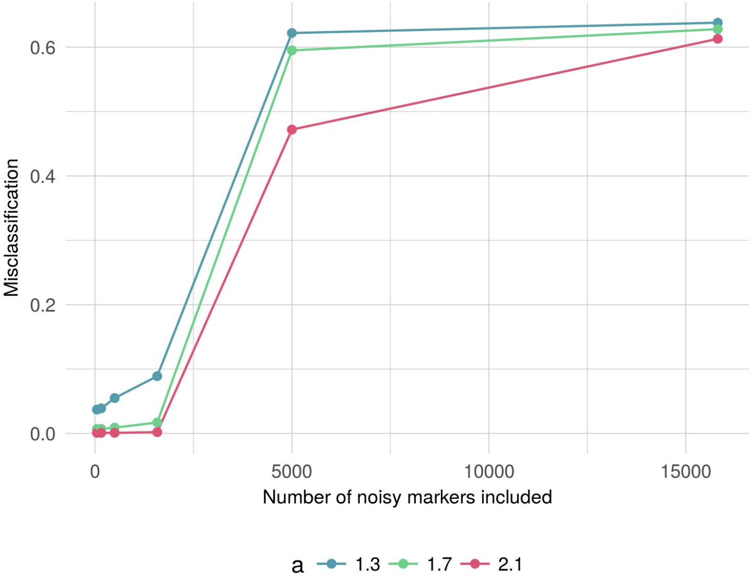


Figure 3.5: Average minimum misclassification as the number of unstructured tumor markers K_C increases, according to the strength of structure, a, in tumor markers related to risk factors.



status are included as risk factors, and estimates are additionally adjusted for study center. Results based on individual gene expression values were reported in Begg *et al.* (2015).

As in the simulation studies, k-means clustering on the full set of gene expression data is performed using 1000 random starts. Then D is calculated for each candidate solution based on a polytomous logistic regression model incorporating all risk factors and using Equation 3.1, and the optimal solution is identified as the one with maximal D. Next, to examine pre-clustering variable selection, D is calculated for each tumor marker individually to test for etiologic heterogeneity based on the gene expression values dichotomized at the median, using 500 permutations of the data, and these p-values are adjusted for multiple comparisons using the false discovery rate method. A reduced gene set is selected for k-means clustering based on the genes that have an adjusted p-value < 0.05. This method differs from that used in the simulation study because in a real data analysis interest is in selecting only genes that are believed to carry a meaningful heterogeneity signal whereas in the simulations the effect of including different numbers of noisy markers was being studied. Because there is a well-established set of four breast cancer molecular subtypes, based on immunohistochemical staining for estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2), clustering is focused on $M = 4$ classes.

Clustering the full set of 202 genes, a value of $D = 0.198$ is obtained. Limiting the gene set to those genes with a permutation-based p-value < 0.05 after adjustment for multiple comparisons results in a reduced set of 33 genes. Clustering the reduced gene set leads to a considerably higher optimal D of 0.331. In line with the simulation results, excluding tumor markers that are not associated with etiologic heterogeneity prior to clustering allows for identification of a more strongly heterogeneous solution. In a real application of this nature misclassification rates cannot be evaluated since the true subtypes are unknown.

However, the alignment of the class solutions can be examined with the established set of four breast cancer molecular subtypes: luminal A (ER+ or PR+, HER2-), luminal B (ER+ or PR+, HER2+), HER2-type (ER-, PR-, HER2+), and triple negative (ER-, PR-, HER2-) (Table 3.1). While there are differences between the class solutions based on the optimal D approach and the standard IHC-based molecular subtypes, there is more alignment of results after performing up front selection of gene expression values to include in the clustering, with the alignment increasing from 47% to 50%. Interestingly, the D estimate for the standard set of subtypes is 0.268, considerably lower than the optimal classification.

Table 3.1: Cross-tabulation of optimal D clustering results on full and reduced gene sets according to a well-established set of four subtypes in the CASH data.

	Standard molecular subtypes			
Optimal class solutions	HER2-type	Luminal A	Luminal B	Triple negative
Full gene set				
1	**23**	102	15	34
2	24	**134**	17	12
3	1	19	**2**	7
4	18	17	11	**82**
Reduced gene set				
1	**46**	60	27	27
2	8	**118**	12	13
3	1	89	**6**	4
4	11	5	0	**91**

3.5 Additional clustering algorithms

In the primary results, the use of a novel clustering strategy that involved performing k-means clustering with 1000 random starts, calculating D for each resulting candidate class solution, and then selecting the solution that results in maximal D as the optimal clustering solution was proposed. However, there are alternative clustering algorithms available. The original results based on this novel clustering strategy, referred to here as optimal D ("optD"), are compared with results produced from standard k-means clustering ("Kmeans"), partitioning around medioids ("PAM"), and model-based expectation-maximization (EM) clustering ("Mclust").

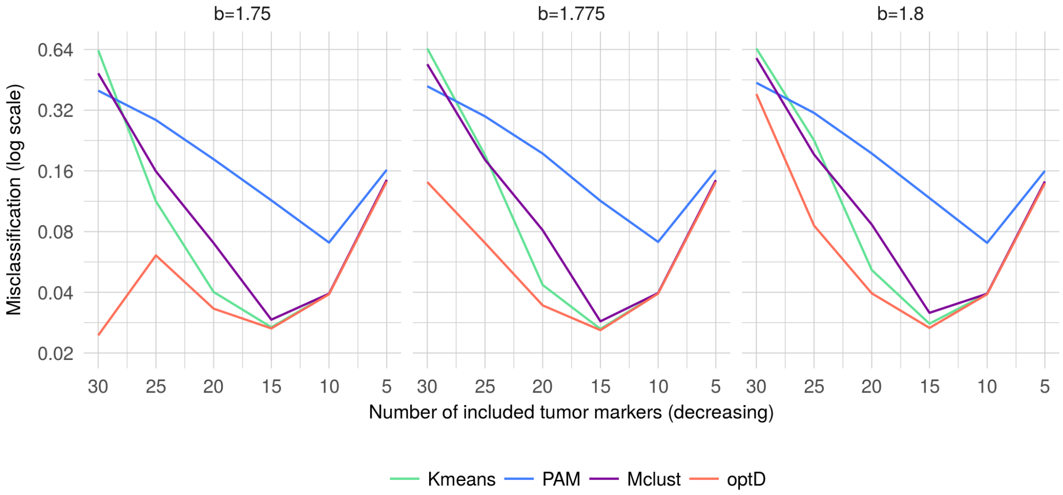
K-means clustering, when used in its standard form, selects the clustering solution that minimizes the ratio of the within-cluster sum of squares to the between-cluster sum of squares based on squared Euclidean distance (MacQueen, 1967). PAM is similar to k-means clustering, except it minimizes a sum of dissimilarities rather than the sum of squared Euclidean distances (Kaufman and Rousseeuw, 1987). Finally, model-based EM clustering (Fraley and Raftery, 2002) relies on Gaussian mixture modeling fitted via the EM algorithm. This approach fits numerous models and then selects the best model according to the Bayesian information criterion (BIC).

In all simulation settings there are $P = 2$ risk factors with the same mean vectors as described in the primary results. The sample size is $N = 2000$ and 1000 simulated datasets are generated. The proportion of controls is $\pi_0 = 0.4$. When k-means clustering is used, it is used with 1000 random starts. Up-front variable selection is performed as described in the primary methods. In all settings there are $K_A = 15$ tumor markers with structure related to the risk factors, which comprise $M = 3$ etiologically distinct subtypes with moderate strength of structure defined by $a = 1.7$.

This section seeks to accomplish two things:

1. To compare optimal D clustering to other clustering methods when there is structure related to the risk factors in addition to counterfeit structure

Figure 3.6: Misclassification rate according to strength of counterfeit structure, comparing clustering methods.



2. To evaluate optimal D clustering in comparison to other clustering methods when assumptions including constant variance, balanced class sizes, and normality do not hold

3.5.1 Clustering comparison in the presence of counterfeit structure

First, optimal D clustering is compared to other clustering methods when there is structure related to the risk factors in addition to counterfeit structure. To accomplish this, the performance of the different clustering algorithms is compared in the presence of $K_B = 15$ tumor markers with counterfeit structure that comprise $L = 3$ classes. Let $b = 1.75, 1.775,$ or 1.8 to explore several strengths of counterfeit structure. There are equal proportions of cases $\pi_1 = \pi_2 = \pi_3 = 0.2$ in each class related to the risk factors.

To assess the ability of the different clustering methods to identify the truly etiologically heterogeneous subtype solution as the strength of counterfeit structure varies, misclassification rates are examined. Figure 3.6 shows that optimal D clustering always performs at least as well as the other clustering algorithms in this setting, and when 30 or 25 of the 30 tumor markers are included in the clustering, optimal D clustering outperforms the other approaches (orange line). When dimension reduction to 20 or 15 of the tumor markers is performed, k-means clustering and model-based EM clustering perform approximately as well as optimal D clustering. PAM does not perform as well as the other approaches even with dimension reduction.

3.5.2 Clustering comparison under assumption violations

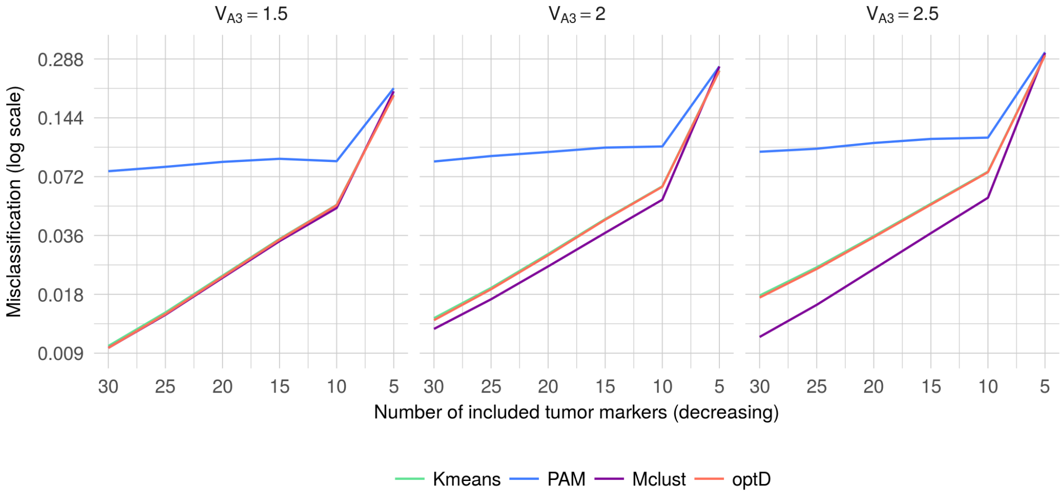
Next, optimal D clustering is evaluated in comparison to other clustering methods when assumptions do not hold. In the following sections counterfeit structure is not included, but rather the performance of the clustering methods under different assumption violations in the presence of $K_C = 15$ unstructured tumor markers is compared.

3.5.2.1 Heteroskedastic data

Because it is widely believed that k-means clustering does not perform as well when data are heteroskedastic, the influence of unequal variance in the tumor markers that comprise the different classes is examined. To accomplish this, in two of the three classes, the variance of the five tumor markers related to each class is fixed at $V_{A1} = V_{A2} = 1$ as in the primary results. However for the third class, the variance of the five tumor markers that comprise this class is varied from $V_{A3} = 1.5$ to $V_{A3} = 2$ to $V_{A3} = 2.5$ to explore the impact of increasing the variance for the tumor markers in only one of the three classes.

Figure 3.7 shows that optimal D clustering, k-means clustering, and model-based EM clustering have similar misclassification rates across all numbers of included tumor markers, though model-based clustering performs slightly better when the variance of the third class is

Figure 3.7: Misclassification rate according to variance of the third class, V_{A3}, comparing clustering methods.



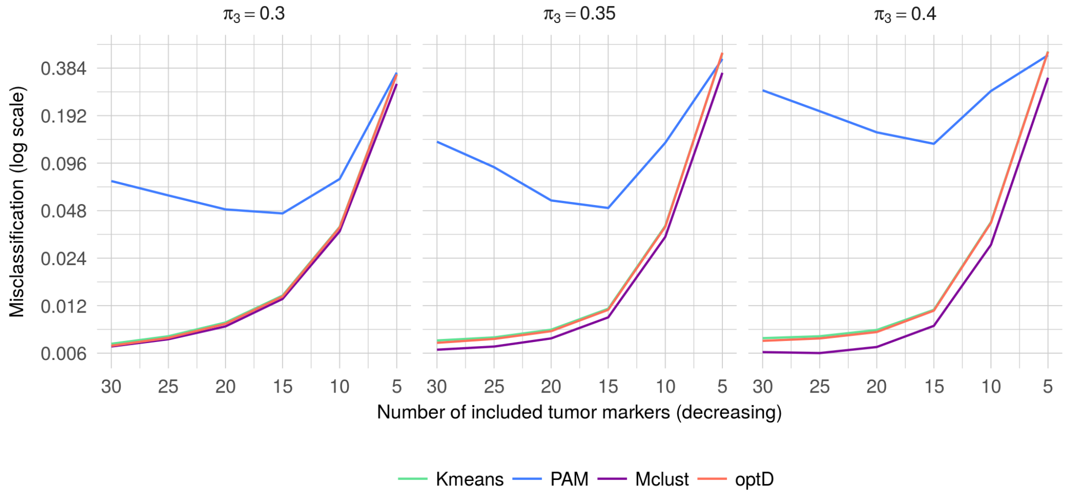
$V_{A3} = 2.5$ (note that the y-axis is on the log scale). PAM always has higher misclassification.

3.5.2.2 Unbalanced class size

Next the impact of unbalanced class sizes is examined. To accomplish this the proportions of cases in each class are varied such that two of the three classes have equal size and the third contains a larger proportion of cases. In the first setting $\pi_1 = \pi_2 = 0.15$ and $\pi_3 = 0.3$, in the second setting $\pi_1 = \pi_2 = 0.125$ and $\pi_3 = 0.35$, and in the third setting $\pi_1 = \pi_2 = 0.1$ and $\pi_3 = 0.4$.

Figure 3.8 shows that optimal D clustering, k-means clustering, and model-based EM clustering have similar misclassification rates, with model-based clustering having slightly lower misclassification when the third class contains $\pi_3 = 0.4$ of the cases. PAM has uniformly higher misclassification.

Figure 3.8: Misclassification rate according to proportion of cases in the third class, π_3, comparing clustering methods.

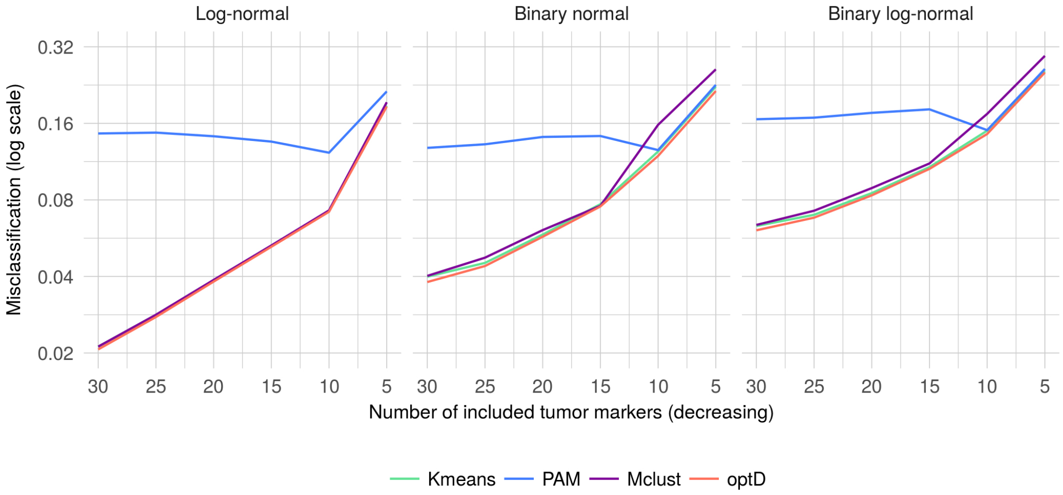


3.5.2.3 Non-normal tumor marker distributions

Because k-means clustering relies on Euclidean distance it is commonly understood that it is optimized for normally distributed data, which has been used in all results in the primary analyses. To examine this, the different clustering methods are compared under a variety of data distributions. Specifically, data from a log-normal distribution with mean 0 and standard deviation 0.5, binary data based on a dichotomization at the median of normally distributed tumor markers with mean 0 and standard deviation 1, and binary data based on a dichotomization at the median of log-normally distributed tumor markers with mean 0 and standard deviation 0.5 are clustered.

Figure 3.9 shows that in most cases, optimal D clustering, k-means clustering, and model-based EM clustering result in very similar misclassification rates. PAM results in uniformly higher misclassification rates.

Figure 3.9: Misclassification rate according to different data distributions, comparing clustering methods.



3.5.3 Clustering comparison conclusions

Overall, because optimal D clustering results in a variety of class solutions from which the solution that maximizes D, a measure of etiologic heterogeneity, is selected, in some circumstances this method is able to identify a class solution with lower misclassification as compared to other clustering methods. Additionally, though the optimal D clustering approach relies on k-means clustering, which in recent years has been utilized less than more modern clustering techniques such as model-based EM clustering, this analysis found that k-means clustering is not impacted by assumption violations more strongly than any of the other clustering methods examined here. Therefore, the novel clustering approach, optimal D clustering, which is based on k-means clustering, can reliably be used across a variety of data types and is able to identify the truly etiologically heterogeneous subtype solution in the presence of counterfeit structure more often than other clustering methods.

3.6 Discussion

In this chapter the performance of the optimal D clustering method was examined with respect to its ability to identify etiologically heterogeneous subtypes. When the structure in the tumor markers defining etiologic heterogeneity is strong, these etiologically distinct subtypes can be identified successfully with low misclassification even in the presence of weaker "counterfeit" structure. As the strength of the true structure decreases, misclassification rates increase. When the strength of the counterfeit structure surpasses that of the true structure, the desired etiologically heterogeneous subtypes can no longer be identified without variable selection to reduce dimension purposefully. Misclassification rates are not substantially impacted by the inclusion of a relatively small number of unstructured tumor markers, but this impact increases as the number of unstructured tumor markers becomes large. Since the ability of the method to identify the truly most etiologically distinct subtypes is impacted by both inclusion of tumor markers with strong structure that are unrelated to the risk factors and inclusion of a large number of unstructured tumor markers, a method to filter out such tumor markers will play an important role in any analysis of this type. Initial selection of tumor markers on the basis of their association with the risk factors led to improved performance across all simulation settings, and in the data application. This suggests that with careful use of up-front selection of tumor markers, the clustering method can reliably identify the truly etiologically distinct subtypes from high dimensional tumor marker data, although clearly the accuracy of the method will depend on the strength of the signal in the etiologic heterogeneity that distinguishes the subtypes. However, there is no obvious strategy for determining where to draw the line in selecting tumor markers for inclusion, and so in practice analyses of this type will require judgment.

Because of the inherent complexity of unsupervised clustering in high-dimensional data of this type, these simulation studies were conducted in a highly simplified context. In reality, genomic tumor marker data frequently have a much higher dimension and will possess much more complex structure than represented by this idealized framework. However, it is difficult to simulate complex structures meaningfully. The intent in this investigation has been to create a framework to permit one to infer generalizable messages that are relevant to the data analytic strategy.

An important area of future work will focus on how to estimate the optimal number of subtypes in a clustering analysis of this type. Estimation of the correct number of clusters is a challenge in any unsupervised clustering analysis, no matter the goal. One popular method is to use the gap statistic, which compares the within-cluster sum of squares to that expected under a null reference distribution for the data (Tibshirani *et al.*, 2002). In the included simulation studies the true number of etiologically distinct disease subtypes was fixed at three and a search for three clusters was specified in the k-means algorithm. However in a real data analysis the true number of subtypes will not be known. Future work is needed to create an appropriate method for estimating the optimal number of subtypes.

In summary, this chapter supports the following conclusion about the use of this clustering method for identifying etiologically heterogeneous subtypes. The method is capable of finding the true subtypes if they exist. However, the accuracy will depend on the strength of the heterogeneity signal, and the method is greatly enhanced with minimal cost by using pre-clustering variable selection of the tumor markers that are observed to be most strongly associated with the risk factors.

Chapter 4

Application to Carolina Breast Cancer Study

The results of this chapter show that by using a novel method to cluster gene expression data and identify disease subtypes that differ maximally with respect to etiologic heterogeneity using data from the Carolina Breast Cancer Study, an etiologically distinct 4-subtype solution was identified in a discovery stage, and in a validation sample showed reasonable validation in terms of both the highest-ranked individual genes and the subtypes formed by selected genes. *PSPHL* was the most important gene in defining etiologically distinct subtypes, and age, postmenopausal body mass index, ever use of oral contraceptives, and race are the risk factors that demonstrate etiologic differences across the optimal subtype solution.

As described in the Introduction (Section 1.3), there are four well-defined subtypes of breast cancer, known as luminal A, luminal B, HER2-type, and basal-like/triple negative. These subtypes have been used in numerous epidemiologic studies of etiologic heterogeneity, but they were originally discovered with the goal of separating patients according to prognosis, not risk. In previous work an approach that combines a search for candidate subtypes of cancer based on genomic information with use of a scalar measure to identify the most etiologically heterogeneous subtype solution was proposed (Begg *et al.*, 2013), and the previous chapter established that with rigorous up front selection of the tumor marker data the method identifies the true subtype solution with high probability. This chapter seeks to apply this approach to data from a large population-based breast cancer case-control study with available gene expression data to determine the optimally heterogeneous subtype solution with respect to risk for disease. Defining etiologically distinct subtypes of disease based on known risk factors will yield improved power to identify new risk factors, especially germline risk factors, that are expected to demonstrate etiologic heterogeneity and therefore will have increased effect sizes associated with certain subtypes of disease, thus leading to smaller and more efficient studies.

4.1 Carolina Breast Cancer Study data

The Carolina Breast Cancer Study (CBCS) was conducted in three phases from 1993 through 2013. The details of the study methodology, including sampling stratification and sampling frequencies, have been previously described in detail (Furberg *et al.*, 2002, 2003; Newman *et al.*, 1995). Briefly, women aged 20-74 living in certain counties in North Carolina and diagnosed with a first primary breast cancer were identified from the North Carolina Central Cancer Registry. Black women and women $<$ 50 years old were over-sampled with specific sampling probabilities. Controls were frequency matched to cases by race and 5-year age group. Phase 2 of the study included cases of DCIS, but the following analyses are limited to invasive breast cancer cases. Additionally, the analyses are limited to cases with available data on a panel of gene expression values. The analysis includes available, known risk factors for breast cancer.

Phases 1 and 2 had case-control designs whereas phase 3 was a case-only study design and thus did not include a sample of matched control subjects. There were a total of 861 cases and 790 controls in phase 1, 947 cases and 774 controls in phase 2, and 2976 cases in phase 3. Because an analysis of this type is already quite complex, methods to account for missing data such as multiple imputation are not feasible to implement, so a complete case analysis was conducted. See Figure 4.1 for details of patient exclusions.

4.1.1 Gene expression processing

The gene expression data in this study were obtained using a custom NanoString codeset for 406 genes of interest. See Section 4.5 for a full list of genes included in this analysis. Performance of the nCounter assay was assessed for efficiency and sub-optimal hybridization. Expression levels below the mean of negative controls were set to the mean background expression. Then positive control normalization multiplied all counts for a sample by the ratio of the average geometric mean of positive controls across all samples to the geometric mean of the sample-specific positive controls. Reference gene normalization was done in a similar way based on a set of 11 housekeeping genes. Batch effects were corrected by calibrating each lot based on a scaling factor calculated as the average geometric mean of endogenous genes across the three lots to the geometric mean of endogenous genes within lot. Finally, expression counts were $\log_2$ transformed.

Visualizations using 1-way dendograms and principal components analysis were used to identify major outliers. A sample was considered a major outlier if, after all of the pre-processing described in the previous paragraph was complete, the sample demonstrated extreme expression across all genes. During the quality control process 126 samples were flagged and excluded from analysis as major outliers according to principal components analysis (Figure 4.1). In a sensitivity analysis clustering a set of cases that included the major outliers, when compared to the results from the primary analysis with the major outliers excluded, between 94% and 99% of cases were classified similarly, suggesting that these major outliers did not comprise a separate etiologically distinct class. Gene expression values were standardized within sample by subtracting the mean gene expression for that sample and dividing by the sample standard deviation. Finally each gene's expression was median centered. Twenty cases from phases 1 and 2 and 20 cases from phase 3 were randomly selected for removal from the case group to test for differences between the various phases of the study, which were conducted at different times, without compromising the overall type I error of the primary results. The overall gene expression distributions between the different phases were compared using histograms and a Wilcoxon rank-sum test.

The final sample sizes for analysis are 83 cases and 739 controls from phase 1, 287 cases and 716 controls from phase 2, and 467 cases from phase 3 (Figure 4.1).

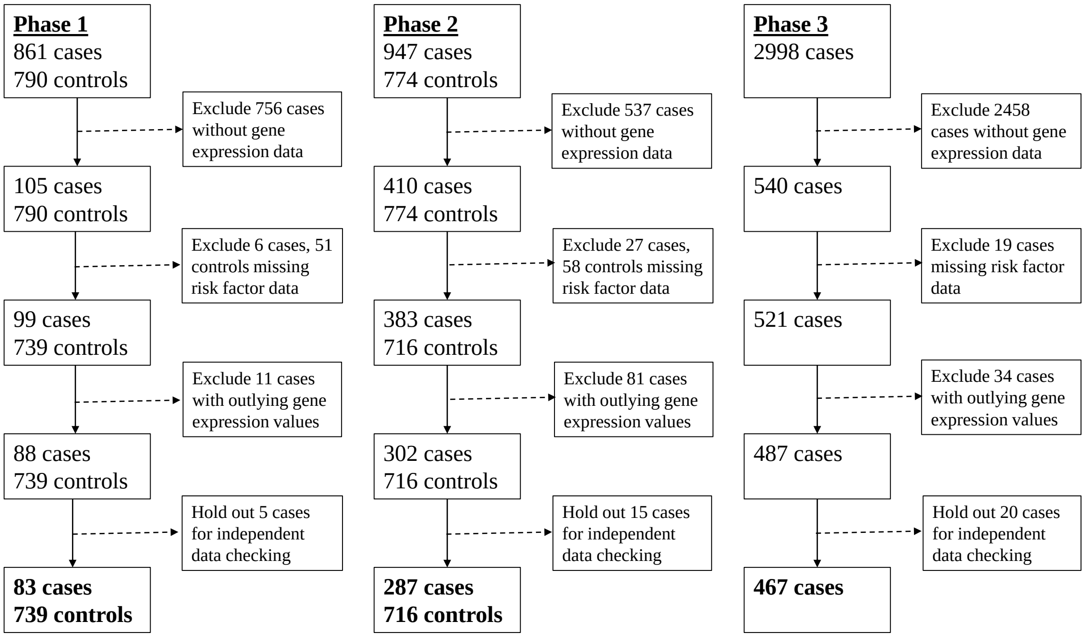
4.2 Methods

The analysis is conducted in two stages:

1. Cluster discovery stage. The 467 phase 3 cases with available gene expression and risk factor data are used to determine the optimally etiologically heterogeneous clustering solution using a case-only analytic setting.

2. Cluster validation stage. The 370 cases with available gene expression and risk factor data and the 1455 controls with available risk factor data from phases 1 and 2 are pooled, and the cases are assigned to a class solution based on the discovery results. Polytomous logistic regression is performed in the case-control setting to identify risk

Figure 4.1: Study exclusions


Phase 1
861 cases
790 controls
Exclude 756 cases without gene expression data
105 cases
790 controls
Exclude 6 cases, 51 controls missing risk factor data
99 cases
739 controls
Exclude 11 cases with outlying gene expression values
88 cases
739 controls
Hold out 5 cases for independent data checking
83 cases
739 controls
Phase 2
947 cases
774 controls
Exclude 537 cases without gene expression data
410 cases
774 controls
Exclude 27 cases, 58 controls missing risk factor data
383 cases
716 controls
Exclude 81 cases with outlying gene expression values
302 cases
716 controls
Hold out 15 cases for independent data checking
287 cases
716 controls
Phase 3
2998 cases
Exclude 2458 cases without gene expression data
540 cases
Exclude 19 cases missing risk factor data
521 cases
Exclude 34 cases with outlying gene expression values
487 cases
Hold out 20 cases for independent data checking
467 cases

factors with heterogeneous effects.

The goal of conducting the analysis in two stages, with discovery followed by independent validation, is to ultimately be able to obtain valid odds ratio estimates and p-values testing for heterogeneity across the subtypes. If the subtypes were discovered using the same data in which testing for heterogeneity was then conducted, the resulting p-values would be over-optimistic, since the risk factor distributions are pivotal in selection of the optimal subtype solution. The data were split into discovery and validation stages based on the original CBCS study design, which in phase 3 collected data only on cases with no matched controls, and in phases 1 and 2 collected data on cases with frequency matched controls. This approach of using the phase 3 data for discovery and the phases 1 and 2 data for validation is consistent with the original design of the study, which collected these data in different years and with different study designs. Use of the phases 1 and 2 data for validation additionally allows for calculation of standard case-control odds ratios.

4.2.1 Clustering methods

In the cluster discovery stage, a novel clustering method that uses unsupervised k-means clustering of the gene expression data in combination with calculation of a scalar measure of etiologic heterogeneity based on all available risk factors is applied to identify the optimally etiologically heterogeneous subtype solution, as detailed in Section 3.1 of Chapter 3. In the setting of a case-control study, the scalar measure of etiologic heterogeneity, denoted D, is calculated according to Equation 3.1. An approximation of this measure, denoted D^*, can be applied in the case-only setting, and details of this approach can be found in Begg *et al.* (2013). Briefly, whereas the variance and covariance terms in Equation 3.1 are averaged over the controls in a case-control setting, in a case-only setting they are averaged over the cases, which represent a risk-biased sample from the population. The goal of an analysis of this type is not to interpret the magnitude of D, but rather to use D to rank different subtyping schemes and identify the one that maximizes etiologic heterogeneity, and rankings based on D and D^* are expected to be broadly similar in practice.

K-means clustering is performed with 1000 random starts on the gene expression data in the discovery cases, to obtain a variety of class solutions. For each candidate solution identified by k-means clustering, D^* is calculated and the solution that maximizes D^* is selected as the optimal solution. To avoid solutions with subtypes with very small sample sizes, clustering solutions where a class had fewer than 20 cases were not considered. Additionally, because the true number of subtypes is unknown, the optimal 2-class, 3-class, 4-class, and 5-class solutions were identified and the ideal number of classes was later selected from these options. Solutions with more than 5 classes were not considered due to sample size limitations and in order to avoid overfitting.

4.2.2 Gene selection

The simulation studies presented in Chapter 3 found that when there exists strong multivariate structure in the tumor marker data that is unrelated to the risk factors of interest, or when there are many tumor markers that simply represent noise, the optimal D clustering method can fail to identify the subtype solution that is truly the most etiologically heterogeneous. However, this problem was relatively easily overcome by performing upfront variable selection on the tumor marker data. To accomplish the selection of genes, the individual D^* value for a 2-class solution is calculated for each gene. The 2-class solution for each gene is identified using standard k-means clustering optimized by inter-cluster distance and searching for two classes in the entire case sample (i.e. all phase 1, 2, and 3 cases combined). The genes are then rank-ordered according to their individual D^* values from the most heterogeneous to the least heterogeneous gene.

Because some genes in the included NanoString codeset are known to be highly correlated, an adjustment to the ranking based on correlation is considered. First, the top-ranked gene was used as the predictor in a linear model and each remaining gene was used as the outcome. The R^2 value was obtained from each linear model, where a larger R^2 represents a situation where the top-ranked gene better predicts the value of the gene under consideration in the outcome of the model. Each gene's individual D^* value was then weighted by the inverse of the resulting R^2 so that genes strongly related to the top-ranked gene would be down-weighted and genes weakly related to the top-ranked gene would be up-weighted. The remaining genes were then re-ranked. Next, the top two genes were used as predictors in a linear model and each remaining gene was used as the outcome. The R^2 values were obtained and used to weight the individual D^* values and adjust the ranking accordingly. This process was repeated until the top 10% of genes, or 40 genes, was obtained for inclusion in the reduced gene set. Solutions with fewer than 40 genes were investigated subsequently, as described in Section 4.2.6.

4.2.3 Validation

In the validation stage, each validation case (i.e. cases from CBCS phases 1 and 2) is assigned to the discovery class solution to which it is most similar. To accomplish this, the cluster centroids are first calculated in the discovery cases by averaging the data points within each of the M subtypes. Next, the Euclidean distance between each validation case and each of the M discovery cluster centroids is calculated. Each validation case is assigned to the closest cluster, defined as the one that demonstrates minimum Euclidean distance. D is calculated for each resulting subtype solution for comparison with the extent of heterogeneity as quantified by the traditional IHC and the traditional PAM50 subtype solutions.

To validate the subtypes identified in the discovery stage, the process of ranking the genes to obtain a reduced set and identifying the optimal solutions of different sizes is repeated in the phases 1 and 2 data, to obtain a validation solution. Both selection of the included genes and clustering of the reduced gene set to identify the optimal solution are conducted independently in the discovery and validation data. The optimal discovery class solution and the optimal validation class solution are then cross-tabulated, with validation cases assigned to discovery cluters as described above, to examine the alignment of validation cases according to the two independent classifications, as a way to assess the replicability of the identified subtypes.

The rest of the analyses are conducted in the validation data using the subtypes defined by the discovery class solution. Univariable associations between risk factors of interest and the resulting subtypes are examined using the Wilcoxon rank-sum test for continuous variables and the Chi-squared test for categorical variables. Then a polytomous logistic regression model is fit for the subtypes versus controls using all risk factors of interest, and adjusting for study phase 1 versus 2. For each risk factor p, $p = 1, \dots, P$, the regression parameters $\hat{\beta}_{pm}$ are obtained from the polytomous logistic regression model (Equation 2.1). These regression parameters are exponentiated to obtain odds ratios $\exp{(\hat{\beta}_{pm})}$ as a measure of effect size. Finally, a heterogeneity p-value is calculated for a test of the null hypothesis $H_{0_\beta} : \beta_{p1} = \cdots = \beta_{pm}$, which addresses the question of whether each risk factor has the same effect across all subtypes of disease. Traditional logistic regression models are additionally fit for each subtype separately versus the controls, allowing for incorporation of the offset terms required to correct for the original sampling design used in CBCS in order to obtain interpretable and generalizable odds ratio estimates. More details of the need for offset terms follow in Section 4.2.5.

4.2.4 Identifying the ideal number of classes

After identifying the optimal 2-class, 3-class, 4-class, and 5-class solutions in the discovery stage, permutation tests were used to test whether the optimal D^* for each class size carries a significant heterogeneity signal. To conduct the permutation tests, for each class size the unique k-means clustering solutions from the 1000 random starts are retained. Then the rows of the candidate class solutions are permuted so that the class label is rendered independent of the risk factor data, D^* is re-calculated for each candidate solution, and the solution that maximizes D^* is selected. This process is repeated 1000 times to obtain a null reference distribution for D^* for each class size. The p-value is calculated as the proportion of times the observed optimal D^* is less than the optimal D^* obtained from the permuted data, and serves as a test of $H_0 : D^* = 0$, i.e. a test of the hypothesis that none of the risk factors have differing effects across the classes. Failure to reject this null hypothesis implies that the candidate discovery class solution did not demonstrate etiologic heterogeneity with respect to the risk factors included in this analysis. Any candidate discovery class solution that results in a failure to reject this null hypothesis was not considered further.

Next the ideal number of classes must be selected from among the candidate discovery class solutions that demonstrate significant etiologic heterogeneity. To do so, the process of ranking the genes to obtain a reduced set and identifying the optimal solutions of different sizes is repeated in the phases 1 and 2 data, as described in Section 4.2.3. The two resulting 2-class, 3-class, 4-class, and 5-class validation class solutions are cross-tabulated with the discovery class solutions. The alignment of the two sets of class labels is used to select the ideal number of classes.

4.2.5 Additional methodologic considerations

Because black and young breast cancer cases were oversampled in CBCS, it is necessary to make some adjustments to the statistical analyses to account for this study design. Sampling weights were defined as the inverse of the sampling probability, and are required for inference to the general population. Offset terms were defined as the natural log of the ratio of the sampling probability for a case in a specific stratum of age and race to the sampling probability for a control in the same stratum of age and race, and are required to obtain valid odds ratio estimates. Polytomous logistic regression does not allow for incorporation of offset terms, so polytomous logistic regression will be utilized in identification of the optimal subtype solution, and to conduct statistical tests of the null hypothesis that risk factor effects are the same across the subtypes. Individual binary logistic regression models comparing each subtype to all controls, and incorporating the offset term, will be used to obtain corrected odds ratios.

4.2.6 Sensitivity analyses

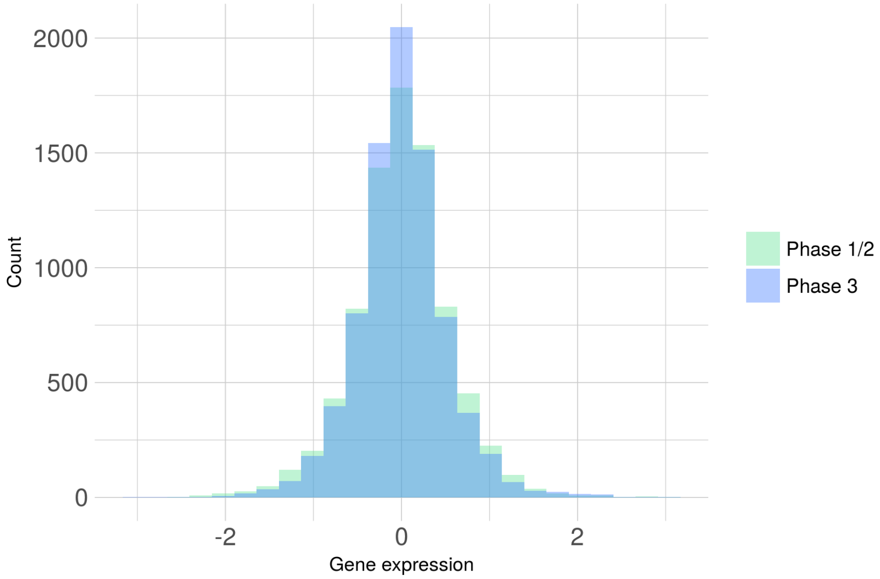
To examine the reliability of the primary results, a number of sensitivity analyses are conducted. The first is to assess the presumption that similar gene rankings would be obtained using the case-only D^* value and the case-control D value. In the primary analysis, the original design of the CBCS study was used to split the data into a discovery stage that included the phase 3 cases and a validation stage that included the phases 1 and 2 cases and controls. Because CBCS phase 3 was a case-only design, a case-only design was used in the discovery stage, where one subtype was used as the reference group in the polytomous logistic regression model to calculate D^* as described in Section 4.2.1. CBCS phases 1 and 2 included controls who were frequency matched to cases, allowing for calculation of D using a case-control design in the validation stage. By ranking the genes and obtaining the optimal subtype solutions using D^* in the discovery stage, and then applying the class labels to the validation cases, it is presumed that the results of a case-only approach are applicable to a case-control setting. This presumption is examined in two ways, with the goal of showing that similar gene rankings would be obtained using either a case-only or case-control design. First, a random sample of the phases 1 and 2 controls is obtained to use as an unmatched control sample for the phase 3 cases, and the genes are re-ranked according to their individual case-control D values calculated in the phase 3 cases and this random sample of phases 1 and 2 controls. This ranking is compared to the primary ranking obtained using the individual gene case-only D^* values based on the phase 3 cases. Second, the genes are re-ranked according to their individual case-only D^* values calculated in the phases 1 and 2 cases. This ranking is compared to the primary ranking obtained using the individual gene case-control D values calculated in the phases 1 and 2 cases and controls. This allows for assessment of the sensitivity of the method to the use of case-only D^* in the discovery stage and case-control D in the validation stage to obtain rankings for the individual genes. Because the primary goal of this sensitivity analysis is to compare rankings of individual genes according to different approaches for quantifying heterogeneity, the individual gene etiologic heterogeneity measures have not been adjusted for correlation.

A second sensitivity analysis assesses the impact of adjusting the selection of included genes for correlation. Some subsets of genes included in the study are known to be highly correlated. Therefore adjustment for correlation was used to avoid a situation where many genes carrying a similar heterogeneity signal were selected for inclusion, thus eliminating other genes that could provide more independent information. To examine the impact of this adjustment for correlation, with the goal of assessing whether a more etiologically heterogeneous solution can truly be identified based on selecting a more independent set of genes for inclusion in the clustering, the top 40 genes ranked by their individual D^* values in the discovery data, without adjustment for correlation, are used in clustering as described in Section 4.2.1. Validation D values based on these clustering results are calculated as described in Section 4.2.3 for the primary results, and the alignment of the class results based on the selected genes accounting for correlation and the selected genes not accounting for correlation is compared.

The third sensitivity analysis assesses whether the level of variable selection, which was somewhat arbitrarily set at 10% of the overall gene list, was sufficiently strict. Simulation studies presented in Chapter 3 revealed that problems of “counterfeit” structure and noise in the data can be overcome through pre-clustering selection of tumor marker data, so the goal of this sensitivity analysis is to try to assess whether the selection was sufficiently stringent, or if more etiologically heterogeneous solutions could be identified with further selection. To examine the impact of more stringent variable selection, the number of genes included in the analysis is systematically reduced. In the discovery cases, beginning with the list of 40 genes ranked from highest to lowest based on their individual D^* values adjusted for correlation, genes are removed from the bottom of the list one at a time. Each time a gene is removed the remaining genes are used in clustering as described in Section 4.2.1 to identify the optimal solution for the various class sizes. The validation cases are then assigned to these optimal classes based on different numbers of included genes as described in Section 4.2.3. The resulting subtype solutions based on different numbers of genes are compared according to their D values, as well as the alignment between discovery and validation class labels.

The etiologic heterogeneity of breast cancer has been examined in a previous study using gene expression data from the Cancer and Stereoid Hormone (CASH) case-control study (Begg *et al.*, 2015), and the final sensitivity analysis compares the rankings of common genes between the CASH and CBCS studies. The CASH study was a case-control study of breast cancer that assessed a panel of 202 gene expression values on the cases. Using the same risk factors as in the primary analysis of the CBCS data, D values are calculated for the individual genes in CASH. For the 38 genes that are in common between the CASH and CBCS studies, the rankings are obtained according to etiologic heterogeneity as measured in the CBCS phases 1 and 2 case-control study and the CASH case-control study. The goal is to determine whether any genes are commonly ranked highly across the two independent

Figure 4.2: Comparison of gene expression distributions in 20 held-out samples from phase 3 and 20 held-out samples from phases 1 and 2



studies, and as such the rankings in this sensitivity analysis have not been adjusted for correlation since interest is not in identifying a reduced gene set.

4.2.7 Software

All statistical analyses in this chapter were conducted using R software (R Core Team, 2018). An R package containing functions to perform specific calculations related to calculation of D and etiologic hetereogeneity p-values is available on GitHub at `https://github.com/zabore/riskclustr`.

4.3 Results

Figure 4.2 shows the distributions of gene expression values for the 20 held-out samples from phase 3 and the 20 held-out samples from phases 1 and 2, which do not differ significantly (p-value $= 0.295$). As a result, the discovery and validation cases can be combined to obtain the 2-class splits for each gene, since expression levels are similarly distributed. Differences in risk factor distributions between the discovery and validation case sets are tested using the full case population (Table 4.1). Validation cases have significantly younger age at first birth, lighter premenopausal and postmenopausal body mass index (BMI), lower frequency of nulliparity, lower frequency of ever oral contraceptive use, and are less frequently black as compared to the discovery cases. However these differences should not impact the primary analyses as some were invoked by the design of the study and interest is in relative measures of risk for breast cancer. Note that the frequencies presented in Table 4.1 are not generalizable to the population as a whole, but rather represent descriptive information about the study sample, as sampling weights have not been taken into account in these calculations.

Before beginning an investigation of subtypes in these data, it is of interest to examine the overall case-control odds ratios for the risk factors using logistic regression with the offset term incorporated to account for the oversampling in the study design, using the validation cases and controls from phases 1 and 2. Table 4.2 shows that increased age at diagnosis, first degree family history of breast cancer, and black versus white race are significantly associated with increased odds of breast cancer in this population. Increased postmenopausal BMI is significantly associated with decreased odds of breast cancer. The other risk factors have been implicated in other studies in the literature, though their effects in the CBCS data are small and not statistically significant (Huang *et al.*, 2000). It is possible that the sampling design of CBCS, which oversampled black women and women < 50 years old, could have led to distributions of hormonal risk factors that are not representative of the general population, thus obscuring common associations with risk for breast cancer.

Table 4.1: Comparison of risk factor distributions between the discovery and validation case sets. Numbers presented are median (minimum, maximum) for continuous variables and frequency (percent) for binary variables.

Variable	Discovery (n = 467)	Validation (n = 370)	p-value
Age at diagnosis	49 (23, 74)	49 (23, 73)	0.77
Age at menarche	12 (8, 18)	13 (8, 21)	0.07
Age at 1st birth	23.5 (13, 44)	22.4 (14, 39)	0.001
Months breastfeeding	0 (0, 95)	0 (0, 58)	0.11
Premenopausal BMI[†]	30 (17.7, 62.5)	29.1 (15.1, 53)	<.001
Postmenopausal BMI[†]	31.2 (17.7, 51.4)	29.7 (14.3, 53.5)	<.001
Nulliparous	88 (18.8)	50 (13.5)	0.04
Postmenopausal	233 (49.9)	188 (50.8)	0.83
Ever use of OCs[‡]	372 (79.7)	250 (67.6)	<.001
Family history*	84 (18)	64 (17.3)	0.86
Black race	275 (58.9)	179 (48.4)	0.003

[†] BMI = body mass index

[‡] OC = oral contraceptive

*First degree family history of breast cancer

4.3.1 Discovery results

First, in Figure 4.3 the correlation among the top 40 genes based on ranking the genes according to their individual D^* values (Figure 4.3A) is compared to the correlation among the top 40 genes based on ranking the genes with adjustment for correlation (Figure 4.3B). The correlation among the top genes was very strong, but after weighting D for correlation and re-ranking the genes, a set of genes that carry more independent information is selected. These top 40 genes, along with their individual D^* values, are listed in Table 4.3. *PSPHL* is the top-ranked gene, with a D^* value of 0.285. Note that some genes known to play a

Table 4.2: Overall case-control logistic regression results in validation data. Additionally adjusted for study phase. OR = odds ratio; CI = confidence interval.

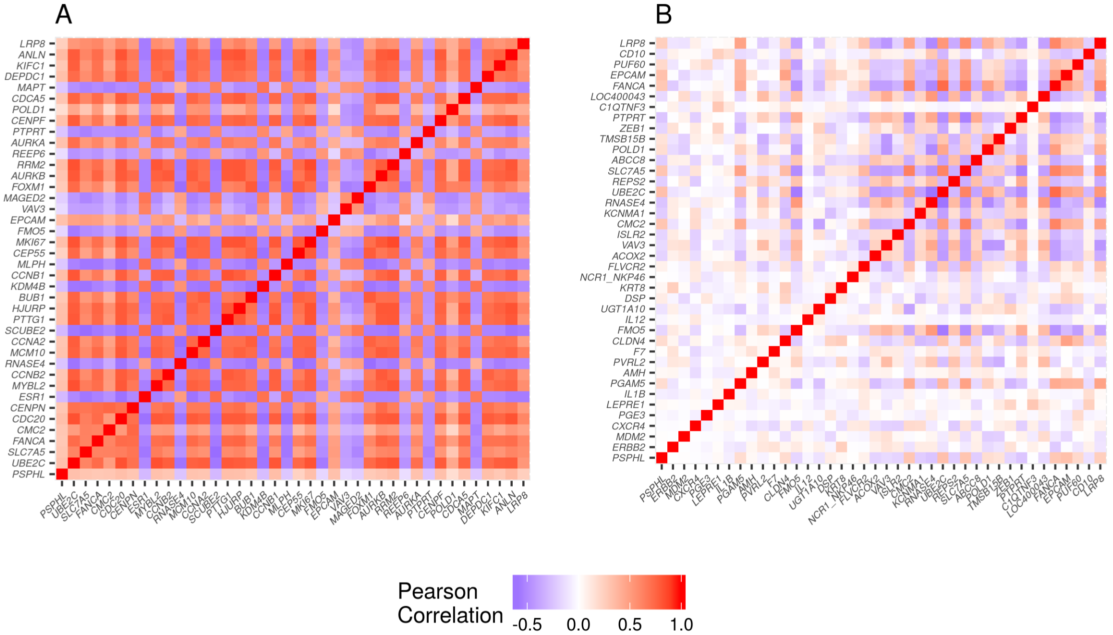
Variable	OR (95% CI)	p-value
Age at diagnosis (per 10 years)	1.68 (1.41 - 1.99)	<.001
Age at menarche (per 2 years)	0.91 (0.78 - 1.06)	0.21
Age at 1st birth (per 5 years)	1.06 (0.93 - 1.20)	0.40
Months breastfeeding (per 6)	0.95 (0.87 - 1.02)	0.17
Premenopausal BMI† (per 20 units)	0.90 (0.56 - 1.46)	0.67
Postmenopausal BMI† (per 20 units)	0.45 (0.27 - 0.77)	0.004
Nulliparous	1.03 (0.71 - 1.49)	0.88
Postmenopausal	1.08 (0.74 - 1.58)	0.69
Ever use of OCs‡	1.17 (0.87 - 1.57)	0.31
Family history*	1.53 (1.11 - 2.13)	0.01
Black vs white	1.31 (1.00 - 1.71)	0.046

†BMI = body mass index

‡OC = oral contraceptive

*First degree family history of breast cancer

Figure 4.3: Heatmaps of correlation among the top 40 genes in the discovery data (A) based on ranking the genes using their individual D^* values and (B) based on ranking the genes with adjustment for correlation.



role in subtyping breast cancer, including *ESR1* with a D^* value of 0.119 and *SCUBE2* with a D^* value of 0.100, are not included in the list of selected genes after accounting for correlation, as they were strongly correlated with the top-ranked gene. Next the selected 40 genes are clustered using the original continuous data, and for each class size the candidate solution that maximizes D^* is selected as the optimal solution. The true optimal D^* value is signficantly greater than the null reference distribution for all class sizes, as indicated by the purple asterisks denoting the observed optimal D^* lying far from the null distribution of D^* (Figure 4.4) and by the significant permutation-based p-values (Table 4.4).

Next the ideal number of classes to use in the remaining analyses is determined by examining the alignment between the optimal solutions identified using the discovery phase 3 data and the optimal solutions identified independently using the validation data from

Table 4.3: Top 40 genes in the discovery data, selected by ranking genes according to D^* values weighted to adjust for correlation.

Gene	D^*	Rank
PSPHL	0.285	1
ERBB2	0.034	2
MDM2	0.027	3
CXCR4	0.040	4
PGE3	0.039	5
LEPRE1	0.008	6
IL1B	0.045	7
PGAM5	0.076	8
AMH	0.052	9
PVRL2	0.051	10
F7	0.060	11
CLDN4	0.074	12
FMO5	0.109	13
IL12	0.015	14
UGT1A10	0.066	15
DSP	0.048	16
KRT8	0.081	17
NCR1_NKP46	0.023	18
FLVCR2	0.053	19
ACOX2	0.068	20
VAV3	0.101	21
ISLR2	0.052	22
CMC2	0.128	23
KCNMA1	0.078	24
RNASE4	0.114	25
UBE2C	0.142	26
REPS2	0.073	27
SLC7A5	0.136	28
ABCC8	0.072	29
POLD1	0.095	30
TMSB15B	0.075	31
ZEB1	0.069	32
PTPRT	0.094	33
C1QTNF3	0.040	34
LOC400043	0.087	35
FANCA	0.138	36
EPCAM	0.097	37
PUF60	0.077	38
CD10	0.046	39
LRP8	0.094	40

Figure 4.4: Histograms of the null reference distributions of D^* for each class size in the discovery data. The purple asterisk denotes the observed optimal D^* values.

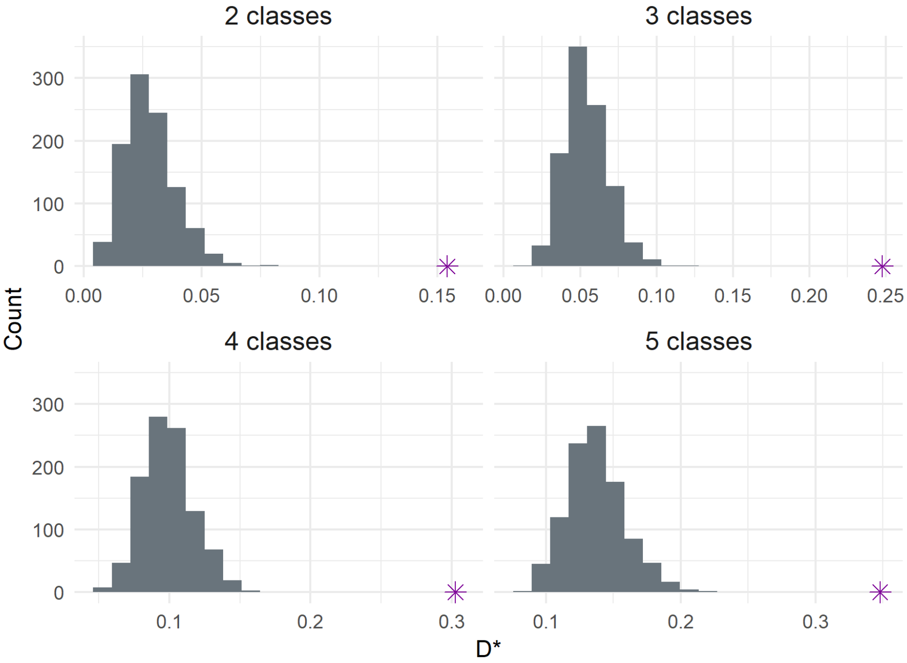
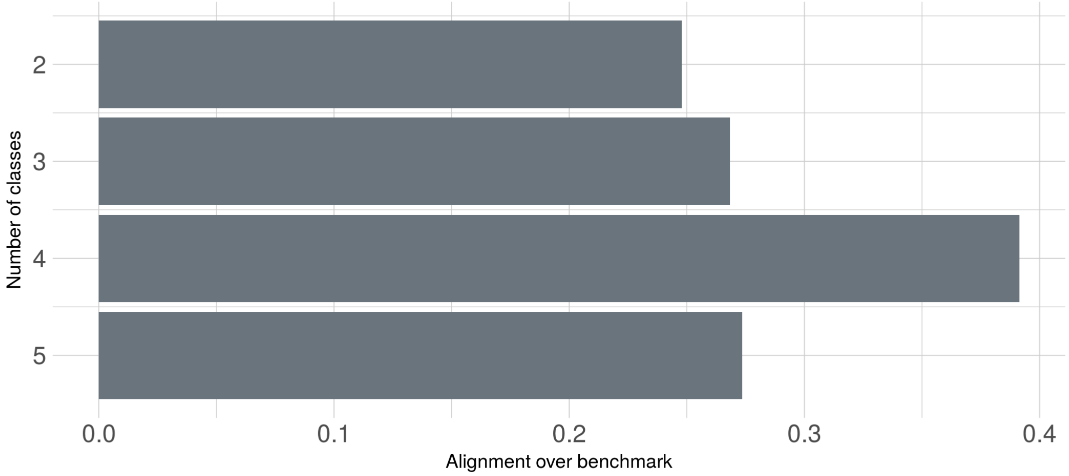


Table 4.4: Optimal D^* in the discovery data for each class size, with permutation-based p-values.

Number of classes	D^*	p-value
2	0.154	$< .001$
3	0.248	$< .001$
4	0.303	$< .001$
5	0.348	$< .001$

Figure 4.5: Alignment over benchmark level between optimal class labels identified independently in the phase 3 and phases 1 and 2 data, by class size.



phases 1 and 2, as described in Section 4.2.4. The alignment is calculated by cross-tabulating the two sets of class labels for each class size and then obtaining the proportion of cases on the diagonal. Since class labels from k-means clustering are arbitrary, the two sets of class labels must first be aligned, as desribed in Section 3.2.4 of Chapter 3. Additionally, because it is easier to achieve alignment when there are fewer classes, a benchmark level of alignment is established for each class size based on the sum of the squared relative frequencies of the discovery classes. Figure 4.5 shows the amount of additional alignment achieved above the benchmark level, and indicates that the 4-class solution achieves the highest level of additional alignment. This strong alignment combined with the fact that the traditional IHC and PAM50 4-class solutions are already well-accepted in breast cancer, the 4-class solution is selected for use in the validation stage of this analysis. In the discovery data there are 114 (24.4%) cases in subtype 1, 174 (37.3%) cases in subtype 2, 135 (28.9%) cases in subtype 3, and 44 (9.4%) cases in subtype 4.

4.3.2 Validation results

Each validation case from the phase 1 and 2 data is assigned to its subtype based on the 4-class discovery solution as described in Section 4.2.3. There are 84 (22.7%) cases assigned to subtype 1, 110 (29.7%) cases assigned to subtype 2, 107 (28.9%) cases assigned to subtype 3, and 69 (18.6%) cases assigned to subtype 4. The high alignment of 66% between the 4-class solutions identified independently in the phase 3 and phases 1 and 2 cases indicates that the subtypes are reasonably replicable (Table 4.5).

Table 4.5: Alignment between the optimal discovery 4-class solution and the optimal validation 4-class solution, in the validation cases.

	Validation 4-class			
Discovery 4-class	1	2	3	4
1	**74**	0	11	12
4	0	**60**	26	7
2	7	46	**66**	6
3	3	4	4	**44**

The D value for the optimal discovery 4-class solution in the validation cases is 0.271. Table 4.6 shows the alignment between the traditional 4-class subtyping system based on three IHC markers (ER, PR, and HER2) and the optimal 4-class solution. 52.5% of cases are classified similarly according to the two subtyping schemes. The D value for the traditional IHC 4-class system is 0.165, which is much lower than the D value of 0.271 for the optimal 4-class solution. Similarly, Table 4.7 shows the alignment between the traditional 4-class subtyping system based on the PAM50 gene expression panel and the optimal validation 4-class solution. 60.2% of cases are classified similarly according to the two subtyping schemes. The D value for the traditional PAM50 4-class system is 0.153, which again is much lower

Table 4.6: Cross-tabulation of traditional 4-class subtypes based on IHC markers and optimal 4-class solution in phase 1 and 2 cases. Note that 10 cases are missing values for the traditional IHC 4-class solution.

	Optimal 4-class			
Traditional IHC 4-class	1	2	3	4
Triple negative	**68**	7	17	21
Luminal A	13	**95**	47	39
HER2-type	1	1	**21**	1
Luminal B	0	3	21	**5**

than the D value of 0.271 for the optimal 4-class solution. These results suggest that while the optimal 4-class solution is fairly well-aligned with the more traditional 4-class subtyping solutions, the most etiologically heterogeneous subtyping solution had not yet been defined, as the identified optimal solution demonstrates a substantially larger etiologic heterogeneity signal. It is of interest to note that optimal class 1 is particularly well aligned with the triple negative and basal-like subtypes, and optimal class 2 is quite strongly aligned with the luminal A subtype.

Table 4.7: Cross-tabulation of traditional 4-class subtypes based on the PAM50 gene expression panel and optimal 4-class solution in phase 1 and 2 cases. Note that 26 cases classified as normal-like by the traditional PAM50 solution are excluded.

	Optimal 4-class			
Traditional PAM50 4-class	1	2	3	4
Basal-like	**79**	0	7	43
Luminal A	1	**93**	43	15
HER2-type	1	0	**32**	2
Luminal B	3	4	18	**3**

Next the univariable associations between the risk factors of interest and the optimal

Table 4.8: Risk factor distributions in phase 1 and 2 cases according to the optimal 4-class solution. Numbers presented are median (minimum, maximum) for continuous variables and frequency (percent) for binary variables.

	Optimal 4-class				
Variable	1 (n = 84)	2 (n = 110)	3 (n = 107)	4 (n = 69)	*p*-value
Age at diagnosis	45 (24, 73)	53 (24, 73)	49 (27, 73)	50 (23, 73)	<.001
Age at menarche	12 (9, 16)	13 (8, 18)	13 (9, 21)	13 (9, 16)	0.13
Age at 1st birth	20 (14, 36)	22.4 (14, 35)	22.4 (14, 39)	22 (14, 36)	0.14
Months breastfeeding	0 (0, 58)	0 (0, 54)	0 (0, 31)	0 (0, 58)	0.18
Premenopausal BMI†	29 (18, 47)	29 (20, 53)	29 (15, 46)	29 (18, 53)	0.44
Postmenopausal BMI†	30 (18, 48)	30 (14, 49)	30 (18, 53)	29 (18, 33)	<.001
Nulliparous	8 (9.5)	17 (15.5)	18 (16.8)	7 (10.1)	0.37
Postmenopausal	28 (33.3)	64 (58.2)	54 (50.5)	42 (60.9)	0.001
Ever use of OCs‡	63 (75)	73 (66.4)	64 (59.8)	50 (72.5)	0.12
Family history*	14 (16.7)	21 (19.1)	18 (16.8)	11 (15.9)	0.95
Black race	54 (64.3)	42 (38.2)	56 (52.3)	27 (39.1)	0.001

†BMI = body mass index

‡OC = oral contraceptive

*First degree family history of breast cancer

4-class solution are presented in Table 4.8. Note that these results are not generalizable to the population as a whole, but rather represent descriptive information about the included study sample, as sampling weights have not been taken into account in these calculations. Age at diagnosis, postmenopausal BMI, postmenopausal status, and race are all significantly associated with the optimal 4-class solution on univariable analysis. Cases in optimal class 1 appear to be younger, less frequently postmenopausal, and more frequently black. Cases in optimal class 2 appear to be older, more frequently postmenopausal, and more frequently white. Odds ratios and 95% confidence intervals from multivariable logistic regression models separately comparing each optimal class to the controls, and accounting for offset terms, are shown in Figure 4.6. The p-values on the plot are from the test for heterogeneity based on a multivariable polytomous logistic regression model, as described in Section 4.2.3. Age at diagnosis, postmenopausal BMI, ever use of oral contraceptives, and race all demonstrate significant heterogeneity across the four subtypes in multivariable analysis. Older women have increased odds of class 2 breast cancer, women with higher postmenopausal BMI have decreased odds of class 4 breast cancer, women who ever used oral contraceptives have decreased odds of class 3 breast cancer, and black women have increased odds of class 1 breast cancer and decreased odds of class 2 breast cancer.

A heatmap of expression values for the 40 included genes according to the optimal 4-class solution in the validation cases is shown in Figure 4.7. Optimal class 4 tends to have lower expression levels, especially for *PSPHL*, a gene with known race associations (Parada *et al.*, 2017). Recall that class 4 has a lower frequency of black women, and *PSPHL* is known to be more highly expressed in black women. Notably, optimal class 3 has higher expression of *ERBB2*, the gene that represents human epidermal growth factor receptor 2 (HER2). Tables 4.6 and 4.7 had previously indicated that HER2-type cancers were almost exclusively classified into optimal class 3.

4.3.3 Sensitivity to use of case-only data to identify subtypes

Because phase 3 of CBCS did not include frequency matched controls, the subtypes in the discovery stage of this analysis were identified in a case-only setting. The phases 1 and 2 cases were then assigned to a discovery class, and risk factor associations were tested in a case-control context. To address this design difference between the discovery and validation stages, a sensitivity analysis is conducted to assess the presumption that similar rankings

Figure 4.6: Odds ratios and 95% confidence intervals from multivariable binary logistic regression with offset term incorporated, in the validation data. Additionally adjusted for study phase. *P*-values are tests for etiologic heterogeneity from a multivariable polytomous logistic regression model. OR = odds ratio; CI = confidence interval.

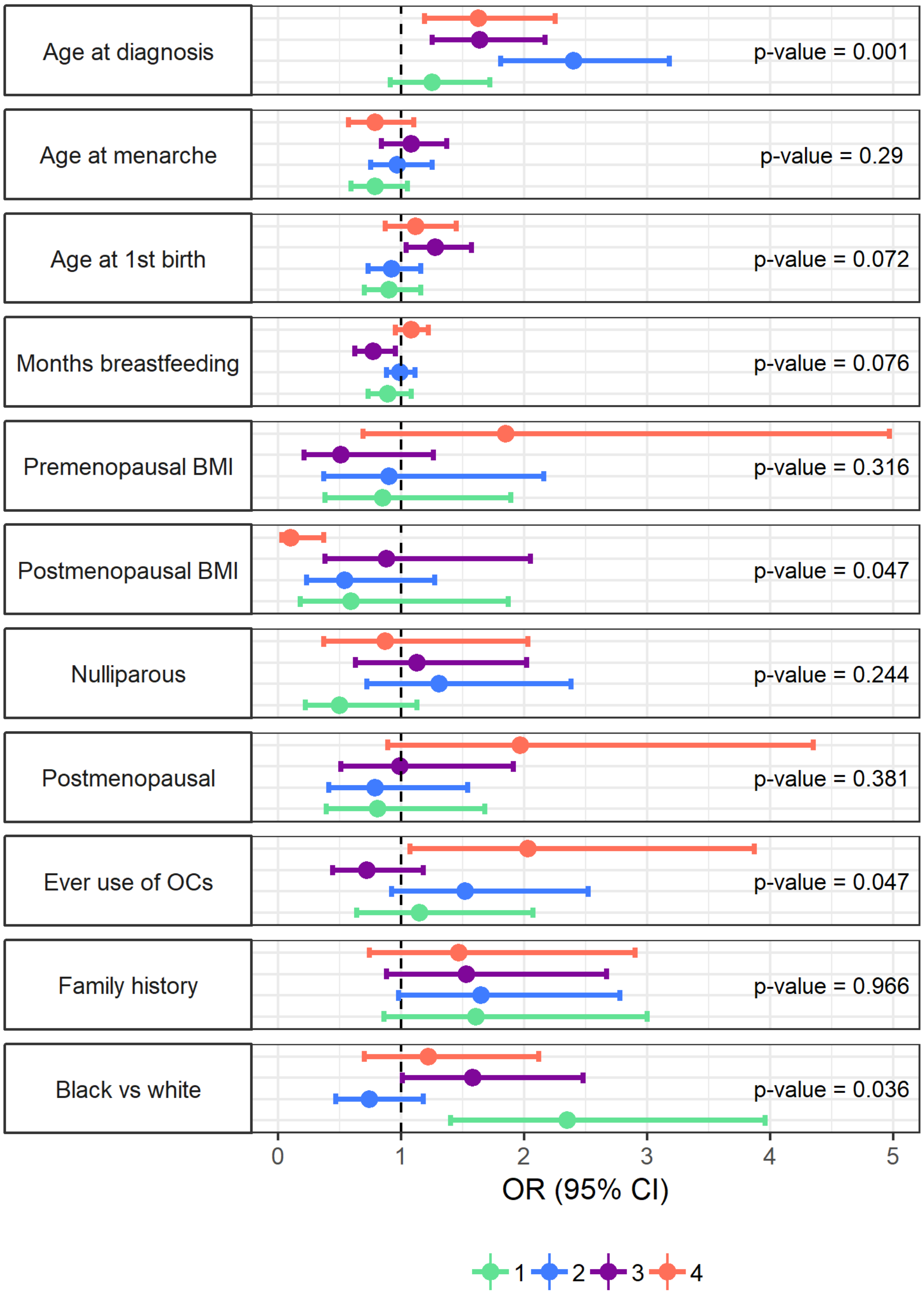


Figure 4.7: Heatmap of expression values for the selected 40 genes according to the optimal 4-class solution in the validation data.

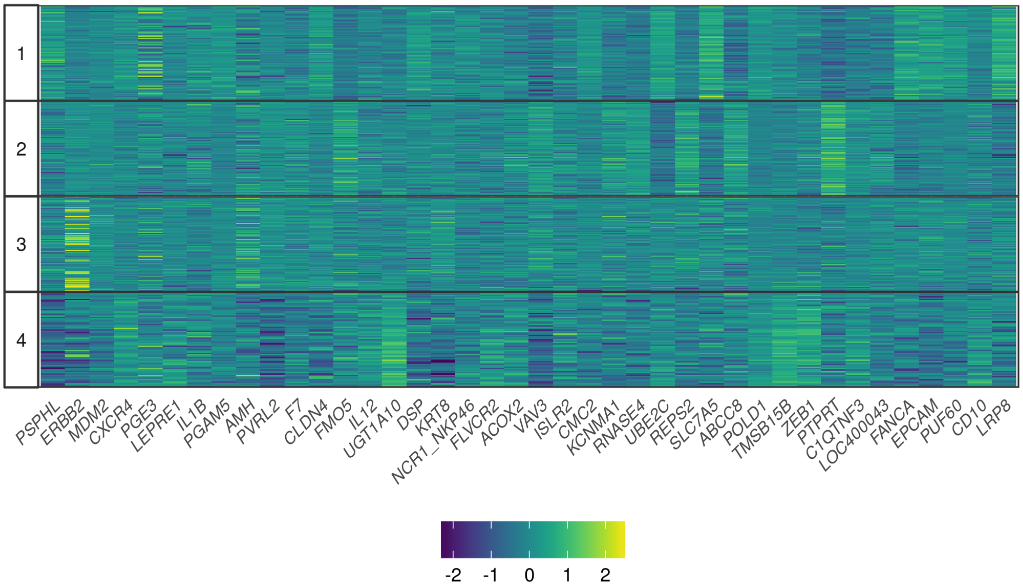


Table 4.9: Comparison of rankings according to case-only D^* in the phase 3 cases, and case-control D in the phase 3 cases and a random sample of phases 1 and 2 controls.

Gene	D^* rank	D rank
PSPHL	1	1
UBE2C	2	2
FANCA	3	4
SLC7A5	4	3
CDC20	5	6
CMC2	6	5
ESR1	7	8
CENPN	8	7
MYBL2	9	9
RNASE4	10	11

would be obtained in a case-only or case-control setting. First, rankings are compared based on individual gene D^* values calculated in the phase 3 cases only versus individual gene D values calculated in the phase 3 cases and a random sample of the phases 1 and 2 control subjects. All 40 of the top 40 genes overlap according to the two rankings, and more specifically, the top-ranked gene is *PSPHL* according to both rankings and 9 of the top 10 genes are the same between the two rankings (Table 4.9). Next, rankings are compared based on individual gene D values calculated in the phases 1 and 2 cases and controls versus individual gene D^* values calculated in the phases 1 and 2 cases only. Thirty-seven of the top 40 genes overlap according to the two rankings, and more specifically, *PSPHL* is again the top-ranked gene according to both rankings, and in both rankings *SLC7A5* is the 2nd ranked gene and *ESR1* is the 3rd ranked gene, and 8 of the top 10 genes are the same according to the two rankings (Table 4.10). These results support the use of D^* in the discovery stage and D in the validation stage, and suggest that the obtained rankings would be similar even if a case-only approach or case-control approach had been consistently used across the two stages of analysis.

4.3.4 Sensitivity to gene selection adjusting for correlation

As some of the genes included in the CBCS codeset were known to be highly correlated, an adjustment for correlation was used when ranking the genes in the primary analysis. A second sensitivity analysis was conducted to assess the impact of accounting for correlation among genes when selecting the top-ranked genes for inclusion in clustering. Table 4.11 shows that the optimal 4-class solution in the validation data based on clustering the top 40 genes ranked by individual D^* ignoring correlation has 62% alignment with the optimal 4-class solution in the validation data based on clustering the top 40 genes after adjusting

Table 4.10: Comparison of rankings according to case-control D in the phases 1 and 2 cases and controls, and case-only D^* in the phases 1 and 2 cases.

Gene	D rank	D^* rank
PSPHL	1	1
SLC7A5	2	2
ESR1	3	3
FOXA1	4	5
PGR	5	4
REEP6	6	6
MAPT	7	17
TMEM158	8	8
UCHL1	9	10
BIRC5	10	11

individual D^* values for correlation. Encouragingly, the validation D of 0.271 for the optimal 4-class solution from the primary results surpasses the D of 0.221 for the optimal 4-class solution based on clustering the top 40 genes selected without adjustment for correlation, suggesting that more risk heterogeneity signal is picked up by selecting a more independent set of genes for inclusion in clustering.

Table 4.11: Cross-tabulation of validation optimal 4-class solutions when correlation is considered or not in selecting the top 40 genes.

	Optimal 4-class			
Ignoring correlation	1	2	3	4
1	**66**	0	3	2
2	0	**83**	15	17
3	1	27	**53**	23
4	17	0	36	**27**

4.3.5 Sensitivity to level of variable selection

In the primary analysis a set of 40 genes, representing the top 10% of genes, were selected for inclusion in the clustering. Next a sensitivity analysis to examine the impact of the extent of variable selection on the results is conducted, as described in Section 4.2.6. As shown in Figure 4.8, there are a number of 4-class solutions with higher D values than the primary 4-class solution based on 40 genes, and the maximum D value of 0.395 is achieved when only the top 12 genes are included. When the top 40 genes are identified separately in the phase 3 discovery data and the phases 1 and 2 validation data, only 9 genes (22.5%) are common between the two lists (Table 4.12). Reassuringly, *PSPHL* is the top-ranked gene when ranking is done independently in the discovery and validation data. *PGE3* is the only other gene included in the top 10 according to both rankings. Figure 4.9 shows the proportion of aligned cases according to class labels assigned based on the optimal discovery 4-class solution and class labels based on the optimal 4-class solution identified independently in the validation cases, as the number of included genes is reduced. The proportion of aligned cases increases as the number of included genes is reduced, such that 85% of cases are classified similarly with an 8-gene or 6-gene solution, 91% of cases are classified similarly with a 4-gene solution, and 95% of cases are classified similarly with a 1-gene solution. These results combine to suggest that perhaps more stringent variable selection, to just the top 8 or 6 genes, would lead to a more etiologically heterogeneous solution with high alignment between two independent class labels, suggesting stability in the clustering based on this small number of genes.

Figure 4.8: D values in the validation data when different numbers of genes are included in identification of the optimal 4-class solution.

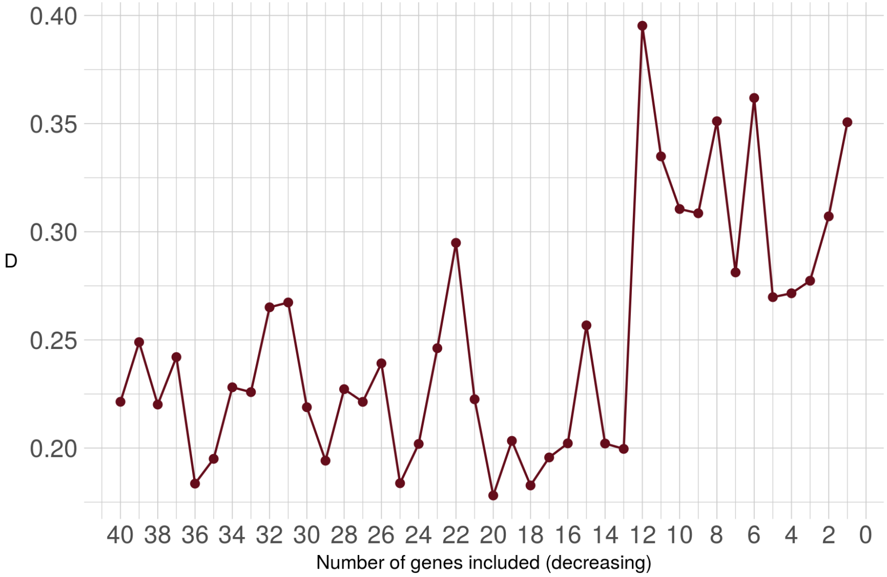


Figure 4.9: Alignment between optimal class labels identified independently in the phase 3 and phases 1 and 2 data, as the number of included genes is reduced.

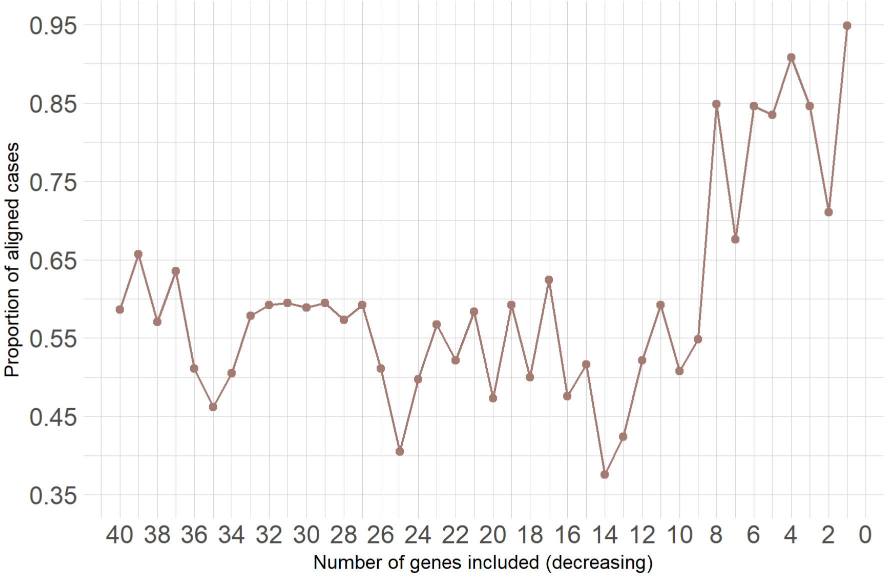


Table 4.12: Similar genes according to discovery and validation rankings.

Gene	Discovery D^*	Discovery rank	Validation D	Validation rank
PSPHL	0.285	1	0.228	1
PGE3	0.039	5	0.076	8
F7	0.060	11	0.063	16
FMO5	0.109	13	0.096	32
DSP	0.048	16	0.076	37
NCR1_NKP46	0.023	18	0.052	20
REPS2	0.073	27	0.088	38
SLC7A5	0.136	28	0.137	18
LOC400043	0.087	35	0.076	17

4.3.6 Comparison of gene rankings in the CBCS and CASH studies

A final sensitivity analysis compares gene rankings according to CBCS phases 1 and 2 case-control rankings and CASH case-control rankings (Table 4.13). Of the 38 common genes between the CBCS and CASH studies, several are consistently highly ranked, including *SLC7A5*, *ESR1*, *PGR*, *IL6ST*, *AR*, *GATA3*, and *BCL2*. The Spearman correlation between the two rankings was 0.449, a moderate correlation that differed significantly from zero (p-value = 0.005). Interestingly only *SLC7A5* remained in the top 40 selected genes in the CBCS phase 3 discovery rankings after adjustment for correlation, as the other genes common between the two studies were all highly correlated with the top-ranked gene, *PSPHL*.

4.4 Discussion

In this data application a novel clustering strategy was used to identify a set of breast cancer subtypes that have clearly distinctive etiology that surpasses that of traditional molecular subtypes in breast cancer. Etiologic differences in these data appeared to be driven by the

Table 4.13: Comparison of gene rankings for the 38 common genes in the CBCS phases 1 and 2 case-control study and the CASH case-control study.

Gene	CBCS phase 1/2	CASH
SLC7A5	1	8
ESR1	2	2
PGR	3	7
IL6ST	4	3
AR	5	13
STC2	6	19
MKI67	7	12
GATA3	8	1
BCL2	9	4
SCGB1D2	10	15
TOP2A	11	23
PTEN	12	27
CCNE1	13	14
TFF3	14	5
IL6	15	25
VEGFA	16	35
EGFR	17	38
CYP19A1	18	30
CCNA2	19	16
BRCA1	20	18
BAG1	21	33
UGT1A10	22	29
KIT	23	26
CDKN1A	24	21
CLDN7	25	20
CDH1	26	36
RAD50	27	10
FANCA	28	34
UGT2B7	29	32
SULT2A1	30	6
ERBB2	31	28
KRT19	32	9
PTGS2	33	37
CCND1	34	11
SULT1E1	35	24
MUC1	36	17
UGT1A4	37	31
RAD17	38	22

PSPHL gene, which was consistently ranked as the top gene across multiple approaches to ranking, and demonstrated a high level of etiologic heterogeneity when used as an individual gene. *PSPHL* is known to be associated with race, such that black women have higher expression of *PSPHL* as compared to white women (Parada *et al.*, 2017; Costantino *et al.*, 2016; Field *et al.*, 2012), and there were clear differences in *PSPHL* across subtypes in this study. *ESR1* and *PGR* were two of the genes identified in the top 10 of the common genes between the CBCS and CASH studies, and these encode estrogen receptor (ER) and progesterone receptor (PR), which are both known to play a role in breast cancer risk and in determining subtypes of breast cancer. ER and PR help form the traditional IHC 4-class subtype solution and similarly *ESR1* and *PGR* are part of the PAM50 gene expression panel that forms the basis of the traditional PAM50 4-class subtype solution. Similarly, *ERBB2* is the gene that encodes human epidermal growth factor receptor 2 (HER2), and was the 2nd-ranked gene in this study after adjustment for correlation. HER2 is also involved in formation of the traditional IHC 4-class subtype solution and, correspondingly, *ERBB2* is included in the PAM50 gene expression panel.

One gene that is thought to play an important role in breast cancer but was not considered in the primary results of this analysis is *TP53.* While not emphasized in the primary results, in the validation cases the 2-class solution based on a *TP53* multigene signature has a D value of 0.029. This would lead to a ranking of 345/407 individual genes, when no adjustment is made for correlation. Because RNA expression of the *TP53* gene is not believed to capture the biologic mechanism through which *TP53* acts in breast cancer, it is therefore not recommended for analysis as a single gene in a study of this type. Instead, a multigene signature for *TP53* is used, and each case is assigned to a “mutant” or “wild-type” class based on a principal components analysis of the multigene signature. While this multigene signature for *TP53* does not appear to contribute strongly to distinguishing etiologically heterogeneous subtypes when considered alone in these data, it is possible that it could be important when considered together with other genes, and future work should determine how best to incorporate this type of information that is on a different scale (i.e. binary versus continuous) into the novel clustering strategy.

Risk factors that contributed the most to distinguishing the optimal 4-class solution included age at diagnosis, postmenopausal BMI, ever use of oral contraceptives, and race. In a pooled analysis across multiple cohort studies, Gaudet *et al.* (2018) found that parity, age at first live birth, years between menarche and first birth among parous women, age at menopause, and first degree family history of breast cancer were risk factors that demonstrated etiologic heterogeneity according to the traditional IHC 4-class subtypes. Race was not included in their analysis. Interestingly, none of these risk factors are the ones that demonstrated etiologic heterogeneity according to the optimal 4-class solution in this analysis. This could be an artifact of the somewhat artificial risk factor distributions in the CBCS study population, induced as a result of the study design that oversampled black women and young women. Alternatively, the differences could be caused by shifting risk factor distributions over time, especially with respect to hormonal risk factors, and changes in the way immunohistochemical markers have been categorized as positive or negative.

There are some remaining methodologic challenges. The first challenge is how to determine the number of tumor markers to include in the clustering analysis. In this application the top 10% were selected based on their individual measures of etiologic heterogeneity, and thus 40 genes were clustered. However, sensitivity analyses indicated that solutions based on even fewer genes led to higher measures of explained variation, as well as better alignment between optimal class solutions identified in two independent datasets. This suggests that the ideal solution in these data may be based on only 8 or 6 genes, rather than the 40 used in the primary analysis. A more objective approach to determining a cutoff for inclusion is needed to determine the appropriate level of upfront variable selection. A second challenge is how to select the ideal number of subtypes. In any unsupervised clustering analysis, the number of classes must be pre-specified, though the true number of underlying classes of interest is unknown. In this data example 2-class, 3-class, 4-class, and 5-class solutions were compared. Solutions with more than five classes were not considered to avoid model overfitting given the number of available cases. However the approach used to select the 4-class solution as ideal, based on the alignment between solutions from two independent datasets as well as permutation tests for significant differences in the amount of explained variation, was somewhat arbitrary. Additionally, the 4-class solution was selected in part based on the knowledge that there are existing 4-class breast cancer subtyping schemes, and so using four classes would make comparisons with these other classification systems easier. More rigorous and objective methods to select the ideal number of classes in an analysis of etiologic heterogeneity are needed, and this will be an area of future work.

Overall, this data application demonstrates that when the proposed novel clustering strategy is used, which combines a search for candidate subtypes with a measure of etiologic heterogeneity based on the available risk factor data, subtype solutions with higher levels of etiologic heterogeneity can be discovered. This is important for epidemiologists who seek to identify solutions that maximize risk heterogeneity across subtypes. While components of this analysis are still subjective, this is in line with the real challenges faced when conducting complex real-world epidemiologic studies, and is therefore not a major limitation.

4.5 Gene expression panel

ABAT
ABCB1
ABCC8
ACADSB
ACTG1P3
ADCY1
AKR7L
ALDH1A1
APBB2
AQP5
AR
AURKB
BLK
BLR1_CXCR5
BMP2
BOP1
BTG3
BUB1
C10ORF116
C11ORF75
C14ORF45
C16ORF45
C1QTNF3
C1ORF106
C2ORF27A
C4A
C4ORF31
C8ORF33
C9ORF98
CACNB3
CALCP
CAPN13
CAPN9
CASKIN1
CCDC103
CCL7
CCNB2
CCR3
CD10
CD19
F7
FAM54A
FAM63A
FAM64A
FANCA
FBXL6
FCRL2
FLJ20152
FMNL2
FMO5
FN1
FOXC2
FOXP3
FPRL1
FSCN1
FUT8
GALT
GCNT2
GFRA1
GPR44
GTSE1
GUCA1
GZMM
HGH1
HJURP
HLA_DOB
HPN
HRC
ICOS
IDO1
IFRD1
IGF2BP2
IGF2BP3
IL12
IL1B
IL2RB
IL5RA
IL6
IL6ST
IL8RA
OCLN
PCSK6
PD_L1
PDCD1
PDSS1
PDZK1
PFKP
PGAM5
PGE3
PINK1
PKIB
PLK1
PPBP
PRF1
PRRG2
PRRT2
PTDSS1
PTGER3
PTGS2
PTPRT
PUF60
PVRL2
RAD54L
RAI2
RBM24
REEP6
REPS2
RIMS4
RNASE4
RPS6KB2
RSPH1
S100A8
SCGB1D2
SCUBE2
SDCBP
SEC14L2
SEMA3B
SERPINB5_MASPIN
SH2D1A
SHCBP1
ADM
AMH
AMHR2
ANGPTL4
ANLN
APH1B
ATAD2
AURKA
AXL
BAG1
BCL2
BIRC5
BLVRA
BRCA1
BTG2
CAV1
MIS18A_C21ORF45
CCNB1
CCND1
CCNE1
CD24
CDC20
CDC25B
CDC25C
CDC6
CDCA7L
CDH3
CDK1
CDKN1A
CDKN3
CENPF
CEP55
CKS1B
CLDN3
CLDN4
CLDN7
CRMP1
CRYAB
CRYBB2
CXXC5
KRT14
KRT17
KRT19
KRT5
KRT8
LEPRE1
LHFP
LOC400043
MAD2L1
MAP2K4
MAPT
MCM3
MDM2
MELK
MET
MIA
CCNA2
MKI67
MLPH
MMP11
MPP1
MSH3
MUC1
MYBL2
MYC
NAT1
NCAPH2
NDC80
NDRG1
NEO1
NPEPPS
NT5E
NUDT1
NUF2
ORC6L
PGR
PHGDH
PIK3CA
PLOD1
PNP

CD2
CD246_CD3Z
CD28
CD3E
CD3G
CD4
CD6
CD68
CD84
CD8A
CD96
CDC45
CDCA5
CDCA7
CDCA8
CDH1
CELSR1
CENPA
CENPN
CMC2
CMYA5
CTSL2
CXCL13
CXCL5
CXCR4
CYBB
CYP19A1
CYP27A1
CYP2D6
CYP3A4
CYP3A5
CYP4B1
CYP7B1
DEPDC1
DLGAP5
DNM2
DOCK3
DTX3
ECE2
EFHD1
ELOVL2
EZH2
IL8RB
INPP4B
IRS1
ISLR2
ITGB5
KCNMA1
KCNN4
KDM4B
KIAA0125
KIF3A
KLHDC9
KLHL7
LAG_3
LCK
LILRB2
LOX
LRG1
LRP8
LRRC50
MAF
MAGED2
MAGI2
MARVELD2
MCM10
MMP1
MMP2
MMP3
MND1
MRPS17
MS4A1
MSR1
MYB
NCAPG
NCR1_NKP46
NCS1
NFKB1
NLN
NME5
NR1H3
NTN4
NXNL2
NXPH4
SHROOM3
SIRPG
SKAP1
SLC1A2
SLC52A2
SLC7A5
SNAI1
SNAI2
SNRPD1
SOX10
STC2
SULT1E1
SULT2A1
SYBU
SYT1
TBC1D9
TFF3
TIM_3
TMSB15B
TNFRSF17
TPX2
TRAF1
TRAT1
TRPC1
TRPM7
TWIEST2
TWIST1
UGT1A10
UGT1A4
UGT1A8
UGT2B7
VAV3
WDR12
WDR19
XBP1
XCL1
ZAP70
ZEB2
ZG16B
ACOX2
ACTR3B
ADHFE1
DAPK1
DDB2
DDIT4
DDR1
DSP
EGFR
EMP3
EPCAM
ERBB2
ERBB3
ERBB4
ESR1
ESRP1
EVI2A
EXO1
F11R
FABP5
FAM177A1
FAM198B
FAM214A_KIAA1370
FBN1
FGFR4
FLVCR2
FNBP1
FOXA1
FOXC1
FOXM1
GAL
GATA3
GGH
GNG11
GPR160
GRB7
GRHL2
GSTP1
GSTT2
JUP
KIAA0040
KIF23
KIF2C
KIFC1
KIT
POLD1
PREP
PSPH
PSPHL
PTEN
PTTG1
PVRL3
RAB25
RAD17
RAD50
RB1
RFC4
RNF103
RRAGD
RRM2
SFRP1
SH2B3
SLC16A3
SLC39A6
SPINT1
SPINT2
SQLE
STK38
TCEAL1
TMEM158
TMEM45B
TNIK
TOP2A
TRIP13
TUBA4A
TYMS
UBE2C
UBE2T
UCHL1
ULK1
VEGFA
VIM
ZEB1

Chapter 5

Conclusion

This dissertation investigated statistical methods related to the study of etiologic heterogeneity. Disease subtyping is increasing in importance, especially in cancer research, due to the rising use of molecular and genomic profiling as part of standard patient care. As a result, statistical methods are needed to identify risk factors that have a differential effect across subtypes of disease, when subtypes may be formed from high dimensional disease characteristic data.

After reviewing existing methods for the study of etiologic heterogeneity, regression-based methods that rely on pre-specified subtypes of disease were compared, including polytomous logistic regression, the two-stage meta-regression method of Wang *et al.* (2015), the two-stage regression with simultaneous estimation method of Chatterjee (2004), and the stratified logistic regression method of Rosner *et al.* (2013). The primary challenge to this was unifying the notation of the various methods so that the similarities and differences could be examined. After doing so it became clear that the methods can all estimate similar parameters $\{\beta_{pm}\}$ to address the question of whether a risk factor of interest has the same effect across all subtypes of disease, and similar parameters $\{\gamma_{pk}\}$ to address the question of whether risk factor effects differ across levels of each individual disease characteristic by which the subtypes are defined. A simplified data example showed that the methods result in similar parameter estimates and conclusions, and simulation studies found that while the stratified logistic regression method of Rosner *et al.* (2013) results in substantial bias in parameter estimation for addressing whether risk factor effects differ across levels of the disease subtype, all methods have similar power to address both questions of interest. These results indicate that polytomous logistic regression, which is easy to implement with standard software, performs at least as well as more complex methods and therefore is an acceptable approach to the study of etiologic heterogeneity when data arise from a case-control study. These results can serve to guide epidemiologists and other researchers seeking to study etiologic heterogeneity in selection of an appropriate statistical method.

Next, the statistical properties of a novel clustering method were examined. Optimal D clustering seeks to identify, from high dimensional disease characteristic data, the subtypes that maximize etiologic heterogeneity. The method is conducted in two stages. In the first stage, the disease characteristic data are clustered using unsupervised k-means clustering with many random starts so that a variety of candidate sets of subtype solutions are found. Then for each candidate solution, a scalar measure of etiologic heterogeneity, denoted D, is calculated based on risk predictions from a polytomous logistic regression model with the candidate class solution as the outcome and the known risk factors as predictors. The subtype solution that maximizes D is selected as the optimal class solution. This approach had been used previously in several applications, including to breast cancer (Begg *et al.*, 2015), melanoma (Mauguen *et al.*, 2017), and kidney cancer (Begg *et al.*, 2014), but this was the first time the statistical properties had been evaluated in detail. Simplified simulation studies found that the method cannot identify the truly etiologically heterogeneous subtype solution when the strength of counterfeit structure surpasses the strength of the truly etiologically heterogeneous structure, or when the number of disease characteristics representing noise is very large. However, this can be overcome with up-front reduction of the set of disease characteristics included in the clustering, selecting the subset of characteristics that show strong heterogeneity signals individually, after which the etiologically distinct subtypes can successfully be identified with high probability.

Finally, the optimal D clustering approach was applied to data from the Carolina Breast Cancer study. The available gene expression data was reduced up-front based on individual gene D values, and only the top 10% of genes according to their individual contributions to risk heterogeneity were included in the clustering, after an adjustment for correlation among genes was applied. A 4-class solution was identified, which contained disease subtypes that are significantly different with respect to the effects of age at diagnosis, postmenopausal BMI, ever use of oral contraceptives, and race. *PSPHL*, a gene with known race associations (Parada *et al.*, 2017), was the gene that was most significant in distinguishing these subtypes under a variety of approaches to gene ranking. *ERBB2*, the gene that encodes HER2, was the second ranked gene in the discovery stage of the analysis. HER2 is known to play a role in subtyping breast cancer. The optimal 4-class solution identified in this application demonstrated a much larger degree of etiologic heterogeneity, as quantified by D, as compared to that seen in the traditional IHC 4-class solution or the traditional PAM50 4-class solution. This result indicates that optimal D clustering can identify more heterogeneous class solutions than existing breast cancer subtypes.

While this work has contributed to understanding the appropriate uses of available statistical methods for the study of etiologic heterogeneity, there are some needed areas of future work. The first relates to the extent of up-front disease characteristic selection.

In the simulation studies disease characteristics were included based on a cut-off value for permutation-based p-values, and in the data application genes were included based on a fixed percentage of top-ranked genes. However, sensitivity analysis in the data application revealed that solutions demonstrating even greater levels of etiologic heterogeneity could be found after more stringent reduction of the gene set. An objective approach to selection of disease characteristics for inclusion in the clustering stage of the analysis is needed. The second area of future work relates to identification of the ideal number of classes. In any unsupervised clustering analysis, the number of classes must be pre-specified. However, the true number of etiologically distinct subtypes is not known in real-world applications. Statistical methods exist to identify the ideal number of classes in traditional clustering analyses, but they are not tailored to this application. An approach that considers both the distance between clusters as well as information from the risk factors is needed. Finally, an R package, referenced throughout this dissertation, is in development to make calculation of D and heterogeneity p-values broadly accessible, and will be freely available for public use.

Bibliography

S. A. Ahrendt, P. A. Decker, E. A. Alawi, Y. R. Zhu Yr, M. Sanchez-Cespedes, S. C. Yang, G. B. Haasler, A. Kajdacsy-Balla, M. J. Demeure, and D. Sidransky. Cigarette smoking is strongly associated with mutation of the k-ras gene in patients with primary adenocarcinoma of the lung. *Cancer*, 92(6):1525–30, 2001.

J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59(1):19–35, 1972.

C. B. Begg and E. C. Zabor. Detecting and exploiting etiologic heterogeneity in epidemiologic studies. *Am J Epidemiol*, 176(6):512–8, 2012.

C. B. Begg and Z. F. Zhang. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prev*, 3(2):173–5, 1994.

C. B. Begg, E. C. Zabor, J. L. Bernstein, L. Bernstein, M. F. Press, and V. E. Seshan. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med*, 32(29):5039–52, 2013.

C. B. Begg, V. E. Seshan, E. C. Zabor, H. Furberg, A. Arora, R. Shen, J. K. Maranchie, M. E. Nielsen, W. K. Rathmell, S. Signoretti, P. Tamboli, J. A. Karam, T. K. Choueiri, A. A. Hakimi, and J. J. Hsieh. Genomic investigation of etiologic heterogeneity: methodologic challenges. *BMC Med Res Methodol*, 14:138, 2014.

C. B. Begg, I. Orlow, E. C. Zabor, A. Arora, A. Sharma, V. E. Seshan, and J. L. Bernstein. Identifying etiologically distinct sub-types of cancer: A demonstration project involving breast cancer. *Cancer Med*, 4(9):1432–9, 2015.

N. E. Breslow and N. E. Day. Statistical methods in cancer research. volume i - the analysis of case-control studies. *IARC scientific publications*, (32):5–338, 1980.

L. A. Brinton, A. S. Felix, D. S. McMeekin, W. T. Creasman, M. E. Sherman, D. Mutch, D. E. Cohn, J. L. Walker, R. G. Moore, L. S. Downs, R. A. Soslow, and R. Zaino. Etiologic heterogeneity in endometrial cancer: evidence from a gynecologic oncology group trial. *Gynecol Oncol*, 129(2):277–84, 2013.

N. Chatterjee, S. Sinha, W. R. Diver, and H. S. Feigelson. Analysis of cohort studies with multivariate and partially observed disease classification data. *Biometrika*, 97(3):683–698, 2010.

Nilanjan Chatterjee. A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association*, 99(465):127–138, 2004.

N. S. Costantino, B. Freeman, C. D. Shriver, and R. E. Ellsworth. Outcome disparities in african american compared with european american women with er+her2- tumors treated within an equal-access health care system. *Ethn Dis*, 26(3):407–16, 2016.

N. Dubin and B. S. Pasternack. Risk assessment for case-control subgroups by polychotomous logistic regression. *Am J Epidemiol*, 123(6):1101–17, 1986.

L. A. Field, B. Love, B. Deyarmin, J. A. Hooke, C. D. Shriver, and R. E. Ellsworth.

Identification of differentially expressed genes in breast tumors from african american compared with caucasian women. *Cancer*, 118(5):1334–44, 2012.

Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

H. Furberg, R. C. Millikan, J. Geradts, M. D. Gammon, L. G. Dressler, C. B. Ambrosone, and B. Newman. Environmental factors in relation to breast cancer characterized by p53 protein expression. *Cancer Epidemiol Biomarkers Prev*, 11(9):829–35, 2002.

H. Furberg, R. C. Millikan, J. Geradts, M. D. Gammon, L. G. Dressler, C. B. Ambrosone, and B. Newman. Reproductive factors in relation to breast cancer characterized by p53 protein expression (united states). *Cancer Causes Control*, 14(7):609–18, 2003.

Mitchell H. Gail, Jay H. Lubin, and Lawrence V. Rubinstein. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, 68(3):703–707, 1981.

M. M. Gaudet, M. F. Press, R. W. Haile, C. F. Lynch, S. L. Glaser, J. Schildkraut, M. D. Gammon, W. Douglas Thompson, and J. L. Bernstein. Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger. *Breast Cancer Res Treat*, 130(2):587–97, 2011.

M. M. Gaudet, G. L. Gierach, B. D. Carter, J. Luo, R. L. Milne, E. Weiderpass, G. G. Giles, R. M. Tamimi, A. H. Eliassen, B. Rosner, A. Wolk, H. O. Adami, K. L. Margolis, S. M. Gapstur, M. Garcia-Closas, and L. A. Brinton. Pooled analysis of nine cohorts reveals breast cancer risk factors by tumor molecular subtype. *Cancer Res*, 2018.

Els Goetghebeur and Louise Ryan. Analysis of competing risks survival data when some failure types are missing. *Biometrika*, 82(4):821–833, 1995.

W. Y. Huang, B. Newman, R. C. Millikan, M. J. Schell, B. S. Hulka, and P. G. Moorman. Hormone-related factors and risk of breast cancer in relation to estrogen receptor and progesterone receptor status. *Am J Epidemiol*, 151(7):703–14, 2000.

L. Kaufman and Peter Rousseeuw. *Clustering by means of medoids*, pages 405–416. North-Holland; Amsterdam, 1987.

B. A. Kohler, R. L. Sherman, N. Howlader, A. Jemal, A. B. Ryerson, K. A. Henry, F. P. Boscoe, K. A. Cronin, A. Lake, A. M. Noone, S. J. Henley, C. R. Eheman, R. N. Anderson, and L. Penberthy. Annual report to the nation on the status of cancer, 1975-2011, featuring incidence of breast cancer subtypes by race/ethnicity, poverty, and state. *J Natl Cancer Inst*, 107(6):djv048, 2015.

Lesnoff, M., Lancelot, and R. *aod: Analysis of Overdispersed Data*, 2012. R package version 1.3.

D. Limsui, R. A. Vierkant, L. S. Tillmans, A. H. Wang, D. J. Weisenberger, P. W. Laird, C. F. Lynch, K. E. Anderson, A. J. French, R. W. Haile, L. J. Harnack, J. D. Potter, S. L. Slager, T. C. Smyrk, S. N. Thibodeau, J. R. Cerhan, and P. J. Limburg. Cigarette smoking and colorectal cancer risk by molecularly defined subtypes. *J Natl Cancer Inst*, 102(14):1012–22, 2010.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Proba-*

bility, Volume 1: Statistics, Fifth Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297. University of California Press, 1967.

C. J. Marsit, B. C. Christensen, E. A. Houseman, M. R. Karagas, M. R. Wrensch, R. F. Yeh, H. H. Nelson, J. L. Wiemels, S. Zheng, M. R. Posner, M. D. McClean, J. K. Wiencke, and K. T. Kelsey. Epigenetic profiling reveals etiologically distinct patterns of dna methylation in head and neck squamous cell carcinoma. *Carcinogenesis*, 30(3):416–22, 2009.

A. Mauguen, E. C. Zabor, N. E. Thomas, M. Berwick, V. E. Seshan, and C. B. Begg. Defining cancer subtypes with distinctive etiologic profiles: An application to the epidemiology of melanoma. *J Am Stat Assoc*, 112(517):54–63, 2017.

M. A. Merritt, M. De Pari, A. F. Vitonis, L. J. Titus, D. W. Cramer, and K. L. Terry. Reproductive characteristics in relation to ovarian cancer risk by histologic pathways. *Hum Reprod*, 28(5):1406–17, 2013.

R. C. Millikan, B. Newman, C. K. Tse, P. G. Moorman, K. Conway, L. G. Dressler, L. V. Smith, M. H. Labbok, J. Geradts, J. T. Bensen, S. Jackson, S. Nyante, C. Livasy, L. Carey, H. S. Earp, and C. M. Perou. Epidemiology of basal-like breast cancer. *Breast Cancer Res Treat*, 109(1):123–39, 2008.

B. Newman, P. G. Moorman, R. Millikan, B. F. Qaqish, J. Geradts, T. E. Aldrich, and E. T. Liu. The carolina breast cancer study: integrating population-based epidemiology and molecular biology. *Breast Cancer Res Treat*, 35(1):51–60, 1995.

S. Ogino, A. T. Chan, C. S. Fuchs, and E. Giovannucci. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. *Gut*, 60(3):397–411, 2011.

Jr. Parada, H., X. Sun, J. M. Fleming, C. R. Williams-DeVane, E. L. Kirk, L. T. Olsson, C. M. Perou, A. F. Olshan, and M. A. Troester. Race-associated biological differences among luminal a and basal-like breast cancers in the carolina breast cancer study. *Breast Cancer Res*, 19(1):131, 2017.

C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–52, 2000.

A. I. Phipps, K. E. Malone, P. L. Porter, J. R. Daling, and C. I. Li. Body size and risk of luminal, her2-overexpressing, and triple-negative breast cancer in postmenopausal women. *Cancer Epidemiol Biomarkers Prev*, 17(8):2078–86, 2008.

A. I. Phipps, K. E. Malone, P. L. Porter, J. R. Daling, and C. I. Li. Reproductive and hormonal risk factors for postmenopausal luminal, her-2-overexpressing, and triple-negative breast cancer. *Cancer*, 113(7):1521–6, 2008.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

G. J. Riely, M. G. Kris, D. Rosenbaum, J. Marks, A. Li, D. A. Chitale, K. Nafa, E. R. Riedel, M. Hsu, W. Pao, V. A. Miller, and M. Ladanyi. Frequency and distinctive spectrum of kras mutations in never smokers with lung adenocarcinoma. *Clin Cancer Res*, 14(18):5731–4, 2008.

B. Rosner, R. J. Glynn, R. M. Tamimi, W. Y. Chen, G. A. Colditz, W. C. Willett, and

S. E. Hankinson. Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. *Am J Epidemiol*, 178(2):296–308, 2013.

J. M. Schildkraut, E. S. Iversen, L. Akushevich, R. Whitaker, R. C. Bentley, A. Berchuck, and J. R. Marks. Molecular signatures of epithelial ovarian cancer: analysis of associations with tumor characteristics and epidemiologic risk factors. *Cancer Epidemiol Biomarkers Prev*, 22(10):1709–21, 2013.

G. P. Sfakianos, E. S. Iversen, R. Whitaker, L. Akushevich, J. M. Schildkraut, S. K. Murphy, J. R. Marks, and A. Berchuck. Validation of ovarian cancer gene expression signatures for survival and subtype in formalin fixed paraffin embedded tissues. *Gynecol Oncol*, 129(1):159–64, 2013.

T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–74, 2001.

C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100(18):10393–8, 2003.

J. A. Taylor, M. A. Watson, T. R. Devereux, R. Y. Michels, G. Saccomanno, and M. Anderson. p53 mutation hotspot in radon-associated lung cancer. *Lancet*, 343(8889):86–7, 1994.

Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

Terry M Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–72, 2002.

R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, N. Traficante, S. Fereday, J. A. Hung, Y. E. Chiew, I. Haviv, D. Gertig, A. DeFazio, and D. D. Bowtell. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res*, 14(16):5198–208, 2008.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

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

M. Wang, A. Kuchiba, and S. Ogino. A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. *Am J Epidemiol*, 182(3):263–70, 2015.

X. R. Yang, R. M. Pfeiffer, M. Garcia-Closas, D. L. Rimm, J. Lissowska, L. A. Brinton, B. Peplonska, S. M. Hewitt, R. W. Cartun, D. Mandich, H. Sasano, D. B. Evans, T. R. Sutter, and M. E. Sherman. Hormonal markers in breast cancer: coexpression, relationship with pathologic characteristics, and risk factor associations in a population-based study. *Cancer Res*, 67(21):10608–17, 2007.

K. Yu, H. Zhang, W. Wheeler, H. N. Horne, J. Chen, and J. D. Figueroa. A robust association test for detecting genetic variants with heterogeneous effects. *Biostatistics*, 16(1):5–16, 2015.

E. C. Zabor and C. B. Begg. A comparison of statistical methods for the study of etiologic heterogeneity. *Stat Med*, 36(25):4050–4060, 2017.