Published in final edited form as:

Stat Med. 2016 December 30; 35(30): 5686–5700. doi:10.1002/sim.7083.

Accounting for measurement error in biomarker data and misclassification of subtypes in the analysis of tumor data

Daniel Nevo^{a,b,*†}, David M. Zucker^a, Rulla M. Tamimi^{c,d}, and Molin Wang^b

^aDepartment of Statistics, Hebrew University, Jerusalem, Israel

^bDepartments of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A.

^cDepartment of Epidemiology, Harvard T.H. Chan School of Public Health, Boston MA, U.S.A.

^dChanning Division of Network Medicine, Brigham and Women's Hospital, Boston MA, U.S.A.

Abstract

A common paradigm in dealing with heterogeneity across tumors in cancer analysis is to cluster the tumors into subtypes using marker data on the tumor, and then to analyze each of the clusters separately. A more specific target is to investigate the association between risk factors and specific subtypes and to use the results for personalized preventive treatment. This task is usually carried out in two steps—clustering and risk factor assessment. However, two sources of measurement error arise in these problems. The first is the measurement error in the biomarker values. The second is the misclassification error when assigning observations to clusters. We consider the case with a specified set of relevant markers and propose a unified single-likelihood approach for normally distributed biomarkers. As an alternative, we consider a two-step procedure with the tumor type misclassification error taken into account in the second-step risk factor analysis. We describe our method for binary data and also for survival analysis data using a modified version of the Cox model. We present asymptotic theory for the proposed estimators. Simulation results indicate that our methods significantly lower the bias with a small price being paid in terms of variance. We present an analysis of breast cancer data from the Nurses' Health Study to demonstrate the utility of our method.

Keywords

risk-factor analysis; measurement error; classification; clustering; heterogeneity

*Correspondence to: Daniel Nevo, Departments of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A.

†danielnevo@gmail.com

Supporting information

Additional supporting information may be found in the online version of this article at the publishers web site.

1. Introduction

Within a given type of cancer, substantial heterogeneity often exists among tumors at the molecular level. This heterogeneity can reflect the existence of disease subtypes representing distinct biological entities. A common approach is to cluster tumors into subtypes (or clusters) based on biomarker data. Breast cancer is one example. In the 1970s, estrogen receptors (ER) status and tumor grade were identified as important prognostic factors. More recently, in the early 2000s, four major breast cancer molecular subtypes were identified: luminal A-like, luminal B-like, HER2+ type, and triple negative (or basal-like). It has been found that different breast cancer subtypes are associated with different etiologies, risk factors, clinical outcomes, and treatment options.

In particular, heterogeneity among tumors at the molecular level is often associated with the effects of risk factors. With tumors clustered into subtypes as described previously, risk factor effects can be examined for each subtype separately. In the case of breast cancer, for example, the effect of Body Mass Index (BMI) on tumor incidence varies according to the levels of ER and progesterone receptors (PR) [1–3]. However, tumor biomarker levels exhibit substantial biological variation. In addition, the measured biomarker levels are often subject to substantial measurement error. Failing to account for these sources of variability may lead to biased estimators of parameters such as the cluster means and to bias in the assessment of risk factor effects. Thus, in analyzing etiological heterogeneity across tumor subtypes, it is necessary to account for variation in observed biomarker levels within tumor subtypes.

A natural approach to carrying this out is to model the distribution of the vector of biomarker levels as a mixture of a family of distributions, e.g., multivariate Gaussians, with parameters varying across the mixture components. Mixture models are reasonable and popular approach for the analysis of heterogeneous data and have been studied extensively. See, for example, [4]. In [5], this approach is described for the analysis of microarray expression data. This paper also suggested a mixture of factor analyzers model to reduce the dimensionality of the model. In [6], the variable selection problem is recast as a model selection problem, and a greedy search algorithm is suggested. This algorithm compares nested models using BIC-based approximation of Bayes factors.

There is a growing literature on how to choose which markers should be used out to identify subtype membership of each observation. The current discussion in the research community centers in the case the number of measured biomarkers is larger than the number of available samples, and hence, a sparse statistical model is needed, e.g., [7–10]. See also [11] for a recent survey of model-based clustering for high-dimensional data. An alternative to model-based clustering is to use a more heuristic approach, and to avoid of assuming a statistical model. One possible heuristic approach is to search for a subspace which describes well the observed heterogeneity in the data. See [12] for a review of subspace clustering methods and [13] for more recent review that also includes additional methods.

This paper has a different focus. We assume the set of relevant biomarkers for defining the disease subtypes (clusters) is given, and we focus on the inter-relationship between the

process of characterizing each subtype, i.e., estimating relevant distribution parameters (which may vary across the subtypes) and the process of estimating risk factor effects for each subtype. In assessing the risk factor effects, our proposed methods account for the uncertainty in subtyping each given disease case that arises from the variation in the tumor biomarker levels within each subtype. This variation includes not only the biologic mechanics of the human body, but also error in the measured biomarker level because of the measurement process itself. Although the Gaussian mixture model we use for characterizing the subtypes is well-established, the inter-relationship between characterizing disease subtypes and estimating risk factor effects for each subtype has not, to our knowledge, been previously investigated.

We consider two settings. In the first, the information on each subject includes a binary disease (tumor) indicator, a risk factor vector, and the measured tumor biomarker levels. The latter is only observed for patients with the disease. However, their disease subtype is unknown. Therefore, while the observed outcome is binary, the underlying model is of a multinomial nature. The second scenario involves survival data. Subjects are followed until the occurrence of either an event or censoring. However, when an event is observed, it is unknown which subtype it is.

The rest of the paper is organized as follows. Section 2 introduces the notation and problem setting. Section 3 first presents an latent multinomial model for observed binary outcomes, then proceeds to describe various two-step methods for estimating the parameters of interest, and finally describes an asymptotically efficient unified likelihood approach. Section 4 presents a simulation study comparing the different methods. In Section 5, a model is developed for survival data, and a partial likelihood is constructed and used to obtain consistent and asymptotically normally distributed estimators. Section 6 presents simulation results for the survival data setting. In Section 7, the partial likelihood approach is applied to breast cancer data in the Nurses' Health Study (NHS) [2]. Section 8 provides a brief discussion.

2. Notation and problem setting

Let $\mathbf{X}$ denote the vector of true expression levels of p markers in a tumor sample. We assume that $\mathbf{X}$ is distributed according to a K -component multivariate normal mixture, with the mixture components representing biologically distinct tumor subtypes. We denote by $\boldsymbol{\mu}_k$ the mean marker vector in Cluster k ($k = 1, \dots, K$), by $\boldsymbol{\Gamma}_k$ the corresponding covariance matrix, and by $\pi_1, \dots, \pi_K$ the mixture probabilities ($\sum \pi_k = 1$). The vector $\mathbf{X}$ is not observed directly. Instead, we observe $\mathbf{W} = \mathbf{X} + \mathbf{U}$ with $\mathbf{U}$ being a measurement error vector, independent of $\mathbf{X}$, that follows a multivariate normal distribution with mean zero and a covariance matrix $\boldsymbol{\Omega}$. Here, we treat $\boldsymbol{\Omega}$ as known. If $\boldsymbol{\Omega}$ is unknown, it can be estimated using a validation study or replicate measurements of $\mathbf{W}$ in the main study itself, as in the analysis in Section 7 of breast cancer data from the NHS. We call the unknown parameters

$$\nu = \{\pi, \boldsymbol{\mu}, \boldsymbol{\Gamma}\} := \{\pi_1, \dots, \pi_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_K\}$$

the clustering parameters. We also denote $\Sigma_k = \Gamma_k + \Omega$, and note that $\mathbf{W}$ is distributed according a K -component normal mixture with the components being $\mathcal{N}(\mu_k, \Sigma_k)$, $k = 1, \dots, K$, with the same mixture probabilities as $\mathbf{X}$.

A vector of risk factors $\mathbf{Q}$ is measured without error, and the analysis is conditioned on the value $\mathbf{Q}$ takes. We denote the tumor subtype (or disease subtype) by Y , where Y equals 0 for no disease, and equals k for disease of type k . We also denote $D = \mathbb{I}_{\{Y>0\}}$ for the indicator of disease occurrence, regardless of the subtype. We use the notation of C for the subtype (cluster) membership of an observation with $D = 1$. We also assume that given C , $\mathbf{W}$ and $\mathbf{Q}$ are independent. That is, given the subtype membership, the biomarker levels and the risk factors are independent.

3. Multinomial logistic model

In the binary data scenario where the outcome variable is the binary variable D , the triplet $(\mathbf{W}_i, D_i, \mathbf{Q}_i)$ is observed for each subject i where $\mathbf{W}_i$ is considered to be missing if $D_i = 0$. However, the assumed underlying cluster structure implies a multinomial model for Y . Let $p_k(\mathbf{q}) = \Pr(Y = k | \mathbf{Q} = \mathbf{q})$ be the probability function of Y given $\mathbf{Q}$. In this paper, we consider the logistic model

$$p_k(\mathbf{q}) = \frac{\exp(\alpha_k + \beta_k^T \mathbf{q})}{1 + \sum_{j=1}^K \exp(\alpha_j + \beta_j^T \mathbf{q})} \quad (1)$$

for $k = 1, \dots, K$, and $p_0(\mathbf{q})$ is simply $1 - \sum_{k=1}^K p_k(\mathbf{q})$. We refer to the parameters $\boldsymbol{\varphi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\} = \{\alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K\}$ as the *risk parameters*. Our problem is not a covariate measurement error problem; rather, the challenge here is addressing misclassification of the *outcome*, which is due in part to natural variation in the biomarkers and in part to measurement error in the biomarkers.

We now present a series of methods for estimating the risk factor parameters. We begin by presenting, as benchmarks, two approaches commonly used in practice, which we call the naive K-means approach and the hard thresholding approach. These approaches take a two-step form, involving an initial clustering step followed by a risk assessment step, where the risk assessment step disregards the possibility of assigning a disease case to the wrong cluster as a result of natural variation in the biomarker levels and measurement error in the observed biomarker levels. We then present our proposed methods, which are designed to take into account this misclassification. We begin with a relatively simple two-step method, which we call the Semi-hard method. We then present a more sophisticated two-step method, which we call the soft method. Finally, we present a unified likelihood method in which the clustering parameters and risk factor parameters are estimated in a single step.

3.1. Benchmark approaches

In the approach, we call the naive K-means method, the analysis begins by applying the well-known K-means clustering algorithm [14] to the biomarker data ($\mathbf{W}$) to assign each of the observations with $D = 1$ to a cluster. We let $\hat{Y}_i^{NV}$ be subject i 's assigned cluster if $D_i = 1$, with $\hat{Y}_i^{NV} = 0$ if $D_i = 0$. With the clusters thus formed, the clustering parameters are estimated as described in Section 4. The risk assessment step involves estimation of the parameters in the risk factor model by maximizing the multinomial likelihood

$$L_{NV}(\varphi) = \prod_{i=1}^n p_{\hat{Y}_i^{NV}}(\mathbf{q}_i).$$

In this naive K-means method, the clustering step is carried heuristically rather than within a formal statistical framework.

In the approach, we call the hard thresholding method, the analysis begins, as with the K-means method, with a clustering step where each of the observations with $D_i = 1$ is assigned to a cluster $\hat{Y}_i^{HT}$, but now the clustering step is carried out within a formal statistical framework. Estimates of the clustering parameters are obtained as the maximizers of the normal mixture (NM) likelihood

$$L_{NM} = \prod_{\{i: D_i=1\}} \left[\sum_{k=1}^K \pi_k \det(\Sigma_k)^{-1/2} \phi(\Sigma_k^{-1/2}(\mathbf{w}_i - \boldsymbol{\mu}_k)) \right] \quad (2)$$

with $\phi(\cdot)$ being the standard p -variate normal density. We note here some details about the maximization process, which will carry over to our proposed methods.

Recall that $\Sigma_k = \Gamma_k + \Omega$ where Ω is a known positive definite matrix. We use a Cholesky decomposition for the parametrization of Γ_k , namely $\Gamma_k = \mathbf{P}_k \mathbf{P}_k^T$ where $\mathbf{P}_k$ is a lower diagonal matrix with positive elements on the diagonal. This approach has good performance both computationally and statistically as described in [15].

In order to ensure identifiability of the model, we assume without loss of generality that $\boldsymbol{\mu}_k$'s satisfy the constraint $\mu_{11} \quad \mu_{21} \quad \dots \quad \mu_{K1}$, where μ_{ki} is the i -th entry in $\boldsymbol{\mu}_k$, the mean vector in cluster k . A second family of linear constraints are $0 < \pi_k < 1$ for all k and $\sum_{k=1}^K \pi_k = 1$ to ensure that the vector $\boldsymbol{\pi} = (\pi_1 \quad \pi_2 \quad \dots \quad \pi_K)$ is indeed a vector of mixture probabilities. Finally, we impose additional linear constraints so the diagonal elements of Γ_k are strictly positive for all k , thereby ensuring that the estimate of Γ_k (the covariance matrix of $\mathbf{X}$ in Cluster k) is a positive definite matrix. Initial values are obtained using the K-means cluster assignments. We can then use any algorithm for linearly constrained maximization. Details are given in Section 4.

A common problem in normal mixture likelihood maximization is the unboundedness of the likelihood function for unwanted solutions. To address this problem, further nonlinear constraints are usually added, and local maximizers have to be carefully examined. See [4], particularly Sections 3.8.1 and 3.9.1. However, because in our case $\Sigma_k = \Gamma_k + \Omega$ and Ω is a known positive definite matrix, we have a lower positive bound for $\det(\Sigma_k)$ and hence, L_{NM} is bounded as a function of $\{\pi_k, \mu_k, \Gamma_k\}$. See Lemma 1 in Web Appendix A for the details.

A popular tool for the maximization of a mixture model likelihood is the EM algorithm [16], or one of its many variants. However, our experience in the present context is that there is no real gain in terms of computation or performance compared with the direct (constrained) maximization approach, and therefore, we use the latter.

The clustering parameters having been estimated, each disease case is assigned to a cluster in the following manner. Let $Z_i^{(k)}$ denote the probability that subject i belongs to cluster k , given the observed biomarker data on subject i , and define $\mathbf{Z}_i = (Z_i^{(1)} Z_i^{(2)} \dots Z_i^{(K)})$. That is,

$$Z_i^{(k)} = \Pr(C_i = k | \mathbf{W}_i, D_i = 1) = \frac{\pi_k \det(\Sigma_k)^{-1/2} \phi(\Sigma_k^{-1/2}(\mathbf{W}_i - \mu_k))}{\sum_{j=1}^K [\pi_j \det(\Sigma_j)^{-1/2} \phi(\Sigma_j^{-1/2}(\mathbf{W}_i - \mu_j))]} \quad (3)$$

with $\mathbf{Z}_i$ is taken to be missing if $D_i = 0$. The quantity $\mathbf{Z}_i$ depends of the cluster parameters ν , but we will often suppress this dependence from the notation. Let

$\tilde{Y}_i^{\text{HT}} = \tilde{Y}_i^{\text{HT}}(\nu) = \arg\max_k Z_i^{(k)}$ be the cluster subject i is most likely belongs to, given his biomarker level vector (and given $D_i = 1$). For observations with $D_i = 0$ put $\tilde{Y}_i^{\text{HT}} = 0$. Let $\hat{\mathbf{Z}}_i$ and $\hat{Y}_i^{\text{HT}}$ be the corresponding quantities using the estimates $\hat{\nu}$ of the clustering parameters obtained in the clustering step by the maximization of L_{NM} . The quantity $\hat{Y}_i^{\text{HT}}$ is the final cluster assignment for observation i .

Moving on to the risk factor assessment step, with the final cluster assignments having been made as described previously, estimates of the risk parameters ϕ are obtained as the maximizers of

$$L_{\text{HT}} = \prod_{i=1}^n p_{\hat{Y}_i^{\text{HT}}}(\mathbf{q}_i).$$

In this hard threshold approach, a statistical model is used in the clustering step, but misclassification in the assignment of tumors to clusters is ignored in the risk factor step.

3.2. Semi-hard method

One way to address misclassification is to incorporate the probability of assigning an observation to Cluster k , while it actually belongs to Cluster k' . Denote this quantity by

$\psi_i(k', k) := \Pr \left(\hat{Y}_i^{\text{HT}} = k' | C_i = k \right)$. In the semi-hard method, the clustering step is the same as for the hard thresholding method, while the risk assessment step is carried out by maximizing

$$L_{\text{SH}} = \prod_{i=1}^n \left(1 - \sum_{k=1}^K p_k(\mathbf{q}_i) \right)^{1-d_i} \left(\sum_{k=1}^K p_k(\mathbf{q}_i) \psi_i(\hat{Y}_i^{\text{HT}}, k) \right)^{d_i}. \quad (4)$$

Note that, unlike in the two methods previously presented, here, the biomarker values are used in the risk assessment step and each observation may contribute to the estimator of the risk parameters in more than one cluster.

If only two clusters are considered, $\psi_i(k', k)$ can be simplified. For example,

$$\psi_i(2, 1) = \Pr \left(Z_i^{(2)} > 0.5 | C = 1 \right) \text{ for } k' = 2 \text{ and } k = 1. \text{ Note that in this case } Z_i^{(1)} = 1 - Z_i^{(2)}.$$

For $K = 2$, the distribution of $Z_i^{(2)}$ can be derived as a transformation of a linear combination of independent χ^2 random variables. Technical details are provided in Web Appendix B. If $K > 2$, the distribution of $\hat{Y}^{\text{HT}}$ can be approximated using simple simulations, given the clustering parameters.

3.3. Soft method

This method works with the actual value of each $\mathbf{Z}_i$. The clustering step is carried out in the same way as for the two previous methods. Let $g_k(\mathbf{z})$ be the (multivariate) density of the first $K - 1$ components of $\mathbf{Z}_i$ given $Y_i = k$. We then replace $\psi_i(\hat{Y}_i^{\text{HT}}, k)$ in (4) with $g_k(\mathbf{z}_i)$ and get

$$L_S = \prod_{i=1}^n \left(1 - \sum_{k=1}^K p_k(\mathbf{q}_i) \right)^{1-d_i} \left(\sum_{k=1}^K p_k(\mathbf{q}_i) g_k(\mathbf{z}_i) \right)^{d_i}.$$

This choice allows an observation to contribute more for the estimation of the risk parameters in a cluster if it is more likely to belong to it. The semi-hard method can be seen as discretized version of this method. Under this method, there is no definite assignment of observations to a specific cluster. This approach resembles the concept of fuzzy clustering (s[17]).

3.4. Unified likelihood

We now turn to a unified likelihood approach. This is a one-step procedure for the estimation of all the parameters in the model at once. It can be viewed as a fully Bayesian method for our model with uninformative priors for the model parameters. Recalling that given the subtype membership, the risk factors and the biomarker levels are independent, the full likelihood of the entire data is

$$\begin{aligned}
L &= \prod_{i=1}^n \Pr(D_i=0 | \mathbf{Q}_i=\mathbf{q}_i)^{1-d_i} (\Pr(D_i=1, \mathbf{W}_i=\mathbf{w}_i | \mathbf{Q}_i=\mathbf{q}_i))^{d_i} \\
&= \prod_{i=1}^n \left(1 - \sum_{k=1}^K p_k(\mathbf{q}_i)\right)^{1-d_i} \left(\sum_{k=1}^K \Pr(Y_i=k | \mathbf{Q}_i=\mathbf{q}_i) \Pr(\mathbf{W}_i=\mathbf{w}_i | Y_i=k)\right)^{d_i} \\
&= \prod_{i=1}^n \left(1 - \sum_{k=1}^K p_k(\mathbf{q}_i)\right)^{1-d_i} \left(\sum_{k=1}^K p_k(\mathbf{q}_i) f_k(\mathbf{w}_i)\right)^{d_i}
\end{aligned} \tag{5}$$

where $f_k(\mathbf{w}) = f(\mathbf{w}; \boldsymbol{\mu}_k, \boldsymbol{\Gamma}_k, \boldsymbol{\Omega})$ is the multivariate normal density with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k + \boldsymbol{\Omega}$. In this model, $\boldsymbol{\pi}$ is no longer a free parameter but a function of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and the distribution of $\mathbf{Q}$. As described in Section 3.1, we use a Cholesky decomposition for $\boldsymbol{\Gamma}_k$. Maximum likelihood estimators are obtained by the implementation of a linear-constrained maximization algorithm where the linear constraints are the same as in Section 3.1 with the exception of the conditions on $\boldsymbol{\pi}$ which are no longer needed.

3.5. Remarks

The methods we have presented exploit the assumption that the biomarkers and the risk factors are conditionally independent given the subtype. We further discuss the plausibility of this assumption in Section 8. We also consider the robustness of the methods with respect to this assumption as part of the simulations presented in Section 4 and in the breast cancer data example in Section 7.

Throughout this section we treated K , the number of subtypes, as known. In practice, this is not necessarily the case. A likelihood-based model selection procedure, e.g., based on AIC [18] or BIC [19], can be applied to determine the appropriate number of subtypes. We recommend BIC, as our goal is to identify the true clusters, that is, to select the true model, and BIC is more suited to this goal than AIC, which is more geared to optimal prediction [20].

Because (5) is a regular likelihood function (recall Lemma 1, Web Appendix A), standard likelihood theory implies that the unified likelihood estimators are consistent, asymptotically normal, and asymptotically efficient. We cannot expect all these properties to hold for the other methods under consideration. In particular, methods that involve hard assignment of observations to clusters are bound to be biased and inconsistent as long as there is a non-negligible probability of misclassification. The semi-hard and the soft methods eliminate the bias but are expected to have a larger asymptotic variance than the unified likelihood estimator. The simulation results presented in the next section bear out these statements.

We now expand on the asymptotic properties of the soft method. If $\boldsymbol{\nu}$ were known, then the analogue $\tilde{L}_S$ of L_S using the true $\boldsymbol{\nu}$ is the conditional likelihood of Y_i given D_i , $Z_i(\boldsymbol{\nu})$, and $\mathbf{Q}_i$. The estimate of $\boldsymbol{\phi}$ that we would get using $\tilde{L}_S$ would therefore be consistent, by standard likelihood theory. Accordingly, the estimate $\hat{\boldsymbol{\phi}}$ based on L_S is a consistent estimator of $\boldsymbol{\phi}$. An outline derivation of the asymptotic distribution of the estimate is provided in Web Appendix C.

Although the soft method is not optimally asymptotically efficient, it offers some advantages. In the soft method, the risk factor assessment process deals only with the $K - 1$ numbers $Z_i^{(k)}$, $k = 1, \dots, K - 1$ and not with the entire p -vector of observed biomarker levels. This is an advantage when the number of markers is substantially greater than the number of clusters, as will often be the case. In particular, when p is substantially greater than K , it can be computationally more practicable to carry out the clustering step and the risk factor step separately rather than using the unified likelihood. In addition, with the soft method, a variety of risk factor models can be explored without redefining the cluster structure each time.

4. Simulation results for binary outcome

We present here results of a simulation study involving all the methods described in the preceding section, for the case of $K = 2$ subtypes. The number of markers is taken to be $p = 2, 3, 5$, or 10 . We take $\mu_1 = 0\mathbf{e}_p$, $\mu_2 = 2.5\mathbf{e}_p$, and $\Gamma_1 = \Gamma_2 = \mathbf{I}_p$ where $\mathbf{e}_p$ is a vector of length p all of whose elements are equal to 1, and $\mathbf{I}_p$ is the identity matrix with dimension p . While we take $\Gamma_1 = \Gamma_2$ in generating the simulation data, our estimation procedure does not assume that this equality holds. The measurement error covariance Ω is taken to be a known diagonal matrix with equal values ω along the diagonal. This was performed for simplicity of presentation. In practice, Ω can have non-zero off-diagonal elements with no effect on the theoretical development. Furthermore, the assumption of known Ω can be relaxed to allow for estimation of Ω from a validation study or a design with replicate biomarker measurements. We considered $\omega = 0.25, 0.50, 0.75, 1.5, 2, 3$ which implies correlation between X_j and W_j for each component j , of $0.89, 0.82, 0.76, 0.63, 0.58, 0.50$, respectively. Regarding the risk parameters, we consider a single binary risk factor Q , and we take $Pr(Q = 1) = 0.4$, $\alpha_1 = \alpha_2 = -2.079$, $\beta_1 = 0.405$ ($= \log 1.5$), and $\beta_2 = 1.609$ ($= \log 5$). The resulting mixture probabilities are $\pi = (1/3, 2/3)$ (π is a function of the other model parameters). We take $n = 1000$ observations per dataset. The design outlined previously results in about 300 tumors per dataset. The number of simulation replications was 1000.

For the naive K-means approach, the cluster parameters are estimated in the following way. Following the assignments of each observation with $D_j = 1$ to a cluster, means are estimated using the sample mean in each cluster. Regarding the covariance matrix Γ_k of the true

biomarker levels, let $\hat{\Sigma}_k^{NV}$ be the sample covariance matrix of the observed biomarker levels for subjects assigned to cluster k and let $\hat{\Sigma}_k^{NV} = \hat{\mathbf{V}}_k \hat{\Lambda}_k \hat{\mathbf{V}}_k^T$ be the usual spectral decomposition of $\hat{\Sigma}_k$ with $\hat{\mathbf{V}}_k$ being a matrix that contains the eigenvectors of $\hat{\Sigma}_k$ and $\hat{\Lambda}_k$ being a diagonal matrix with the eigenvalues of $\hat{\Sigma}_k$ as its diagonal. We then estimate Γ_k using $\hat{\Gamma}_k^{NV} = \hat{\mathbf{V}}_k (\hat{\Lambda}_k - \Omega) \hat{\mathbf{V}}_k^T$ and replace negative or near-zero entries in the diagonal of $\hat{\Lambda}_k - \Omega$ with an arbitrary small number. The obtained estimates from the naive K-means are used as initial values for the rest of the methods examined.

The simulations results are presented in Web Tables I and II in Web Appendix D. Figures 1 and 2 present the mean and the standard deviations in the estimation of β_1 and β_2 ,

respectively, under the various methods. The results indicate that the methods that do not carry out a hard assignment of each observation to a specific cluster have significantly lower bias, with a small price being paid in terms of variance. While the unified likelihood approach is theoretically more efficient than the soft and the semi-hard method, in the finite sample cases, we considered the three methods performed comparably.

In the simulation results we have presented, we have excluded outlier simulations, defined as simulations in which $|\hat{\beta}_j| > 5$ for $j = 1, j = 2$ or both, or simulations where optimization failed. The percentage of outliers is generally small, but for the combination of a low number of markers (2) and large measurement error ($\omega > 1.5$) the percentage of outliers is more substantial. See Web Appendix E for more details. The methods we have presented assume that the biomarker vector $\mathbf{X}$ follows a multivariate normal distribution. While this assumption may be plausible in practice, the robustness of our methods to this assumption is of interest. Web Tables 6 and 7 of Web Appendix F present simulation results for our methods when the assumed distribution of the biomarkers is multivariate normal, but the correctly-measured biomarker vector actually follows a multivariate t distribution, while the other model components, including the normal measurement error, are left unchanged. A marginal increase, if at all, was observed for the bias while a minor increase of about 5% was observed in the standard deviations. In practice, the assumed multivariate distribution can be replaced by another parametric distribution, where a notable change is that $f_k(\mathbf{w})$ becomes a convolution of the assumed distribution for $\mathbf{X}$ with the normal distribution of the measurement error. While the distribution of Z in the case of two clusters is not as shown in Web Appendix B, the associated density can be computed by simulation.

As discussed before, the robustness of the methods to the assumption of conditional independence of the biomarkers and the risk factors is also of interest. We thus considered the following two scenarios where this assumption does not hold. In the first scenario, we increased the mean of all biomarkers by 0.1 whenever $Q = 1$, and for both clusters. For example, for an observation belonging to Cluster 2, with $Q = 1$, the mean was taken to be 2.6 for all biomarkers. In the second scenario, we increased the value of the first biomarker only by 0.5 whenever $Q = 1$, and for both clusters. The results, presented in Web Tables 8–11 of Web Appendix F suggest that for a small number of biomarkers and more than minimal measurement error, the added bias is substantial. The increased bias was greater for the unified likelihood approach than for the soft method. For both scenarios, the bias was greatly reduced as the number of biomarkers grew. We further consider this assumption and its plausibility in Sections 7 and 8.

5. Survival data model

The proposed unified likelihood approach can be extended to models other than the one described in Section 3. In this section, we present a partial likelihood approach for survival data subject to right censoring and left truncation. The proposed partial likelihood is of a form similar to the Cox partial likelihood [21] and is derived using a similar argument. In what follows, we first recall relevant notation from Section 2, with adjustments made for the survival data case. We then present our unified partial likelihood approach. A similar extension can be carried out for the the soft approach.

As before, denote the true biomarker level by $\mathbf{X}$ and the observed level by $\mathbf{W}$, where $\mathbf{W} = \mathbf{X} + \mathbf{U}$ and $\mathbf{U}$ is measurement error vector independent of $\mathbf{X}$ with the multivariate normal distribution $N_p(0, \mathbf{\Omega})$. The vector $\mathbf{X}$ again follows a mixture distribution with components $N(\boldsymbol{\mu}_k, \mathbf{\Gamma}_k)$. Recall that $Y = 0$ for no disease and $Y = k$ for disease of subtype k .

The observed data for each individual i is $(T_i, \mathbb{D}_i, \mathbf{Q}_i, \mathbf{W}_i)$ where T_i is the observed follow-up time, $\mathbb{D}_i$ is the event indicator, $\mathbf{Q}_i$ is the value of a risk factor vector, possibly time-dependent (and then, it is denoted by $\mathbf{Q}_i(t)$), and $\mathbf{W}_i$ is the marker levels vector, observed if $\mathbb{D}_i = 1$ and otherwise regarded as missing. Note that because an event of only one type can occur for each subject, we have a competing risk model. See [22] for a basic introduction to competing risk models.

Let $\lambda_k(t, \mathbf{q}(t))$ be the crude cause-specific hazard function for the k -th cluster when $\mathbf{Q}(t) = \mathbf{q}(t)$, that is,

$$\lambda_k(t, \mathbf{q}(t)) = \lim_{\Delta \downarrow 0} \Pr(T \in [t, t + \Delta], Y = k | T \geq t, \mathbf{Q}(t) = \mathbf{q}(t)).$$

We use a proportional hazard model for the crude hazard function for each subtype k ,

$$\lambda_k(t, \mathbf{q}(t)) = \lambda_{0k}(t) \exp \left(\boldsymbol{\beta}_k^T \mathbf{q}(t) \right). \quad (6)$$

This is the model discussed in Section 3.3.1 of [22]. We write the cause-specific baseline hazard function as $\lambda_{0k}(t) = \lambda_{0K}(t) h_k(t)$, where $h_k(t) = \lambda_{0k}(t) / \lambda_{0K}(t)$ is the baseline hazard ratio for Cluster k relative to Cluster K and with $h_K(t) = 1$. We work with a parametric model for $h_k(t)$; i.e., we take $h_k(t) = h(t; \boldsymbol{\eta}_k)$ for a vector of unknown parameters $\boldsymbol{\eta}_k$ ($k = 1, \dots, K - 1$). We leave $\lambda_{0K}(t)$ unspecified, as in the classical proportional hazard model. Similar to Cox's partial likelihood, $\lambda_{0K}(t)$ cancels out of our partial likelihood; the hazard ratio functions $h_k(t)$, however, do not. Our formulation is more flexible than assuming a parametric model for all of the baseline hazard functions $\lambda_{0k}^\eta(t)$, $k = 1, \dots, K$.

The density function for $\mathbf{W}_i$ given that subject i belongs to Cluster k is denoted by $f(\mathbf{w}; \boldsymbol{\nu}_k)$, where $\boldsymbol{\nu}_k$ is a vector containing the parameters of the relevant normal distribution. That is,

$$f(\mathbf{w}; \boldsymbol{\nu}_k) = \sum_k^{-1/2} \phi \left(\sum_k^{-1/2} (\mathbf{w} - \boldsymbol{\mu}_k) \right). \text{ Let } \boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K), \boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_K), \text{ and } \boldsymbol{\nu} = (\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_K). \text{ Finally, denote the vector of all unknown parameters by } \boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\nu}).$$

Let $\xi_i(t)$ be equal 1 if person i is at risk at time t and 0 otherwise, and define the event counting process in the usual way: $N_i(t) = \mathbb{D}_{\{T_i \leq t\}}$. Let $\mathbf{W}_i(t)$ be the process defined as $\mathbf{W}_i$ if $N_i(t) = 1$, and arbitrarily defined as some constant when $N_i(t) = 0$. We introduce the filtration $\mathcal{F}_t = \sigma(\xi_i(u), \mathbf{Q}_i(u), N_i(u), \mathbf{W}_i(u), i = 1, \dots, n, u \in [0, t])$. We denote the maximum follow-up time by τ and use the notation $\mathcal{F}$ to denote the filtration as a whole, i.e., $\mathcal{F} = \{\mathcal{F}_t, t \in [0, \tau]\}$.

We now construct our partial likelihood. Consider a small interval of time $[t - \Delta, t]$, where Δ is small. Let $\mathcal{E}(t, \Delta)$ denote the set of study realizations in which an event of some type occurred to some individual in $[t - \Delta, t]$, and let $\mathcal{O}(t, \Delta, i, k, \mathbf{w})$ denote the subset of $\mathcal{E}(t, \Delta)$ in which subject i is the one who suffered the event, the event is of type k , and the ensuing biomarker vector is $\mathbf{w}$. That is, $\mathcal{O}(t, \Delta, i, k, \mathbf{w}) = \{T_i \in [t - \Delta, t], Y_i = k, \mathbf{W}_i = \mathbf{w}\}$, and $\mathcal{E}(t, \Delta) = \bigcup_{i=1}^n \bigcup_{k=1}^K \mathcal{O}(t, \Delta, i, k, \mathbf{w})$. Similar to the derivation of the Cox partial likelihood [21], we work with the conditional likelihood of $\mathcal{O}(t, \Delta, i, k, \mathbf{w})$ given $\mathcal{F}_{t-\Delta}$ and $\mathcal{E}(t, \Delta)$. We can write

$$\Pr(\mathcal{O}(t, \Delta, i, k, \mathbf{w}) | \mathcal{F}_{t-\Delta}) \doteq \xi_i(t - \Delta) \lambda_{0k}^\eta(t - \Delta) \exp \left(\beta_k^T \mathbf{Q}_i(t - \Delta) \right) f_k(\mathbf{w}) \Delta.$$

where the joint distribution of $(T_i, Y_i, \mathbf{W}_i)$ given the history was expressed in terms of the joint distribution of (T_i, Y_i) given the history and the distribution of $\mathbf{W}_i$ given Y_i exploiting the assumption of conditional independence of the biomarkers with other quantities. Accordingly,

$$\Pr(\mathcal{E}(t, \Delta) | \mathcal{F}_{t-\Delta}) \doteq \left[\sum_{j=1}^n \xi_j(t - \Delta) \sum_{k=1}^K \lambda_{0k}^\eta(t - \Delta) \exp \left(\beta_k^T \mathbf{Q}_j(t - \Delta) \right) \right] \Delta.$$

Thus,

$$\begin{aligned} \Pr(\mathcal{O}(t, \Delta, i, k, \mathbf{w}) | \mathcal{F}_{t-\Delta}, \mathcal{E}(t - \Delta)) &\doteq \frac{\xi_i(t - \Delta) \lambda_{0k}^\eta(t - \Delta) \exp \left(\beta_k^T \mathbf{Q}_i(t - \Delta) \right) f_k(\mathbf{w})}{\sum_{j=1}^n \xi_j(t - \Delta) \sum_{k'=1}^K \lambda_{0k'}^\eta(t - \Delta) \exp \left(\beta_{k'}^T \mathbf{Q}_j(t - \Delta) \right)} \\ &= \frac{\xi_i(t - \Delta) h(t - \Delta; \boldsymbol{\eta}_k) \exp \left(\beta_k^T \mathbf{Q}_i(t - \Delta) \right) f_k(\mathbf{w})}{\sum_{j=1}^n \xi_j(t - \Delta) \sum_{k'=1}^K h(t - \Delta; \boldsymbol{\eta}_{k'}) \exp \left(\beta_{k'}^T \mathbf{Q}_j(t - \Delta) \right)} \end{aligned}$$

Now, let T_r^* , $r = 1, \dots, R$ denote the ordered observed event times, H_r the index of the person who suffered the event at time T_r^* , Y_r^* be the (unobserved) cluster index for the event that occurred at time T_r^* , and $\mathbf{W}_r^*$ the biomarker vector for the event that occurred at time T_r^* . Let $\mathcal{E}_r$ represent the occurrence of an event of some type at time T_r^* . From the calculations previously, we obtain (letting Δ go to 0)

$$\begin{aligned} \Pr \left(H_r = \kappa_r, \mathbf{W}_r^* = \mathbf{w}_r^* | \mathcal{F}_{T_r^* -}, \mathcal{E}_r \right) &= \frac{\sum_{k=1}^K \Pr \left(Y_r^* = k, H_r = \kappa_r, \mathbf{W}_r^* = \mathbf{w}_r^* | \mathcal{F}_{T_r^* -}, \mathcal{E}_r \right)}{\sum_{j=1}^n \xi_j(T_r^*) \sum_{k=1}^K \Pr \left(Y_r^* = k, H_r = j | \mathcal{F}_{T_r^* -}, \mathcal{E}_r \right)} \\ &= \frac{\sum_{k=1}^K h(T_r^*; \boldsymbol{\eta}_k) \exp \left(\beta_k^T \mathbf{Q}_{h_r}(T_r^*) \right) f_k(\mathbf{w}_r^*)}{\sum_{j=1}^n \xi_j(T_r^*) \sum_{k=1}^K h(T_r^*; \boldsymbol{\eta}_k) \exp \left(\beta_k^T \mathbf{Q}_j(T_r^*) \right)} \end{aligned}$$

and the partial likelihood is then

$$L(\boldsymbol{\theta}) = \prod_{r=1}^R \frac{\sum_{k=1}^K h(T_r^*; \boldsymbol{\eta}_k) \exp \left(\boldsymbol{\beta}_k^T \mathbf{Q}_{h_r}(T_r^*) \right) f_k(\mathbf{W}_r^*)}{\sum_{j=1}^n \xi_j(T_r^*) \sum_{k=1}^K h(T_r^*; \boldsymbol{\eta}_k) \exp \left(\boldsymbol{\beta}_k^T \mathbf{Q}_j(T_r^*) \right)}. \quad (7)$$

We can write L in the following alternate form:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n \left[\frac{\sum_{k=1}^K h(T_i; \boldsymbol{\eta}_k) \exp \left(\boldsymbol{\beta}_k^T \mathbf{Q}_i(T_i) \right) f_k(\mathbf{W}_i)}{\sum_{j=1}^n \xi_j(T_i) \sum_{k=1}^K h(T_i; \boldsymbol{\eta}_k) \exp \left(\boldsymbol{\beta}_k^T \mathbf{Q}_j(T_i) \right)} \right]^{\mathbb{D}_i}. \quad (8)$$

Maximum (partial) likelihood estimator for $\boldsymbol{\theta}$ is obtained by maximizing (8) under the same linear constraints as described in Section 3. In particular, the constraints impose model identifiability and positive definiteness of the covariance matrix estimates $\hat{\boldsymbol{\Gamma}}_k$, $k = 1, \dots, K$. Generalizing the hard, semi-hard and soft method to the time-to-event design is straightforward. The clustering step utilizes L_{NM} defined in (2), and is identical to the binary data case, and the risk assessment is carried out by replacing $f_k(\mathbf{W}_j)$ in the denominator of (8) with the appropriate $g_k(\mathbf{Z}_j)$ and then maximizing the resulting expression as a function of the risk parameters.

Under suitable regularity conditions, the maximum (partial) likelihood estimator using either the unified likelihood or the soft methods are consistent and asymptotically normally distributed with covariance matrix that can be estimated from the data. Consistency is shown by noting that the score vector $\mathbf{u}(\boldsymbol{\theta})$ corresponding to $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ converges to a limit $\mathbf{u}^*(\boldsymbol{\theta})$ which is equal to zero at the true value of $\boldsymbol{\theta}$. Asymptotic normality is shown by an argument along the lines of [23]. Proofs are provided in Web Appendix G. For the extension of the soft method of Subsection 3.3 to the survival setting, the asymptotic properties can be established by combining the argument in Web Appendix G with the asymptotic argument made for the soft method in Web Appendix C.

In the next section, we present simulation results for the unified partial likelihood approach and the soft method and compare them with a naive K-means approach for survival data. And then, in Section 7, we present a case study involving breast cancer data from the NHS.

6. Simulation results for survival data

For the simulation design, we take $p = 3$ markers, $K = 2$ subtypes and the same clustering parameters, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Gamma}_1$, and $\boldsymbol{\Gamma}_2$, as in the simulation for the binary data. The matrix $\boldsymbol{\Omega}$ is again a diagonal matrix with equal values ω along the diagonal. Q is taken to be a single binary covariate with $P(Q = 1) = 0.4$. Regarding the risk factor effects, we take $\beta_1 = 0.405$ ($= \log 1.5$) and $\beta_2 = 1.609$ ($= \log 5$).

We take the baseline hazard for each subtype k to be of parametric form, and more specifically Weibull with parameters m_k and γ_k . That is,

$$\lambda_{0k}(t) = \frac{m_k}{\gamma_k} \left(\frac{t}{\gamma_k} \right)^{m_k-1}$$

which implies that $h(t, \boldsymbol{\eta}) = \eta_1 t^{\eta_2}$ with the appropriate η_1 and η_2 . For creating the dataset, we take the values $\gamma_1 = 175$, $\gamma_2 = 110$, $m_1 = 1.5$, and $m_2 = 2.5$ and accordingly $\eta_1 \approx 0.03$ and $\eta_2 = 1$. We also add an exponential right-censoring to the data (mean=50) and censor all observation at $t = 90$ if no event or censoring has occurred before. In total, we take 1000 observations per dataset. The design outlined previously results in about 300 events per data set, 100 for $k = 1$ and 200 for $k = 2$. For each of the datasets created, we estimate the parameters using the naive K-means approach and using the maximum partial likelihood described in Section 5, as well as the soft method appropriately modified for survival data. The number of simulation replications was 1000.

In implementing the naive K-means method, we first used the K-means algorithm to cluster observations with $\mathbb{D} = 1$ into two clusters. With the clusters thus formed, the means in each cluster ($\boldsymbol{\mu}_k$) were estimated using the within-cluster sample means, and the covariance matrices were estimated in the same way as described in Section 4 for the binary data case. The risk assessment step was then carried out by fitting a Weibull regression model to a modified version of the data using the ‘phreg’ function in the ‘eha’ package in **R** [24]. For each subtype k , observations with $\hat{Y}^{NV} = k$ were taken to be censored and estimates for γ_k , m_k and β_k were obtained using a Weibull regression model for censored data. Thereafter, estimates for η_1 and η_2 were calculated by substituting in the estimates for γ_1 , γ_2 , m_1 and m_2 .

The partial likelihood estimates were obtained by maximizing (7) under linear constraints similar to those described for the binary data. The identifiability constraints (after reparameterization) and the constraint $\eta_1 > 0$ are simply lower bounds for the model parameters. Therefore, a standard quasi-Newton optimization algorithm can be implemented. We used the ‘constrOptim’ function in **R**. In cases the algorithm did not converge, the slower Nelder-Mead algorithm was used instead (about 5% of cases). We eliminated less than 5% of the original 1000 iterations per simulation scenario (per ω value), according to the following criteria: The Nelder-Mead algorithm did not converge; Weibull fitting failed to converge. See Web Appendix E for more details. Simulation results are presented in Web Table III of Web Appendix D for the unified (partial) likelihood, the soft and the naive K-means methods. Both the unified (partial) likelihood and the soft methods appear to outperform the naive K-means approach in terms of bias. The unified (partial) likelihood seems to have much better performance in terms of bias comparing with the soft method, especially for large measurement error. The bias when using the soft method should vanish as the sample size grows.

7. Analysis of nurses’ health study data

We applied the methods we have described on data from the NHS [2]. The NHS began in 1976 when 121,700 female nurses aged 30–55 returned a questionnaire about their lifestyle and their health. Participants answered a biennial questionnaire and their answers identified

the occurrence of various clinical events and formed the vector of time dependent covariates. The clinical event we considered here is incidence of breast cancer. Tumor tissue was stored, and later several biomarker levels were measured. In our analysis, we considered the following tumor biomarkers: **ER**, **PR**, and **Ki67** protein. Previous research identified breast cancer subtypes using these biomarkers (cf. [3]). The biomarkers were assessed using the **Definiens** software as the proportion of nuclei that fell above the Immunohistochemistry marker threshold for positive cells on the tumor core of the total number of nuclei detected.

Some individuals have two or three repeated measurements of at least one biomarker. Measurement error arises here from two sources. The first source is tumor tissue heterogeneity, because the repeated measures were taken from different tumor cores located in different places in the tumor. The second source is the technical laboratory error of the measurement. However, we model the pooled measurement error arising from both sources as a single source.

Because breast cancer exhibits different properties depending on whether it occurs pre-menopause or post-menopause, we restricted our analysis to post-menopausal records only. The risk factor we considered was BMI, expressed in terms of three categories: Normal or Low weight ($\text{BMI} < 25$), Overweight (BMI between 25 and 29) and Obesity ($\text{BMI} > 29$). We also include various known confounders in the model. In the succeeding discussion, we list these confounders, and their definitions.

- **Parity and Age at 1st Birth:** Categorical variable describing parity and age at first birth. Reference category is 'Nulliparous'; the rest of the categories are crossproduct of parity and age at first birth categories.
- **Age at Menar:** Categorical variable, age at menarche (' ≤ 12 ', ' $= 13$ ', ' ≥ 14 '). Reference category ' ≤ 12 '.
- **Alcohol:** Binary variable, alcohol drinking.
- **Postmenopausal hormon use:** Categorical variable, postmenopausal hormon use ('never used', 'current user', 'past user'). Reference category is 'never used'.

We used the missing-indicator method for categorical variables. We used records from the first eight questionnaire cycles. In total, our analysis included 556,472 records, representing 93,104 distinct individuals. This cohort includes 683 breast cancer cases.

In order to improve the fit of the normal model to the biomarker data, we applied an arcsine-root transformation for the biomarkers' values. This transformation is helpful when the observed values are between 0 and 1, and are relatively concentrated in the edges, as appeared to be the case here. After the transformation, the biomarker measurements were normalized to have mean 0 and standard deviation 1. The measurement error covariance matrix $\mathbf{\Omega}$ was assumed to be diagonal, and was estimated by running a random effects ANOVA model on the replicate biomarker data (using the **R** package 'nmle'[25]). The resulting estimated measurement error variances were 0.20 for **ER**, 0.10 for **PR**, and 0.21 for **Ki67**. Because we have replicate measurements of the biomarker levels, the model described in

Section 5 is slightly modified. However, no additional parameters are introduced, and the technical changes, given in Web Appendix H, apply only to $f_k(\mathbf{w})$.

We considered a model with two clusters. Initial estimates for the parameters were obtained using the naive K-means approach. The implementation of the K-means algorithm on the biomarkers values led to partition of the 683 cases into two groups, with 311 cases in Cluster 1 and 372 cases in Cluster 2. Means and covariance matrices were then estimated as described in Section 4.

In constructing the partial likelihood for the risk factor model, we stratified by questionnaire cycle ($S = 1, \dots, 8$). The unified partial likelihood (8) then becomes

$$L(\boldsymbol{\theta}) = \prod_{s=1}^S \prod_{i=1}^{n_s} \left[\frac{\sum_{k=1}^2 h_s(T_{si}; \boldsymbol{\eta}_{sk}) \exp \left(\boldsymbol{\beta}_k^T \mathbf{Q}_{si}(T_{si}) \right) f_k(\mathbf{W}_{si})}{\sum_{j=1}^{n_s} \xi_{sj}(T_{si}) \sum_{k=1}^2 h(T_{si}; \boldsymbol{\eta}_{sk}) \exp \left(\boldsymbol{\beta}_k^T \mathbf{Q}_{sj}(T_{si}) \right)} \right]^{\mathbb{D}_{si}} \quad (9)$$

with stratum-specific definitions for all relevant quantities (parameters, random variables, and values). For the baseline hazard ratio, we took $h_s(t) = h_s$, that is, a different constant for each stratum.

Initial estimates for h_s , $s = 1, \dots, 8$ and for the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were obtained as follows. We temporarily converted all events in Cluster 2 to censored records, while events from Cluster 1 and non-event records were left untouched. We then ran a stratified exponential regression under the counting process framework taking the Cluster 1 cases as events. We then carried out a similar exponential regression analysis with the roles of the clusters reversed, so that Cluster 2 cases were regarded as events and Cluster 1 cases were regarded as censored observations. Initial estimates for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were obtained from these exponential regression analyses. Initial estimates for h_s were obtained by taking the ratio of the estimated constant baseline hazard ratio in each regression, for each stratum separately.

We first implemented the procedure we just described on a sample of the data, containing all 683 cases and a stratified sample of 7,000 non cases. For this sampled data, the naive K-means estimates were taken as initial values in an optimization algorithm aimed at maximizing the partial likelihood (9) under linear constraints that ensure identifiability of the model and positive definiteness of $\hat{\boldsymbol{\Gamma}}_1$ and $\hat{\boldsymbol{\Gamma}}_2$. We also considered at this point the option of more than two clusters. Using this sample of 7,683 individuals, BIC directed us to the model with two clusters (BIC values of 16,295, 16,644, and 16,956 for $K = 2$, $K = 3$ and $K = 4$, respectively). We then took estimates obtained and used them as initial values for optimization using the full data. In the succeeding discussions, we compare between results obtained by the naive K-means and the unified partial likelihood approaches, both implemented for the full data.

Table I presents medians and interquartile ranges (IQRs) in the markers original scales. Looking at this table, we see that **PR** median estimates obtained using our method are smaller. IQR estimates using our method are larger for **ER** and **PR** in Cluster 1 and for **PR** in Cluster 2. Table II displays Kendall's τ for each pair of biomarkers. We use Kendall's τ

correlation because it is invariant to transformations. Our results implies moderate correlation between **PR** and **ER** in both clusters, while correlations using K-means estimates are negative for cluster 2 (−0.27) and positive in Cluster 1 (0.25). There are also apparent differences in the **PR/Ki67** correlation estimates. Inference for β_k in the naive K-means methods was carried out under the standard exponential regression model, ignoring the possibility of misclassification. The partial likelihood estimators are consistent and normally distributed (Web Appendix G). The variance of $\hat{\beta}$ was estimated using an extension of the sandwich formula given in Web Appendix G that allows for repeated measurements of biomarkers and a stratified hazard model.

Table III presents β estimates using both methods. Our main concern here is the **BMI** effect. In Cluster 1, **BMI** did not have a significant effect (hazard ratios 0.94 and 0.85, p -values 0.94 and 0.41, respectively) and in Cluster 2, Overweight increases this risk (hazard ratio 1.39, p -value < 0.01) while Obesity did not have a significant effect (hazard ratio 1.06, p -value 0.68).

In this data analysis example, we have assumed that the biomarkers are conditionally independent of the risk factors, given the true subtypes. As a sensitivity analysis, we redid the analysis three additional times, each time allowing the mean of a different biomarker (**ER**, **PR**, or **Ki67**) to depend not only on the subtype, but also on the categorical variable of interest, **BMI**. The covariance matrix was kept the same for all three **BMI** levels. The sensitivity analyses were carried out by replacing $f_k(\mathbf{w})$ in (9) with $f_{k,q}(\mathbf{w})$, where q is the **BMI** level (1,2, or 3), and $f_{k,q}(\mathbf{w})$ is the multivariate normal density with mean $\mu_{k,q}$ and covariance matrix Σ_k . In each reanalysis, we allowed the mean of a different biomarker to depend on the **BMI** level by setting the j -th component of $\mu_{k,q}$ to be equal to $\mu_{kj} + \rho_q$, for $k = 1, 2$ and $q = 1, 2, 3$. We took ρ_1 equal to zero for identifiability. When we carried out this modification, most estimates did not change substantially. In particular, the coefficients for the effect of Overweight and Obesity originally estimated as $\{-0.01, 0.16\}$ in Cluster 1, were estimated as $\{0.03, -0.16\}$, $\{0.09, -0.15\}$, and $\{-0.01, -0.16\}$ when allowed the mean of **ER**, **PR**, and **Ki67**, respectively, to depend on **BMI**. In Cluster 2, the corresponding original estimates $\{0.33, 0.06\}$ were slightly changed to $\{0.31, 0.06\}$, $\{0.28, 0.06\}$, and $\{0.33, 0.06\}$ when we allowed the mean of **ER**, **PR**, and **Ki67**, respectively, to depend on **BMI**.

8. Discussion

We have developed a new approach to risk factor analysis of cancer data when the risk factor effects can vary across tumor subtypes. Although the subtypes are not known *a priori*, measured biomarker levels aid the researcher in identifying the subtypes, with the parameters of the biomarker distribution depending on the subtype. Here, it was assumed the biomarker level vector is multivariate normal.

A common practice is to carry out a hard classification of observations to clusters and then estimate risk factor effects in each cluster separately. However, because of misclassification, this approach yields inconsistent estimators. Under a multinomial logistic model for the risk factor effects on the tumor subtypes, we described a unified likelihood approach that

accounts for the natural biological variability and measurement error in the biomarker data. We also considered two intermediate two-step methods that provide consistent estimators, a semi-hard method that carries out a hard classification but incorporates misclassification probabilities into the parameter estimation, and a soft method that avoids a hard classification and instead works with cluster membership probabilities (fuzzy clustering). Simulation results indicated that, in comparison with the common approach taken in practice, our proposed approaches lead to estimates with greatly reduced bias, with a small price being paid in terms of variance.

As between the two consistent two-stage estimators, the semi-hard and soft estimators, the soft method makes more precise use of the data and is therefore generally to be preferred.

By standard likelihood theory, the unified likelihood method is asymptotically optimally efficient, whereas the soft method is not. On the other hand, the soft method offers some advantages, as we described in Section 3.5. In light of our simulation results, the soft method should be preferred over the unified likelihood method, when the sample size is low to moderate and the measurement error is small. The soft method could also be used to provide initial values for the unified likelihood method. When the measurement error is large, the soft method can be used as an approximate method in order to investigate expeditiously certain questions of interest, for example covariate selection for the regression model.

Our methods rely on the assumption that given the subtype membership, the risk factors are not informative regarding the biomarker. While we think this assumption is often plausible, its soundness should be assessed in a specific application. As we demonstrated in the breast cancer data example, the unified likelihood method can be extended to cover this dependence, although it may be impractical to fit more complicated models because of the large number of parameters needed. If the non-informativeness assumption is violated, then any method that attempts to incorporate dependence between the biomarkers and the risk factors is bound to be complicated and to include a large number of parameters. Regarding the suggested two-stage methods, namely semi-hard and soft methods, these methods cannot be extended directly to cover this dependence, because of the separation of the cluster step and the risk factor step. If one suspects that a specific side variable is related to both the biomarker and the risk factor, a practical approach is to adjust the biomarker for the side variable prior to the clustering step and then to add the side variable as a background covariate in the risk factor step.

We also considered a survival data model (for right-censored and left-truncated survival data) in which the cluster-specific crude hazard functions are taken to be of proportional hazards form. We developed a unified partial likelihood approach with an appropriately-defined partial likelihood function. This partial likelihood approach allows for time-dependent covariates. We also presented the appropriate implementation of the soft method for survival data. The resulting estimators were shown to be consistent and asymptotically normal. Simulation results yielded the same conclusions as in the multinomial case. In the analysis of the NHS breast cancer data, we have shown how the partial likelihood approach can be applied to estimate risk factor effects in different cancer subtypes. A possible

extension of our approach would be to combine our method with variable selection procedure in order to identify new subtypes when many biomarkers are available.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Althuis MD, Fergenbaum JH, Garcia-Closas M, Brinton LA, Madigan MP, Sherman ME. Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. *Cancer Epidemiology Biomarkers & Prevention*. 2004; 13(10):1558–1568.
2. Tamimi RM, Colditz GA, Hazra A, Baer HJ, Hankinson SE, Rosner B, Marotti J, Connolly JL, Schnitt SJ, Collins LC. Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. *Breast Cancer Research and Treatment*. 2012; 131(1):159–167. [PubMed: 21830014]
3. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, Gaudet M, Schmidt MK, Broeks A, Cox A, Fasching PA, Hein R, Spurdle AB, Blows F, Driver K, Flesch-Janys D, Heinz J, Sinn P, Vrieling A, Heikkinen T, Aittomäki K, Heikkilä P, Blomqvist C, Lissowska J, Peplonska B, Chanock S, Figueroa J, Brinton L, Hall P, Czene K, Humphreys K, Darabi H, Liu J, Van't Veer LJ, van Leeuwen FE, Andrulis IL, Glendon G, Knight JA, Mulligan AM, O'Malley FP, Weerasooriya N, John EM, Beckmann MW, Hartmann A, Weihbrecht SB, Wachter DL, Jud SM, Loehberg CR, Baglietto L, English DR, Giles GG, McLean CA, Severi G, Lambrechts D, Vandrope T, Weltens C, Paridaens R, Smeets A, Neven P, Wildiers H, Wang X, Olson JE, Cafourek V, Fredericksen Z, Kosel M, Vachon C, Cramp HE, Connley D, Cross SS, Balasubramanian SP, Reed MW, Dörk T, Bremer M, Meyer A, Karstens JH, Ay A, Park-Simon TW, Hillemanns P, Arias Pérez JJ, Menéndez Rodríguez P, Zamora P, Benítez J, Ko YD, Fischer HP, Hamann U, Pesch B, Brüning T, Justenhoven C, Brauch H, Eccles DM, Tapper WJ, Gerty SM, Sawyer EJ, Tomlinson IP, Jones A, Kerin M, Miller N, McInerney N, Anton-Culver H, Ziogas A, Shen CY, Hsiung CN, Wu PE, Yang SL, Yu JC, Chen ST, Hsu GC, Haiman CA, Henderson BE, Le Marchand L, Kolonel LN, Lindblom A, Margolin S, Jakubowska A, Lubinski J, Huzarski T, Byrski T, Górski B, Gronwald J, Hooning MJ, Hollestelle A, van den Ouweland AM, Jager A, Kriege M, Tilanus-Linthorst MM, Collée M, Wang-Gohrke S, Pylkäs K, Jukkola-Vuorinen A, Mononen K, Grip M, Hirvikoski P, Winqvist R, Mannermaa A, Kosma VM, Kauppinen J, Kataja V, Auvinen P, Soini Y, Sironen R, Bojesen SE, Ørsted DD, Kaur-Knudsen D, Flyger H, Nordestgaard BG, Holland H, Chenevix-Trench G, Manoukian S, Barile M, Radice P, Hankinson SE, Hunter DJ, Tamimi R, Sangrajrang S, Brennan P, McKay J, Odefrey F, Gaborieau V, Devilee P, Huijts PE, Tollenaar RA, Seynaeve C, Dite GS, Apicella C, Hopper JL, Hammet F, Tsimiklis H, Smith LD, Southey MC, Humphreys MK, Easton D, Pharoah P, Sherman ME, Garcia-Closas M. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. *Journal of the National Cancer Institute*. 2011; 103(3):250–263. [PubMed: 21191117]
4. McLachlan, G., Peel, D. *Finite Mixture Models*. John Wiley & Sons; New York: 2004.
5. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*. 2002; 18(3):413–422. [PubMed: 11934740]
6. Raftery AE, Dean N. Variable selection for model-based clustering. *Journal of the American Statistical Association*. 2006; 101(473):168–178.
7. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. *Bioinformatics*. 2003; 19(9):1061–1069. [PubMed: 12801866]
8. Maugis C, Celeux G, Martin-Magniette ML. Variable selection for clustering with gaussian mixture models. *Biometrics*. 2009; 65(3):701–709. [PubMed: 19210744]
9. Pan W, Shen X. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*. 2007; 8:1145–1164.
10. Witten DM, Tibshirani R. A framework for feature selection in clustering. *Journal of the American Statistical Association*. 2010; 105(490):713–726. [PubMed: 20811510]

11. Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: a review. *Computational Statistics & Data Analysis*. 2014; 71:52–78.
12. Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*. 2004; 6(1):90–105.
13. Kriegel HP, Kröger P, Zimek A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2009; 3(1):1–58.
14. MacQueen, J. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. Oakland, CA, USA: 1967. Some methods for classification and analysis of multivariate observations; p. 281–297.
15. Pinheiro JC, Bates DM. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*. 1996; 6(3):289–296.
16. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1977; 39:1–38.
17. Yang MS. A survey of fuzzy clustering. *Mathematical and Computer Modelling*. 1993; 18(11):1–16.
18. Akaike H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*. 1974; 19(6):716–723.
19. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6(2):461–464.
20. Shao J. An asymptotic theory for linear model selection. *Statistica Sinica*. 1997; 7:221–242.
21. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1972; 34(2):187–220.
22. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*. 2007; 26(11):2389–2430. [PubMed: 17031868]
23. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*. 1989; 84(408):1074–1078.
24. Broström, G. eha: Event history analysis. 2014. <http://CRAN.R-project.org/package=eha>, R package version 2.4-1
25. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. R Core Team. nlme: Linear and nonlinear mixed effects models. 2014. <http://CRAN.R-project.org/package=nlme>, R package version 3.1-118

Figure 1.

Means and standard deviations of β_1 estimates in simulated data. The black solid horizontal line represents the true value $\beta_1 = 0.405$.

Figure 2.

Means and standard deviations of β_2 estimates in simulated data. The black solid horizontal line represents the true value $\beta_2 = 1.609$.

Table 1

Estimated medians and IQRs.

	Cluster 1			Cluster 2		
	ER	PR	Ki67	ER	PR	Ki67
Naïve K-means	Median	0.42	0.10	0.07	0.53	0.68
	IQR	0.72	0.58	0.21	0.70	0.64
Unified partial likelihood	Median	0.46	0.06	0.09	0.61	0.62
	IQR	0.78	0.56	0.26	0.69	0.71

IQR An asymptotic theory for linear model selections, interquartile ranges; ER, estrogen receptors; PR, progesterone receptors.

Table IIEstimated Kendall's τ .

		Cluster 1	Cluster 2
Naive K-means	ER/PR	0.25	−0.27
	ER/Ki67	−0.06	0.09
	PR/Ki67	−0.11	−0.08
Unified partial likelihood	ER/PR	0.44	0.42
	ER/Ki67	0.04	0.10
	PR/Ki67	0.19	0.12

ER, estrogen receptors; PR, progesterone receptors.

Table III

β estimates when using naive K-means and unified partial-likelihood methods.

	Naive K-means			Unified partial-likelihood			
	Estimate(exp)	p	CI(95%)	Estimate(exp)	p	CI(95%)	
Cluster 1	Parity 1-4 & AgeAt1stBirth < 25	-0.22 (0.80)	0.37	(-0.70, 0.26)	-0.22 (0.80)	0.43	(-0.77, 0.33)
	Parity 1-4 & AgeAt1stBirth 25 - 30	-0.02 (0.98)	0.93	(-0.51, 0.46)	-0.09 (0.91)	0.74	(-0.63, 0.45)
	Parity 1-4 & AgeAt1stBirth > 30	0.17 (1.19)	0.54	(-0.37, 0.72)	0.03 (1.03)	0.94	(-0.59, 0.64)
	Parity > 4 & AgeAt1stBirth	-0.05 (0.95)	0.87	(-0.61, 0.51)	-0.19 (0.83)	0.56	(-0.84, 0.46)
	Parity > 4 & AgeAt1stBirth > 25	0.23 (1.26)	0.45	(-0.34, 0.82)	-0.05 (0.95)	0.89	(-0.74, 0.65)
	Parity or AgeAt1stBirth missing	-0.03 (0.97)	0.95	(-0.89, 0.83)	0.11 (1.11)	0.82	(-0.83, 1.05)
	AgeAtMenar = 13	0.10 (1.11)	0.44	(-0.16, 0.36)	0.22 (1.24)	0.15	(-0.08, 0.52)
	AgeAtMenar >= 14	0.06 (1.06)	0.70	(-0.23, 0.35)	0.03 (1.03)	0.85	(-0.31, 0.38)
	AgeAtMenar missing	0.21 (1.24)	0.71	(-0.93, 1.36)	0.43 (1.54)	0.46	(-0.72, 1.59)
	BMI 25 - 29	0.10 (1.11)	0.42	(-0.15, 0.35)	-0.01 (0.94)	0.94	(-0.32, 0.30)
	BMI > 29	-0.26 (0.77)	0.13	(-0.59,-0.08)	-0.16 (0.85)	0.41	(-0.53, 0.22)
	Alcohol	0.07 (1.07)	0.64	(-0.21, 0.34)	0.06 (1.06)	0.72	(-0.26, 0.38)
	Alcohol missing	-0.18 (0.84)	0.42	(-0.62, 0.26)	-0.19 (0.82)	0.45	(-0.70, 0.31)
	PMHuse: past	-0.12 (0.89)	0.46	(-0.44, 0.20)	-0.20 (0.82)	0.28	(-0.56, 0.16)
	PMHuse: current	0.06 (1.07)	0.65	(-0.21, 0.34)	-0.13 (0.88)	0.46	(-0.48, 0.22)
	PMHuse missing	-0.17 (0.84)	0.37	(-0.54, 0.20)	-0.19 (0.83)	0.37	(-0.61, 0.23)
	Cluster 2	Parity 1-4 & AgeAt1stBirth < 25	-0.13 (0.87)	0.56	(-0.58, 0.31)	0.01 (1.01)	0.96
Parity 1-4 & AgeAt1stBirth 25 - 30		0.05 (1.05)	0.84	(-0.40, 0.49)	0.07 (1.08)	0.75	(-0.37, 0.51)
Parity 1-4 & AgeAt1stBirth > 30		0.35 (1.41)	0.17	(-0.15, 0.84)	0.30 (1.35)	0.22	(-0.18, 0.78)
Parity > 4 & AgeAt1stBirth < 25		-0.42 (0.65)	0.14	(-0.99, 0.14)	-0.20 (0.82)	0.46	(-0.73, 0.33)
Parity > 4 & AgeAt1stBirth > 25		-0.09 (0.91)	0.77	(-0.68, 0.50)	0.01 (1.01)	0.98	(-0.56, 0.58)
Parity or AgeAt1stBirth missing		0.21 (1.23)	0.59	(-0.54, 0.95)	0.15 (1.16)	0.71	(-0.62, 0.92)
AgeAtMenar = 13		-0.13 (0.88)	0.29	(-0.36, 0.11)	-0.19 (0.82)	0.10	(-0.42, 0.04)
AgeAtMenar >= 14		-0.23 (0.79)	0.10	(-0.51, 0.05)	-0.25 (0.78)	0.06	(-0.51, 0.01)
AgeAtMenar: missing		-0.08 (0.92)	0.89	(-1.22, 1.06)	-0.25 (0.78)	0.67	(-1.39, 0.89)
BMI 25 - 29		0.32 (1.38)	0.01	(0.10, 0.55)	0.33 (1.39)	<0.01	(0.11, 0.54)
BMI > 29	0.16 (1.17)	0.29	(-0.13, 0.44)	0.06 (1.06)	0.68	(-0.22, 0.34)	

	Naïve K-means			Unified partial-likelihood		
	Estimate(exp)	p	CI(95%)	Estimate(exp)	p	CI(95%)
Alcohol	0.10 (1.10)	0.45	(-0.15, 0.34)	0.12 (1.13)	0.32	(-0.12, 0.36)
Alcohol missing	-0.46 (0.63)	0.04	(-0.91, -0.02)	-0.41 (0.67)	0.07	(-0.84, 0.03)
PMHuse past	0.00 (1.0)	0.99	(-0.28, 0.28)	-0.17 (0.84)	0.23	(-0.44, 0.10)
PMHuse current	-0.13 (0.87)	0.31	(-0.39, 0.12)	0.02 (1.02)	0.85	(-0.22, 0.27)
PMHuse missing	-0.37 (0.69)	0.04	(-0.73, -0.01)	-0.48 (0.62)	0.01	(-0.84, -0.12)