Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

A test for detecting etiologic heterogeneity in epidemiological studies

S. Karagulle & Z. Kalaylioglu

To cite this article: S. Karagulle & Z. Kalaylioglu (2016) A test for detecting etiologic heterogeneity in epidemiological studies, Journal of Applied Statistics, 43:3, 538-549, DOI: 10.1080/02664763.2015.1070808

To link to this article: https://doi.org/10.1080/02664763.2015.1070808

Journal of Applied Statistics, 2016
Vol. 43, No. 3, 538–549, http://dx.doi.org/10.1080/02664763.2015.1070808

Taylor & Francis
Taylor & Francis Group

A test for detecting etiologic heterogeneity in epidemiological studies

S. Karagulle and Z. Kalaylioglu*

Department of Statistics, Middle East Technical University, 06800 Ankara, Turkey

(*Received 3 September 2014; accepted 6 July 2015*)

Current statistical methods for analyzing epidemiological data with disease subtype information allow us to acquire knowledge not only for risk factor-disease subtype association but also, on a more profound account, heterogeneity in these associations by multiple disease characteristics (so-called etiologic heterogeneity of the disease). Current interest, particularly in cancer epidemiology, lies in obtaining a valid *p*-value for testing the hypothesis whether a particular cancer is etiologically heterogeneous. We consider the two-stage logistic regression model along with pseudo-conditional likelihood estimation method and design a testing strategy based on Rao's score test. An extensive Monte Carlo simulation study is carried out, false discovery rate and statistical power of the suggested test are investigated. Simulation results indicate that applying the proposed testing strategy, even a small degree of true etiologic heterogeneity can be recovered with a large statistical power from the sampled data. The strategy is then applied on a breast cancer data set to illustrate its use in practice where there are multiple risk factors and multiple disease characteristics of simultaneous concern.

Keywords: cancer; disease subtype; odds ratio; polychotomous logistic regression; risk factor; score test

1. Introduction

One central goal of epidemiological studies of cancer is to study the etiologic heterogeneity in its subtypes. Etiologic heterogeneity in the subtypes is explained as follows. Consider breast cancer which is characterized by its histological type (ductal/lobular/tubular/mixed carcinoma), tumor size (≤ 2 cm, 2 cm), tumor grade (1/2/3), nodal status (+/−), estrogen receptor (ER) status (+/−), and progesterone receptor (PR) status (+/−). Subtypes are the breast cancer classifications according to these characteristics. The subtypes are etiologically heterogeneous, if the effect of exposures are different for different subtypes. Etiologic heterogeneity of breast cancer, in particular, has been under investigation, see [11–13,18,19,29,30] for recent findings. Such studies have also long been concerned with other types of cancer including ovarian cancer, colorectal cancer, and non-Hodgkins lymphoma as in [20,22, 23,28].

*Corresponding author. Email: kzeynep@metu.edu.tr

In epidemiological case-control studies, statistical approaches to study etiologic heterogeneity of the disease, where reference is the disease-free situation represented by the study controls, commonly consist of (i) logistic regression analysis for each disease characteristic, (ii) polychotomous logistic regression analysis for a categorical variable whose levels are the disease subtypes. In the first approach, etiological heterogeneity in the disease subtypes is studied by means of heterogeneity in disease's defining characteristics in terms of their relation with the risk factors. This approach estimates the odds ratios for each disease characteristic separately ignoring the relation between them. This in fact hampers direct inference on etiological heterogeneity with respect to disease subtypes. On the other hand, the latter approach provides direct inference on etiological heterogeneity with respect to disease subtypes as desired. However, it suffers from high-dimension problem for diseases with large number of subtypes, for example, number of breast cancer subtypes constructed by grouping the disease according to the characteristics described in the first paragraph is $4 \times 2 \times 3 \times 2 \times 2 \times 2 = 192$ leading to a polychotomous logistic regression with 384 parameters (including the intercepts) for a single explanatory variable in the model. Obviously, the dimension of the parameter space will increase substantially with the inclusion of all risk factors of interest. In such situations, not only does the analysis suffer from computational burden, but it also results in decreased statistical power and inflated bias. The bias increases with increasing p/n ratio, where p is the total number of regression coefficients and n is the number of study patients [5]. Cordeiro and McCullagh [7] show for logistic models that the bias is equal to $p\beta/n$, where β is the vector of unknown regression coefficients. Also, this approach appears to be inefficient when odds ratios by disease characteristics is desired in addition to the odds ratios by disease subtypes. Challenges encountered in the statistical analysis of epidemiological data concerned with cancer subtypes are reviewed by Troester and Swift-Scanlan [26]. Power analysis of a subtype based approach for etiologic heterogeneity is given in [1].

Chatterjee [6] developed a two-stage logistic regression (TS-LR) addressing the issues described above. This model tames the high-dimensionality issue of approach (ii) (e.g. for the example above, number of association coefficients is 192 in approach (ii) versus 10 if the two-stage approach was used), readily estimates etiologic heterogeneity in each disease characteristic adjusted for the others, and provides inference on etiologic heterogeneity in terms of disease subtypes. The method has been widely used to understand etiologic heterogeneity of colorectal cancer [3,15–17,21] and breast cancer [10,24,27]. As the emphasis on etiologic heterogeneity increases in cancer epidemiology studies, the need for formal statistical tests for it emerges. Testing etiological heterogeneity can be accomplished using the TS-LR model given its advantages summarized above and detailed in [6]. In the current article, a strategy based on score test (ST) statistic is developed to test etiological heterogeneity with respect to the disease defining characteristics. The main advantages of Rao's ST are that it is invariant to different formulations of the null hypothesis and requires maximum likelihood estimation on a lower dimension parameter space [2,4]. Statistically significant etiologic heterogeneity in terms of the disease defining characteristics indicates etiologic heterogeneity between the disease subtypes with respect to the corresponding characteristics. Detecting etiologic heterogeneity in disease subtypes ultimately leads to a better understanding of risk profiles by subtypes (1) and improves effective treatment options.

In this article, we investigate the statistical properties of the ST in TS-LR and illustrate its use. We hope that the results of our simulation study ultimately guide the researcher in determining the sufficient number of study cases from each subtype and use a powerful test for detecting etiological heterogeneity. Next section reviews the TS-LR model while articulating additional interpretations and lays out the testing strategy devised. Section 3 gives the results of the simulation experiment carried out to assess the performance of the testing strategy in terms of its type I error and power. Section 4 employs the devised strategy to test for etiologic heterogeneity

in breast cancer for Turkish patients and also shows how it is used in the presence of multiple covariates. A brief discussion is given in the last section.

2. Notation and testing strategy

2.1 *Review of TS-LR*

The TS-LR model is a hierarchical model consisting of the following two stages:

$$P(Y_i = m \mid \boldsymbol{X}_i) = \frac{\exp(\alpha_m + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_m)}{1 + \sum_{m=1}^{M} \exp(\alpha_m + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_m)}, \tag{1}$$

$$\beta_m = \{\beta_{i_1 i_2 \cdots i_K}\} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})}, \tag{2}$$

where $i = 1, 2, \ldots, N$ indexing the study subjects and $m = 1, 2, \ldots, M$ indexing the disease subtypes, N and M are the number of study subjects and disease subtypes, respectively, α_m $(m = 1, 2, \ldots, M)$ determine the baseline prevalence of the different disease subtypes, and β_m denotes the $P \times 1$ vector of regression coefficients associated with P covariates X_i. Below, we expound on the model basics and advise the reader to refer to [6], for further elaborations. For the time being, assume a single covariate for ease in illustrating the idea. First stage is the response model in which multi-level categorical disease subtype variable, Y_i, is modeled using a standard logistic regression given the covariate X_i. In model (1), $Y_i = m$ means that disease subtype of subject i is the one coded by m, while $m = 0$ refers to the disease-free category. Second stage consists of models for the first-stage regression coefficients (i.e. for the β coefficients). In the current article, we concentrate on the diseases in which etiologic heterogeneity with respect to one characteristic does not depend on the others and thus β_m $(m = 1, \ldots, M)$ is modeled as in model (2). Otherwise, interaction terms are included in this model but this is a concern of another paper. Each β_m can also be denoted by $\beta_{i_1 i_2 \cdots i_K}$, where K is the number of categorical disease defining characteristics and each of i_k, $k = 1, 2, \ldots, K$ indexes the specific level of each characteristic. Model (2) is a design-like deterministic model consisting of a linear combination of contrast parameters denoted by θs. Each $\theta^{(1)}_{\cdot}$ parameter in the model is a first-order contrast parameter. Here, $\theta^{(1)}_{k_1(i_{k_1})}$ is the logOR associated with having i_{k_1}th level of the characteristic k_1 relative to the reference level of the same characteristic for a one unit change in the particular covariate. To illustrate simply, suppose a cancer described by the following two characteristics; tumor size (small(0)/medium(1)/large(2)) and nodal status (yes(1)/no(0)). Cross-classifying the levels of the two characteristics results in a six-level subtype with levels being (small,yes), (small,no), (medium,yes), (medium,no), (large,yes), and (large,no) of which the reference level is (small,no). Assuming a single risk factor, first-stage slope parameters are $\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}, \beta_{20}$,

Table 1. Reparameterization of the first stage slopes.

m	β	Tumor size	Nodal status	Second stage model
1	$\beta_1 = \beta_{00}$	small(=0)	no(=0)	$\theta^{(0)} + \theta^{(1)}_{1(1)} + \theta^{(1)}_{2(1)}$
2	$\beta_2 = \beta_{01}$	small(=0)	yes(=1)	$\theta^{(0)} + \theta^{(1)}_{1(1)} + \theta^{(1)}_{2(2)}$
3	$\beta_3 = \beta_{10}$	medium(=1)	no(=0)	$\theta^{(0)} + \theta^{(1)}_{1(2)} + \theta^{(1)}_{2(1)}$
4	$\beta_4 = \beta_{11}$	medium(=1)	yes(=1)	$\theta^{(0)} + \theta^{(1)}_{1(2)} + \theta^{(1)}_{2(2)}$
5	$\beta_5 = \beta_{20}$	large(=2)	no(=0)	$\theta^{(0)} + \theta^{(1)}_{1(3)} + \theta^{(1)}_{2(1)}$
6	$\beta_6 = \beta_{21}$	large(=2)	yes(=1)	$\theta^{(0)} + \theta^{(1)}_{1(3)} + \theta^{(1)}_{2(2)}$

and β_{21}. Reparameterization of those in terms of θs are given in Table 1. The contrast arrays $(\theta^{(1)}_{1(1)}, \theta^{(1)}_{1(2)}, \theta^{(1)}_{1(3)})^{\mathrm{T}}$ and $(\theta^{(1)}_{2(1)}, \theta^{(1)}_{2(2)})^{\mathrm{T}}$ are related with tumor size and nodal status, respectively. The reference level contrasts are set at 0 for estimability (i.e. $\theta^{(1)}_{1(1)} = \theta^{(1)}_{2(1)} = 0$). Each non-reference contrast parameter represents the association between the covariate and the odds of a certain disease characteristic being at a certain level in relation to its reference level. For instance, $\theta^{(1)}_{1(2)}$ represents the association between the covariate and the odds of tumor being medium relative to tumor being small. Also, under the estimability condition, $\theta^{(0)}$ is the coefficient specific to the reference disease subtype and represents the logOR associated with probability of the disease at the reference level relative to absence of the disease. In the case of multiple covariates, above ideas are replicated for each covariate with the proper indexing. Proper notations for multi covariate situations are given in [6]. The TS-LR model is especially beneficial when logOR estimates by cancer characteristics is desired. The θ parameters readily deliver logOR estimates by disease characteristics, whereas one would have to go through extra calculations to obtain them when a standard polychotomous logistic regression (approach *ii*) is used.

2.2 *ST for etiologic heterogeneity in TS-LR*

Source of etiologic heterogeneity in disease subtypes in terms of their associations with the risk factors is the etiologic heterogeneity in each disease characteristic. Etiologic heterogeneity in terms of the disease defining characteristics indicates etiologic heterogeneity between the disease subtypes with respect to the corresponding characteristics. Therefore, determining etiologic heterogeneity of the disease boils down to testing the difference between the related θ parameters for each disease characteristic. For instance, admitting to the illustrative example above, assuming a single covariate for the time being, the null hypothesis for testing etiological heterogeneity in tumor size is $H_0 : \theta^{(1)}_{1(2)} = \theta^{(1)}_{1(3)}$. The generalization is that for the disease characteristic k with m_k levels, testing its etiological heterogeneity is accomplished by setting $H_0 : \theta^{(1)}_{k(2)} = \theta^{(1)}_{k(3)} = \cdots = \theta^{(1)}_{k(m_k)}$ versus $H_1 : \theta^{(1)}_{k(i)} \neq \theta^{(1)}_{k(j)}$ for at least one $i, j \epsilon \{1, 2, \ldots, m_k\},\ i \neq j$. We derive a ST to test this hypothesis under the models (1) and (2). For a recent informative account of Rao's ST refer to [2]. Result of the test sheds light on risk factors with significant differential effects with respect to disease characteristic levels.

For parameter estimation, pseudo-conditional likelihood (PCL) method, the basics of which are recapitulated here, is used. Please refer to [6] for further methodological and theoretical details. This methodology uses a likelihood function where the building blocks are conditional probabilities. PCL is free of the intercept parameters, namely α_m's and depends only on the regression coefficients β_m's. In the resulting PCL, β_m's are replaced by the forms in model (2) and score equations for θ parameters are obtained. Resulting maximum likelihood estimators of θ parameters were proved to be asymptotically normal.

ST statistics, denoted by T_s, is given in Equation (3).

$$T_s = \boldsymbol{S}_1(\tilde{\boldsymbol{\theta}})^{\mathrm{T}} (\tilde{\boldsymbol{I}}_{T,11} - \tilde{\boldsymbol{I}}_{T,12} \tilde{\boldsymbol{I}}^{-1}_{T,22} \tilde{\boldsymbol{I}}_{T,21})^{-1} \boldsymbol{S}_1(\tilde{\boldsymbol{\theta}}). \tag{3}$$

In this formula, $\boldsymbol{S}_1(\tilde{\boldsymbol{\theta}}) = (\partial/\partial\tilde{\boldsymbol{\zeta}}) \log L_{\mathrm{PCL}}(\tilde{\boldsymbol{\theta}})$; $\log L_{\mathrm{PCL}}$: natural logarithm of PCL function for TS-LR model [6]; $\boldsymbol{\zeta} = (\theta^{(1)}_{k(2)}, \theta^{(1)}_{k(3)}, \ldots, \theta^{(1)}_{k(m_k)})^{\mathrm{T}}$, that is, vector of θ parameters being tested; $\boldsymbol{\theta}$: all the θ parameters in the model; $\tilde{\boldsymbol{\theta}}$: $\boldsymbol{\theta}$ estimates (i.e. maximum likelihood estimates based on L_{PCL}) constrained by H_0; $\boldsymbol{I}_T$: total information matrix; $\tilde{\boldsymbol{I}}_T$: $\boldsymbol{I}_T$ evaluated at $\tilde{\boldsymbol{\theta}}$; $\boldsymbol{I}_{T,11}$: partition of $\tilde{\boldsymbol{I}}_T$ associated with ζ; $\boldsymbol{I}_{T,22}$: partition of $\tilde{\boldsymbol{I}}_T$ associated with θ parameters other than ζ; $\boldsymbol{I}_{T,12}$: off-diagonals of $\tilde{\boldsymbol{I}}_T$.

Following [2,25], one can easily confirm that asymptotic distribution of T_s in the TS-LR model is χ^2 with degrees of freedom $m_k - 1$. This implies that when subtype frequencies are sufficiently

large, one can use suitable χ^2 quantiles for the test. Then two customary crucial questions follow: How large is sufficiently large? How should the test proceed if subtype frequencies are not large enough? Our Monte Carlo-based investigation sheds light on these questions. The results of the simulation study infer a condition for using χ^2 (the asymptotic distribution) for performing the test and the condition is based on minimum of the expected subtype frequencies, that is, minimum of the subtype frequencies that would have been expected if the disease characteristic of interest was non-heterogeneous with respect to its association with the particular risk factor of interest. For study designs in which the condition for asymptotic test is not satisfied, a permutation-based testing procedure is recommended. We derive the permutation algorithm and provide an outline for it here using the following example. For a $3 \times 2 \times 2$ disease where there are 3 characteristics with number of levels being 3, 2, and 2, respectively, binary dummy variables for each characteristic are constructed. To test for etiological heterogeneity of the disease characteristic – 1, for example, rows of its corresponding dummy variables are permuted (reshuffled) holding the rows of the other characteristics and the covariates untouched. A new column holding the disease subtype information is then constructed based on the resulting reshuffled levels. We used 10,000 such permuted data sets. The idea behind the permutation method is that the finite sample distribution of a test statistic under the null hypothesis can be realized by randomly shuffling the data several times obeying the condition presented in the null hypothesis and calculating the test statistic for each data set obtained as such (see [9,14]). The basis for our choice of resampling procedure over Bootstrap is that the rationale for permutation method appears to be clearer for hypothesis testing in linear models. For performing a level-α ST, critical points are set at $(1-\alpha)$th sample quantile of the resulting empirical distribution of the ST statistic. In case of multiple covariates, the method above is repeated in a similar manner for each covariate.

3. Simulation study

Our purpose is to get a rough idea about the minimum expected subtype frequencies necessary for asymptotic ST and power of permutation-based ST when expected subtype frequency does not meet the minimum condition. We try to address those via a simple Monte Carlo simulation study. To study type I error and power, two representative sample sizes are considered, $N = 500$ and $N = 1000$ with equal number of cases and controls. Two separate scenarios are considered for disease subtypes. First one concerns with a disease with 3 categorical characteristics each having 2 levels, resulting in 8 subtypes ($2 \times 2 \times 2$). The second scenario concerns with a disease with 3 categorical characteristics having 4, 2, and 2 levels respectively, resulting in 16 subtypes ($4 \times 2 \times 2$). For convenience, a single covariate is considered and covariate data are generated from the standard normal distribution. To generate the responses (i.e. disease subtypes), true values for θ parameters are set as explained in the following two paragraphs and then true β parameters are obtained using model (2). Then, given βs and X_is, Y_is (i.e. disease subtypes including disease-free case) are generated from a multinomial distribution with cell probabilities given in Equation (1). Hypothesis of interest is etiologic heterogeneity of a particular disease characteristic, for example, the first one, and denoted by $H_0 : \theta^{(1)}_{1(2)} = 0$ and $H_0 : \theta^{(1)}_{1(2)} = \theta^{(1)}_{1(3)} = \theta^{(1)}_{1(4)}$ for the first and second scenarios, respectively. The aim of the simulation study is to investigate the statistical properties of the asymptotic- and permutation-based STs for testing this type of hypotheses. Nominal significance level is set at 0.05.

We designed the following two-part Monte Carlo experiment to profile empirical false positive rates (type I errors) of asymptotic- and permutation-based ST for different minimum expected subtype frequencies. The aim of the first part is to determine true values for θ parameters so that they lead to a certain minimum expected average subtype frequency under H_0 of interest. Then in the second part, data are generated as explained in the previous paragraph using the θ values

predetermined by the first part so that we have a control over the minimum expected subtype frequencies in our experiment. In the first part, the sample size N and number of simulated data sets J are set at large numbers namely 10,000 and 20,000, respectively. Let $\tilde{p}_{1i}^{(j)}, \tilde{p}_{2i}^{(j)}, \ldots, \tilde{p}_{mi}^{(j)}$ denote the probabilities for each disease subtype for subject i in the jth simulated data set which is estimated using models (1) and (2) under the null hypothesis. Then, for each simulated data set, the probabilities associated with each subtype are averaged over the subjects to obtain an average probability of having a certain disease subtype. For the jth simulated data set, the average probability of being in subtype-m is $\bar{p}_{m\cdot}^{(j)} = (1/N)\sum_{i=1}^{N} \tilde{p}_{mi}^{(j)}$, $m = 1, \ldots, M$. Ultimately for each subtype, these average probabilities are averaged once more over 20,000 Monte Carlo iterations as $\bar{\bar{p}}_{m\cdot} = (1/J)\sum_{j=1}^{J} \bar{p}_{m\cdot}^{(j)}$. In order to obtain the average expected subtype frequencies, resulting values are multiplied by the number of the cases n_{case} as $n_{\text{case}} \times \bar{\bar{p}}_{m\cdot}$. Minimum of these frequencies are recorded. This work provided an idea about the θ values and the minimum subtype frequency they eventually lead to in the long run. These particular θ values are then used in the data generation process to study empirical type I error rates and their relationship with minimum subtype frequencies for the scenarios considered.

For the power study, true θ parameters are chosen to satisfy H_1. The first column of Table 4 shows the alternative values for $\theta_{2(2)}^{(1)}$ that correspond to varying degrees of etiologic heterogeneity most of which are local alternatives. For the case of $4 \times 2 \times 2$, the first two columns of Table 5 imply a different alternative hypothesis each time. Here, Δ_1 and Δ_2, respectively, are $\theta_{1(2)}^{(1)} - \theta_{1(3)}^{(1)}$

Table 2. Empirical type I error rates.

N	Min.avr. exp. subtype freq.	Asymptotic	Permutation based
500	13	0.0566	0.0460
	12	0.0554	0.0540
	11	0.0582	0.0520
	8	0.0566	0.0500
	7	0.0570	0.0520
	6	0.0609	0.0610
	5	0.0612	0.0560
	4	0.0601	0.0540
1000	25	0.0597	0.0400
	24	0.0638	0.0560
	23	0.0592	0.0570
	12	0.0691	0.0570
	11	0.0652	0.0540
	10	0.0673	0.0530

Note: Disease subtypes= $2 \times 2 \times 2$.

Table 3. Empirical type I error rates.

Min.avr. exp. subtype freq.	Asymptotic	Permutation based
12	0.0683	0.0540
11	0.0655	0.0580
10	0.0690	0.0630
8	0.0711	0.0640
7	0.0730	0.0660
6	0.0795	0.0630
5	0.0820	0.0620

Note: Disease subtypes $= 4 \times 2 \times 2$; $N = 1000$.

and $\theta^{(1)}_{1(3)} - \theta^{(1)}_{1(4)}$ which jointly correspond to different degrees of etiological heterogeneity in a characteristic with 4 levels. For power study 1000 iterations are used.

Results:

The aim of the study in Section 3.1. is to investigate the finite sample properties of the tests when there are subtype categories with particularly low expected frequencies. Tables 2 and 3 reported the empirical type I error rates for the cases $2 \times 2 \times 2$ and $4 \times 2 \times 2$, respectively. In these tables, the proportion of the minimum expected subtype frequencies vary between about 1% and 5% of the cases. The results show that, when the minimum subtype frequency is as low as $1 - 5\%$ of the cases, asymptotic ST is liberal in terms of empirical type I error rate and fails to maintain the nominal significance level. Practical implication of this finding is that asymptotic ST should be used when the proportion of study patients in each disease subtype expected under the null hypothesis is larger than the ones viewed here. Otherwise, the analyst is likely to result in false positive decision about etiologic heterogeneity. On the other hand, the same tables show that permutation-based ST maintains the nominal significance level better. The results for $4 \times 2 \times 2$

Table 4. Empirical power.

	Min. avr. exp. subtype freq. = 25		Min.avr. exp. subtype freq . = 10
$\theta^{(1)}_{2(2)}$	Asymptotic	Permutation based	Permutation based
0.500	0.9986	0.9980	0.9800
0.400	0.9800	0.9890	0.9210
0.300	0.8813	0.8720	0.7520
0.200	0.5793	0.5730	0.4340
0.100	0.2064	0.2350	0.1700
0.090	0.1799	0.1700	0.1580
0.080	0.1534	0.1460	0.1280
0.070	0.1336	0.1290	0.1050
0.060	0.1145	0.1010	0.1000
0.050	0.0978	0.0840	0.0880
0.040	0.0867	0.0860	0.0780
0.030	0.0755	0.0690	0.0640
0.020	0.0675	0.0480	0.0590
0.010	0.0636	0.0540	0.0640
0.001	0.0604	0.0430	0.0620

Note: Disease subtypes = $2 \times 2 \times 2$; $N = 1000$.

Table 5. Empirical power (permutation based).

Δ_1	Δ_2	Min.avr. exp. subtype freq . = 5	Min.avr. exp. subtype freq . = 12
1.500	1.500	1.0000	1.0000
1.000	1.000	1.0000	1.0000
0.400	0.400	1.0000	1.0000
0.300	0.300	0.9790	0.9800
0.200	0.200	0.7820	0.7980
0.100	0.100	0.2500	0.2950
0.050	0.050	0.1050	0.1090
0.040	0.040	0.0970	0.0970
0.030	0.030	0.0790	0.0800
0.020	0.020	0.0740	0.0610
0.010	0.010	0.0530	0.0550
0.001	0.001	0.0520	0.0570

Note: Disease subtypes = $4 \times 2 \times 2$; $N = 1000$.

case with $n = 500$ is disregarded as they did not add any further information. These results altogether imply that analyst should prefer the permutation-based test when minimum subtype frequency does not meet the guidelines produced here for the asymptotic $\chi^2_{(1)}$ to hold.

Power comparisons are given in Tables 4 and 5. From Table 4, when there is a subtype category with expected frequency as low as 5% of the cases (that is the block with 25) in a moderate situation such as $2 \times 2 \times 2$, asymptotic ST offers slightly larger power than the permutation-based ST, but at the cost of higher type I error rate. The second block of the table shows that when there is a subtype category with expected frequency as low as 2% of the cases, the preferred permutation-based ST achieves 92% power to detect even a small effect size such as 0.4 towards etiologic heterogeneity (i.e. towards 0). From Table 5, when there is a subtype category with expected frequency as low as 1% or 2% of the cases, the permutation-based ST achieves about at least 80% power to detect an effect of size at least (0.2,0.2) towards etiologic heterogeneity.

4. Illustrative example

The testing procedure is applied to understand the heterogeneity in risk factor-breast cancer subtype associations in the Turkish female breast cancer patients. The data set is obtained from a case-control study in Ankara Oncology Research and Education Hospital. There are 500 females in the study, of whom 249 have breast cancer. Reader should refer to [8] for further study details. Major tumor characteristics considered for our illustration are tumor size (extended to chest wall or skin, >50, >20 and ≤ 50, ≤ 20 mm), tumor type (invasive ductal, invasive lobular, and tubular), NA status (metastatis in axillary nodes or not), ER status $(+,-)$, PR status $(+,-)$, and Her2/neu receptor status $(+,-)$ where last levels are the reference levels. Cross-classifying these levels correspond to $4 \times 3 \times 2 \times 2 \times 2 \times 2 = 192$ breast cancer subtypes. Covariates in this study consist of the risk and the adjusting factors. Risk factors, namely age at menopause, age at first menstruation, number of births, age at first birth, age at last birth, duration of breast feeding, and duration of smoking are ascertained and indexed by $p = 1, \ldots, 7$. Adjusting factors typical to breast cancer studies, namely age and body mass index, are included in the model and indexed by $p = 8, 9$. A TS-LR model and maximum PCL estimation is used. Second stage main contrast parameters for each of these covariates are estimated and the ones of the main concern are given in Tables 6 and 7.

Developed testing strategy is employed to investigate the etiologic heterogeneity of the disease subtypes by means of heterogeneity in disease characteristic-risk factor association. Null

Table 6. Estimates (standard errors) of the second stage parameters.

Covariate	$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{1(3)}$	$\theta^{(1)}_{1(4)}$	$\theta^{(1)}_{2(2)}$
Age at menopause	−0.0776	0.0532	0.0497	0.0142	0.1328
	(0.0032)	(0.0020)	(0.0047)	(0.0113)	(0.0072)
Age at first menstruation	−0.2553	0.0411	−0.0693	0.2469	−0.3381
	(0.0297)	(0.0174)	(0.0413)	(0.0857)	(0.0537)
Number of births	−0.1558	0.0282	0.0252	−0.1795	0.1165
	(0.0547)	(0.0308)	(0.0746)	(0.1519)	(0.0740)
Age at first birth	0.0038	−0.0201	0.0380	−0.0184	0.0771
	(0.0040)	(0.0023)	(0.0053)	(0.0133)	(0.0087)
Age at last birth	0.0777	−0.0377	0.0233	−0.0127	−0.1510
	(0.0035)	(0.0020)	(0.0053)	(0.0133)	(0.0084)
Duration of breast feeding	−0.0160	0.0078	0.0016	0.0119	0.0167
	(0.0002)	(0.0001)	(0.0002)	(0.0004)	(0.0002)
Duration of smoking	−0.0038	0.0097	−0.1027	−0.1356	0.0837
	(0.0009)	(0.0005)	(0.0039)	(0.0256)	(0.0006)

Table 7. Estimates (standard errors) of the second stage parameters-(continued).

Covariate	$\theta^{(1)}_{2(3)}$	$\theta^{(1)}_{3(2)}$	$\theta^{(1)}_{4(2)}$	$\theta^{(1)}_{5(2)}$	$\theta^{(1)}_{6(2)}$
Age at menopause	− 0.0642	− 0.0410	0.0267	0.0763	− 0.0017
	(0.0055)	(0.0016)	(0.0017)	(0.0015)	(0.0021)
Age at first menstruation	0.2123	0.2005	0.0686	− 0.1204	0.0961
	(0.0491)	(0.0125)	(0.0146)	(0.0115)	(0.0156)
Number of births	− 0.3316	− 0.0537	0.1192	− 0.1081	− 0.1071
	(0.0822)	(0.0202)	(0.0244)	(0.0194)	(0.0241)
Age at first birth	− 0.1124	0.0003	− 0.0210	0.0553	− 0.0592
	(0.0078)	(0.0018)	(0.0021)	(0.0017)	(0.0026)
Age at last birth	0.0791	− 0.0238	0.0347	− 0.0353	0.0015
	(0.0056)	(0.0015)	(0.0018)	(0.0015)	(0.0020)
Duration of breast feeding	− 0.0001	0.0079	0.0003	0.0101	0.0138
	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Duration of smoking	0.0178	0.0003	0.0092	− 0.0488	0.0062
	(0.0014)	(0.0004)	(0.0005)	(0.0004)	(0.0006)

hypotheses of interest for the pth risk factor then are $H_{0,1,p} : \theta^{(1)}_{1(4)p} = \theta^{(1)}_{1(3)p} = \theta^{(1)}_{1(2)p}$, $H_{0,2,p} : \theta^{(1)}_{2(3)p} = \theta^{(1)}_{2(2)p}$, $H_{0,3,p} : \theta^{(1)}_{3(2)p} = 0$, $H_{0,4,p} : \theta^{(1)}_{4(2)p} = 0$, $H_{0,5,p} : \theta^{(1)}_{5(2)p} = 0$, and $H_{0,6,p} : \theta^{(1)}_{6(2)p} = 0$, respectively, for tumor size, tumor type, NA status, ER status, PR status, and Her2/neu receptor status, for $p = 1, \ldots, 7$. These are indexed by p to indicate that all these need to be run for each covariate.

Permutation-based ST is applied as it is proved to be more guarded against false discovery compared to its asymptotic counterpart. The reshuffling procedure is repeated 10,000 times and at each time, a ST statistic corresponding to each risk factor and each hypothesis is calculated and stored. These values are then used to obtain the required empirical quantile of distribution of the ST statistic under the null hypothesis for each risk factor. The observed test statistics and the permutation-based critical points are presented in Tables 8 and 9 herein and Tables 1–4 in the

Table 8. ST results for detecting etiologic heterogeneity for tumor size.

Covariate	T_s	95th sample quantile
Age at menopause	0.2360	6.6463
Age at first menstruation	0.8123	4.8740
Number of births	0.0237	5.5642
Age at first birth	4.6794	4.8546
Age at last birth	4.0438	5.8492
Duration of breast feeding	0.2561	4.6109
Duration of smoking	5.7204	7.2364

Table 9. ST results for detecting etiologic heterogeneity for tumor type.

Covariate	T_s	95th sample quantile
Age at menopause	1.1109	4.2178
Age at first menstruation	3.1816	3.2365
Number of births	0.4960	3.6309
Age at first birth	0.4106	3.1248
Age at last birth	2.3382	3.8619
Duration of breast feeding	0.5991	2.9056
Duration of smoking	2.0982	4.3071

supplemental file, each corresponding to a particular disease characteristic as seen. The results reveal that at the 5% significance level, there is statistically significant evidence in the data set that PR status and Her2/neu receptor status are etiologically heterogeneous with respect to their relationship with smoking and breastfeeding durations, respectively (see supplemental material). That means, smoking has a different effect on PR positive and PR negative breast cancer. Similarly, duration of breastfeeding–breast cancer association is significantly different for Her2/neu receptor positive and negative cancers. Looking back at Table 7, estimate (-0.0488) given in (duration of smoking, $\theta^{(1)}_{5(2)}$) suggests that the duration of smoking is more strongly associated with PR negative-type breast cancers. Also estimate (0.0138) given in (duration of breastfeeding, $\theta^{(1)}_{6(2)}$) suggests that the duration of breastfeeding is more strongly associated with Her2/neu receptor positive-type breast cancers.

5. Discussion

The suggested testing strategy consists of two stages:

Stage 1: Estimate subtype frequencies that would have been expected if the null hypothesis (etiologically non-heterogeneous) was true. The decision as to whether use asymptotic- or permutation-based test is then made given the minimum of the expected frequencies. The rule of thumb to which we are led following the results of our simulation study is that avoid using asymptotic test for etiologic heterogeneity and prefer permutation-based test instead if minimum expected subtype frequency is less than about 5% of the study cases.

Stage 2: Apply the test chosen at stage 1. Use χ^2 tables with appropriate degrees of freedom for asymptotic test. For permutation test, carry out the permutations as outlined to obtain exact distribution of the ST statistic under the null hypothesis; use the appropriate empirical quantile of this distribution to finalize the test.

We considered single covariate in the simulations for clear interpretation of the results. Inclusion of multiple covariates in the model surely would have changed parameter and information matrix estimates. However, there is no reason to believe that it would have altered the conclusion derived from the simulation results regarding its power properties should we had multiple covariates. Because each hypothesis is concerned with etiological heterogeneity related with each covariate separately. Nevertheless, applications with multiple covariate case is addressed in the application section.

The sample sizes and the number of subtypes utilized may be considered as a semblance of limitation. However, it does not curb the intuition obtained from this study as to whether to proceed with asymptotic- or permutation-based test, nor does it preclude generalizing the results and what this paper aims to contribute to the epidemiological data analysis.

Acknowledgments

We thank Dr Nilanjan Chatterjee, Biostatistics Branch of Division of Cancer Epidemiology and Genetics, National Cancer Institute, for bringing our attention the need for a statistical procedure testing the etiologic heterogeneity in diseases with multiple characteristics particularly in breast cancer.

Disclosure statement

No potential conflict of interest was reported by the authors.

Supplemental data

Supplemental data for this article can be accessed at doi:10.1080/02664763.2015.1070808.

References

[1] C.B. Begg and E.C. Zabor, *Detecting and exploiting etiologic heterogeneity in epidemiologic studies*, Am. J. Epidemiol. 176 (2012), pp. 512–518.

[2] A.K. Bera and Y. Bilias, *Rao's score, Neyman's C(α) and Silvey's LM tests: An assay on historical developments and some new results*, J. Statist. Plann. Inference 97 (2001), pp. 9–44.

[3] S.I. Berndt, W.Y. Huang, N. Chatterjee, M. Yeager, R. Welch, S.J. Chanock, J.L. Weissfeld, R.E. Schoen, and R.B. Hayes, *Transforming growth factor beta 1 (TGFB1) gene polymorphisms and risk of advanced colorectal adenoma*, Carcinogenesis. 28 (2007), pp. 1965–1970.

[4] D.D. Boos and L.A. Stefanski, *Essential statistical inference, theory and methods*, Springer-Verlag, New York, 2013.

[5] S.B. Bull, C. Mak, and C.M.T. Greenwood, *A modified score function estimator for multinomial logistic regression in small samples*, Comput. Stat. Data. Anal. 39 (2002), pp. 57–74.

[6] N. Chatterjee, *A two stage regression model for epidemiological studies with multivariate disease classification data*, J. Amer. Statist. Assoc. 99 (2004), pp. 127–138.

[7] G.M. Cordeiro and P. McCullagh, *Bias correction in generalized linear models*, J. Roy. Statist. Soc. B. 53 (1991), pp. 629–643.

[8] L. Dogan, Z. Kalaylioglu, N. Karaman, C. Ozaslan, C. Atalay, and M. Altinok, *Relationships between epidemiological features and tumor characteristics of breast cancer*, Asian Pac. J. Cancer Prev. 12 (2011), pp. 3375–3380.

[9] M.D. Ernst, *Permutation methods: A basis for exact inference*, Statist. Sci. 19 (2004), pp. 676–685.

[10] M. Garcia-Closas, L.A. Brinton, J. Lissowska, N. Chatterjee, B. Peplonska, W.F. Anderson, N. Szeszenia-Dabrowska, A. Bardin-Mikolajczak, W. Zatonski, A. Blair, Z. Kalaylioglu, G. Rymkiewicz, D. Mazepa-Sikora, R. Kordek, S. Lukaszek, and M.E. Sherman, *Established breast cancer risk factors by clinically important tumour characteristics*, Br. J. Cancer. 95 (2006), pp. 123–129.

[11] M. Garcia-Closas and S. Chanock, *Genetic susceptibility loci for breast cancer by estrogen receptor status*, Clin. Cancer Res. 14 (2008), pp. 8000–8009.

[12] M. Garcia-Closas, P. Hall, H. Nevanlinna, K. Pooley, J. Morrison, D.A. Richesson, S.E. Bojesen, B.G. Nordestgaard, C.K. Axelsson, J.I. Arias, R.L. Milne, G. Ribas, A. González-Neira, J. Benítez, P. Zamora, H. Brauch, C. Justenhoven, U. Hamann, Y.-D. Ko, T. Bruening, S. Haas, T. Dörk, P. Schürmann, P. Hillemanns, N. Bogdanova, M. Bremer, J.H. Karstens, R. Fagerholm, K. Aaltonen, K. Aittomäki, K. von Smitten, C. Blomqvist, A. Mannermaa, M. Uusitupa, M. Eskelinen, M. Tengström, V.-M. Kosma, V. Kataja, G. Chenevix-Trench, A.B. Spurdle, J. Beesley, X. Chen, P. Devilee, C.J. van Asperen, C.E. Jacobi, R.A.E.M. Tollenaar, P.E.A. Huijts, J.G.M. Klijn, J. Chang-Claude, S. Kropp, T. Slanger, D. Flesch-Janys, E. Mutschelknauss, R. Salazar, S. Wang-Gohrke, F. Couch, E.L. Goode, J.E. Olson, C. Vachon, Z.S. Fredericksen, G.G. Giles, L. Baglietto, G. Severi, J.L. Hopper, D.R. English, M.C. Southey, C.A. Haiman, B.E. Henderson, L.N. Kolonel, L. Le Marchand, D.O. Stram, D.J. Hunter, S.E. Hankinson, D.G. Cox, R. Tamimi, P. Kraft, M.E. Sherman, S.J. Chanock, J. Lissowska, L.A. Brinton, B. Peplonska, J.G.M. Klijn, M.J. Hooning, H. Meijers-Heijboer, J.M. Collee, A. van den Ouweland, A.G. Uitterlinden, J. Liu, L.Y. Lin, L. Yuqing, K. Humphreys, K. Czene, A. Cox, S.P. Balasubramanian, S.S. Cross, M.W.R. Reed, F. Blows, K. Driver, A. Dunning, J. Tyrer, B.A.J. Ponder, S. Sangrajrang, P. Brennan, J. McKay, F. Odefrey, V. Gabrieau, A. Sigurdson, M. Doody, J.P. Struewing, B. Alexander, D.F. Easton, and P.D. Pharoah, *Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics*, Australian Ovarian Cancer Management Group; Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer, PLoS Genet. 4 (2008), p. e1000054.

[13] M.M. Gaudet, M.F. Press, R.W. Haile, C.F. Lynch, S.L. Glaser, J. Schildkraut, M.D. Gammon, T.W. Douglas, 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 (2011), pp. 587–597.

[14] P.I. Good, *Permutation, parametric, and bootstrap tests of hypotheses*, Springer-Verlag, New York, 2005.

[15] L. Hou, N. Chatterjee, W.Y. Huang, A. Baccarelli, S. Yadavalli, M. Yeager, R.S. Bresalier, S.J. Chanock, N.E. Caporaso, B.T. Ji, J.L. Weissfeld, and R.B. Hayes, *CYP1A1 Val462 and NQO1 Ser187 polymorphisms, cigarette use, and risk for colorectal adenoma*, Carcinogenesis. 26 (2005), pp. 1122–1128.

[16] W.Y. Huang, N. Chatterjee, S. Chanock, M. Dean, M. Yeager, R.E. Schoen, L.F. Hou, S.I. Berndt, S. Yadavalli, C.C. Johnson, and R.B. Hayes, *Microsomal epoxide hydrolase polymorphisms and risk for advanced colorectal adenoma*, Cancer Epidemiol. Biomarkers Prev. 14 (2005), pp. 152–157.

[17] L.E. Moore, W.Y Huang, N. Chatterjee, M. Gunter, S. Chanock, M. Yeager, B. Welch, P. Pinsky, J. Weissfeld, and R.B. Hayes, *GSTM1, GSTT1, and GSTP1 polymorphisms and risk of advanced colorectal adenoma*, Cancer Epidemiol. Biomarkers Prev. 14 (2005), pp. 1823–1827.

[18] M.F. Munsell, B.L. Sprague, D.A. Berry, G. Chisholm, and A. Trentham-Dietz, *Body mass index and breast cancer risk according to postmenopausal estrogen-progestin use and hormone receptor status*, Epidemiol. Rev. 36 (2014), pp. 114–136.

[19] H.B. Nichols, A. Trentham-Dietz, R.R. Love, J.M. Hampton, P.T. Hoang Anh, D.C. Allred, S.K. Mohsin, and P.A. Newcomb, *Differences in breast cancer risk factors by tumor marker subtypes among premenopausal Vietnamese and Chinese women*, Cancer Epidemiol. Biomarkers Prev. 14 (2005), pp. 41–47.

[20] R. Nishihara, T. Morikawa, A. Kuchiba, P. Lochhead, M. Yamauchi, X. Liao, Y. Imamura, K. Nosho, K. Shima, I. Kawachi, Z.R. Qian, C.S. Fuchs, A.T. Chan, E. Giovannucci, and S. Ogino, *A prospective study of duration of smoking cessation and colorectal cancer risk by epigenetics-related tumor classification*, Am. J. Epidemiol. 178 (2013), pp. 84–100.

[21] U. Peters, N. Chatterjee, K.A. McGlynn, R.E. Schoen, T.R. Church, R.S. Bresalier, M.M. Gaudet, A. Flood, A. Schatzkin, and R.B. Hayes, *Calcium intake and colorectal adenoma in a US colorectal cancer early detection program*, Am. J. Clin. Nutr. 80 (2004), pp. 1358–1365.

[22] M.P. Purdue, D.G. Bassani, N.S. Klar, M. Sloan, and N. Kreiger; Canadian Cancer Registries Epidemiology Research Group; *Dietary factors and risk of non-Hodgkin lymphoma by histologic subtype: A case-control analysis*, Cancer Epidemiol. Biomarkers Prev. 13 (2004), pp. 1665–1676.

[23] J.M. Schildkraut, E.S. Iversen, L. Akushevic, 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 (2013), pp. 1709–1721.

[24] M.E. Sherman, D.L. Rimm, X.R. Yang, N. Chatterjee, L.A. Brinton, J. Lissowska, B. Peplonska, N. Szeszenia-Dabrowska, W. Zatonski, R. Cartun, D. Mandich, G. Rymkiewicz, M. Ligaj, S. Lukaszek, R. Kordek, Z. Kalaylioglu, M. Harigopal, L. Charrette, R.T. Falk, D. Richesson, W.F. Anderson, S.M. Hewitt, and M. García-Closas, *Variation in breast cancer hormone receptor and HER2 levels by etiologic factors: A population-based analysis*, Int. J. Cancer. 121 (2007), pp. 1079–1085.

[25] S.D. Silvey, *The lagrangian multiplier test*, Ann. Math. Stat. 30 (1959), pp. 389–407.

[26] M.A. Troester and T. Swift-Scanlan, *Challenges in studying the etiology of breast cancer subtypes*, Breast Cancer Res. Treat. 11 (2009), pp. 1–2.

[27] S.S. Tworoger, A.H. Eliassen, B. Rosner, P. Sluss, and S.E. Hankinson, *Plasma prolactin concentrations and risk of postmenopausal breast cancer*, Cancer Res. 64 (2004), pp. 6814–6819.

[28] H.P. Yang, B. Trabert, M.A. Murphy, M.E. Sherman, J.N. Sampson, L.A. Brinton, P. Hartge, A. Hollenbeck, Y. Park, and N. Wentzensen, *Ovarian cancer risk factors by histologic subtypes in the NIH-AARP Diet and Health Study*, Int. J. Cancer. 131 (2012), pp. 938–948.

[29] X. R. Yang, J. Chang-Claude, E. L. Goode, F. J. Couch, H. Nevanlinna, R. L. Milne, M. Gaudet, M. K. Schmidt, A. Broeks, A. Cox, P. A. Fasching, R. Hein, A. B. Spurdle, F. Blows, K. Driver, D. Flesch-Janys, J. Heinz, P. Sinn, A. Vrieling, T. Heikkinen, K. Aittomaki, P. Heikkila, C. Blomqvist, J. Lissowska, B. Peplonska, S. Chanock, J. Figueroa, L. Brinton, P. Hall, K. Czene, K. Humphreys, H. Darabi, J. Liu, L. J. Van 't Veer, F. E. van Leeuwen, I. L. Andrulis, G. Glendon, J. A. Knight, A. M. Mulligan, F. P. O'Malley, N. Weerasooriya, E. M. John, M. W. Beckmann, A. Hartmann, S. B. Weihbrecht, D. L. Wachter, S. M. Jud, C. R. Loehberg, L. Baglietto, D. R. English, G. G. Giles, C. A. McLean, G. Severi, D. Lambrechts, T. Vandorpe, C. Weltens, R. Paridaens, A. Smeets, P. Neven, H. Wildiers, X. Wang, J. E. Olson, V. Cafourek, Z. Fredericksen, M. Kosel, C. Vachon, H. E. Cramp, D. Connley, S. S. Cross, S. P. Balasubramanian, M. W. R. Reed, T. Dork, M. Bremer, A. Meyer, J. H. Karstens, A. Ay, T.-W. Park-Simon, P. Hillemanns, J. I. Arias Perez, P. M. Rodriguez, P. Zamora, J. Benitez, Y.-D. Ko, H.-P. Fischer, U. Hamann, B. Pesch, T. Bruning, C. Justenhoven, H. Brauch, D. M. Eccles, W. J. Tapper, S. M. Gerty, E. J. Sawyer, I. P. Tomlinson, A. Jones, M. Kerin, N. Miller, N. McInerney, H. Anton-Culver, A. Ziogas, C.-Y. Shen, C.-N. Hsiung, P.-E. Wu, S.-L. Yang, J.-C. Yu, S.-T. Chen, G.-C. Hsu, C. A. Haiman, B. E. Henderson, L. Le Marchand, L. N. Kolonel, A. Lindblom, S. Margolin, A. Jakubowska, J. Lubinski, T. Huzarski, T. Byrski, B. Gorski, J. Gronwald, M. J. Hooning, A. Hollestelle, A. M. W. van den Ouweland, A. Jager, M. Kriege, M. M. A. Tilanus-Linthorst, M. Collee, S. Wang-Gohrke, K. Pylkas, A. Jukkola-Vuorinen, K. Mononen, M. Grip, P. Hirvikoski, R. Winqvist, A. Mannermaa, V.-M. Kosma, J. Kauppinen, V. Kataja, P. Auvinen, Y. Soini, R. Sironen, S. E. Bojesen, D. Dynnes Orsted, D. Kaur-Knudsen, H. Flyger, B. G. Nordestgaard, H. Holland, G. Chenevix-Trench, S. Manoukian, M. Barile, P. Radice, S. E. Hankinson, D. J. Hunter, R. Tamimi, S. Sangrajrang, P. Brennan, J. McKay, F. Odefrey, V. Gaborieau, P. Devilee, P. E. A. Huijts, R. Tollenaar, C. Seynaeve, G. S. Dite, C. Apicella, J. L. Hopper, F. Hammet, H. Tsimiklis, L. D. Smith, M. C. Southey, M. K. Humphreys, D. Easton, P. Pharoah, M. E. Sherman, and M. Garcia-Closas, *Associations of breast cancer risk factors with tumor subtypes: A pooled analysis from the Breast Cancer Association Consortium studies*, J. Natl. Cancer. Inst. 103 (2011), pp. 250–263.

[30] X.R. Yang, M.E. Sherman, D.L. Rimm, J. Lissowska, L.A. Brinton, B. Peplonska, S.M. Hewitt, W.F. Anderson, N. Szeszenia-Dabrowska, A. Bardin-Mikolajczak, W. Zatonski, R. Cartun, D. Mandich, G. Rymkiewicz, M. Ligaj, S. Lukaszek, R. Kordek, and M. García-Closas, *Differences in risk factors for breast cancer molecular subtypes in a population-based study*, Cancer Epidemiol. Biomarkers Prev. 16 (2007), pp. 439–443.