Comparison of statistical methods for detection of serum lipid biomarkers for mesothelioma and asbestos exposure

Aim: We compared three statistical methods in selecting a panel of serum lipid biomarkers for mesothelioma and asbestos exposure. **Materials & methods:** Serum samples from mesothelioma, asbestos-exposed subjects and controls (40 per group) were analyzed. Three variable selection methods were considered: top-ranked predictors from univariate model, stepwise and least absolute shrinkage and selection operator. Crossed-validated area under the receiver operating characteristic curve was used to compare the prediction performance. **Results:** Lipids with high crossed-validated area under the curve were identified. Lipid with mass-to-charge ratio of 372.31 was selected by all three methods comparing mesothelioma versus control. Lipids with mass-to-charge ratio of 1464.80 and 329.21 were selected by two models for asbestos exposure versus control. **Conclusion:** Different methods selected a similar set of serum lipids. Combining candidate biomarkers can improve prediction.

First draft submitted: 13 March 2017; Accepted for publication: 2 May 2017; Published online: 23 May 2017

Keywords: asbestos exposure • biomarker • LASSO • mesothelioma • PCA • stepwise regression

Mesothelioma is a rare cancer that can develop in the pleural lining, stomach, heart and testis. 5-year survival rates of mesothelioma are quite poor – usually less than 5% [1]. Although early identification can increase the likelihood of a successful treatment success, with a long latency period of the disease, early detection is often very challenging [2]. Therefore, there is interest in identifying biomarkers for mesothelioma to assist identification of potential cases and understand the mechanism of the disease. Asbestos exposure has been found to be associated with incidence of mesothelioma [3–5], and is considered a critical risk factor. Exposure to asbestos can be both occupational and environmental. High-risk populations occupationally exposed to asbestos includes workers in asbestos products manufacturing and asbestos mining; environmental exposure may include household and residential

exposures. Residential exposure is often a result of close proximity to asbestos mining and manufacturing site, and the most common source of household exposure is asbestos dust brought home from a family member working closely with asbestos [3]. According to the meta-analysis by Bourdes *et al.* [5], the relative risk for household exposure was estimated to be 8.1 (95% CI: 5.3–12), and 7 (95% CI: 4.7–11) for residential exposure. Developing biomarkers for asbestos exposure will also benefit early detection of the disease among high-risk population.

Different methods, including genomics, proteomics and immunomics, have been utilized to identify biomarkers for mesothelioma [6]. Some of the biomarkers reported to date include mesothelin, osteopontin, fibulin-3 and HMGB-1 [7–14]. Mesothelin is a glycosylphosphatidylinositol-linked cell surface glycoprotein with 40 kDa in mesothelial

Rengyi Xu^{1,2}, Clementina Mesaros¹, Liwei Weng¹, Nathaniel W Snyder³, Anil Vachani^{1,4}, Ian A Blair¹ & Wei-Ting Hwang^{*1,2}

¹Penn Superfund Research Program & Center of Excellence in Environmental Toxicology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104–6160, USA

²Department of Biostatistics & Epidemiology, Perelman School of Medicine, University of Pennsylvania, PA, USA

³AJ Drexel Autism Institute, Drexel University, Philadelphia, PA 19104, USA

⁴Pulmonary, Allergy & Critical Care Division, Perelman School of Medicine, University of Pennsylvania, PA, USA

*Author for correspondence: whwang@mail.med.upenn.edu

Future
Medicine part of

fsg

cells [11]. Robinson *et al.* identified soluble mesothelin-related peptide highly elevated in mesothelioma patients, with a sensitivity of 83% and specificity of 95% compared with healthy controls [7]. Despite the high specificity, its relatively low sensitivity limited its use in early detection [6]. Osteopontin is a glycoprotein overexpressed in lung cancer and is regulated in cell-signaling pathways associated with asbestos-induced carcinogenesis. Serum osteopontin levels comparing subjects exposed to asbestos without cancer and patients exposed to asbestos with pleural mesothelioma had a sensitivity of 78% and specificity of 86% [10]. Plasma fibulin-3 levels were found to be significantly higher in patients with pleural mesothelioma; Pass *et al.* reported that compared with control subjects with asbestos exposure, plasma fibulin-3 levels distinguished stage I or II mesothelioma patients with a sensitivity of 100% and specificity of 94%. Effusion fibulin-3 levels, on the other hand, were found to have very poor correlation with plasma levels [8]. However, Creaney *et al.* reported a sensitivity of only 22% at a specificity of 95% in a more recent study [6]. Other recent research also showed that fibulin-3 had a low diagnostic accuracy, with area under the curve (AUC) of 0.63 and 0.56 for detecting mesothelioma cases and asbestos exposure, respectively [13]. More recently, Napolitano *et al.* stated that hyperacetylated HMGB1 serum levels were significantly higher in patients with mesothelioma and asbestos-exposed subjects, compared with unexposed controls [14] and can be a potential biomarker. The sensitivity and specificity of using serum hyperacetylated HMGB1 to discriminate mesothelioma and controls were 100%. But the validity of this new biomarker remains to be confirmed further with other samples.

In this report, we applied and compared commonly used statistical methods to identify a panel of lipid biomarkers for mesothelioma and for asbestos exposure.

Materials & methods

Sample sources

Subjects consisted of 40 mesothelioma patients randomly selected from the Penn Lung Center and the Penn Mesothelioma and Pleural Program, 40 hospital-based control subjects without cancer and 40 randomly selected from a cohort of 150 participants from the Philadelphia Insulators and Asbestos Workers Local 14 Union who had been heavily exposed to asbestos. IRB approval was obtained from the University of Pennsylvania (PA, USA) for the study (protocol #819932).

Liquid chromatography–high-resolution mass spectrometry analysis

Serum samples from each subject were processed and analyzed by liquid chromatography (LC) high-resolution mass spectrometry exactly as described by Sny-

der *et al.* [15]. Differential analysis was done by SIEVE software (Thermo Scientific) and then raw LC–mass spectrometry data were analyzed in Xcalibur v2.6 and TraceFinder v3.2 (Thermo Fisher Scientific, CA, USA). Quantification was conducted using high-resolution, accurate mass (HR/AM)-MS1 with a 5-ppm window.

Relative abundance of each lipid was quantified by area ratio of the lipid candidate relative to an internal standard. A total of 35 serum candidate lipids were identified after differential analysis. Among which four were duplicates thus were removed from the further analysis. The area ratio of each lipid candidate biomarker ranged from 0.001 to 32 across all subjects. The mean of the 31 lipids across all subjects ranged from 0.08 to 10.15 and lipid-specific standard deviation ranged from 0.49 to 5.58. We used the mass-to-charge ratio (m/z) to denote each lipid candidate. For example, L244.23 denotes the lipid with m/z of 244.23.

Logistic regression model

Because biomarkers distinguishing mesothelioma from controls may or may not be useful to distinguish asbestos-exposed subjects from controls, two sets of analysis were considered separately; one analysis to compare mesothelioma versus controls and, another analysis to compare asbestos exposure versus controls. Because the outcome in both analyses was a binary indicator of whether subject had mesothelioma (or exposed to asbestos in the second analysis), logistic regression model was employed. Logistic regression is a regression technique used for modeling a binary outcome variable as a function of predictors, in this case area ratio of lipid biomarkers. Let outcome Y_i denote the indicator of whether subject i has mesothelioma (or being exposed to asbestos in the second analysis) and $X_{1,i}$, $X_{2,i}$, ..., $X_{k,i}$ denote the predictors from subject i , then the logistic regression has the following form:

$$\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i}$$

Regression coefficient s are usually estimated by maximum likelihood estimation method and have the log odds ratio interpretation representing the magnitude of the association between the predictor and outcome. To compare the predictive values between logistic regression models with different sets of predictors, area under the receiver operating characteristic (ROC) curve was used. AUC measures a model's ability in discriminating between mesothelioma cases (or asbestos-exposed subjects) and controls. AUC can also be interpreted as the probability of the prediction of the disease status made by the model is in concordance of the observed status [16].

Selection of important lipids

We considered three different statistical approaches: univariate, stepwise and least absolute shrinkage and selection operator (LASSO) approaches. In the univariate analysis, separate logistic regression model was fitted with only one predictor X (i.e., lipid candidate) in the model each time. Then the top five lipids with the highest AUCs among all the potential biomarker lipids were chosen. In the stepwise regression approach, we used the forward selection approach to choose lipids sequentially with efforts to minimize the Akaike information criterion (AIC). The AIC statistic equals to $-2 \log\text{-likelihood function} - 2 \times (\text{number of parameter estimated in the model})$. It rewards a better fit of a model via minimizing the log-likelihood function but at the same time penalizes models for having a large number of predictors. In forward selection, a predictor is added into a model only if it reduces AIC of the current model, which indicates an improvement in the overall performance of model. Similarly, LASSO regression is also a type of penalized regression model that a penalty term is added to the log-likelihood function to force the sum of the estimated regression coefficients (in the absolute value) from all selected predictors to be less than a fixed value [17]. Using this penalty term constrains, the number of predictors is selected in the final model such that regression coefficients for some of the predictors shrink to zero. Then the final model includes only the variables with nonzero coefficients, instead of using all of the potential predictors, and thus reducing collinearity and overfitting issues due to including too many and potentially correlated predictors. The equation below shows the negative log-likelihood of a LASSO regression with k candidate predictors and n observations, where the last term ($\lambda \sum_{j=1}^k \|\beta_j\|$) is the penalty term with a tuning parameter, λ . The set of β s that minimizes the equation is the final estimated coefficients from the LASSO regression. The value of the tuning parameter λ was selected to be λ that maximized the AUC under a five-fold cross-validation process. To further reduce sampling variations in each of the cross-validation run, we use the median value of the λ s from 30 runs instead. There are other types of penalized regression approaches, such as elastic net model that uses different penalty terms to handle a large number and highly correlated predictors. Considering that LASSO is more well-known and commonly used among researchers, we choose to focus on LASSO model in this report.

$$\min_{\beta} - \sum_{j=1}^k \left(\sum_{i=1}^n y_i x_{ij} \beta_j \right) + \sum_{i=1}^n \log \left[1 + \exp \left(\sum_{j=1}^k \beta_j x_{ij} \right) \right] + \lambda \sum_{j=1}^k \|\beta_j\|$$

Handling collinearity among predictors

We considered two sets of variables as our predictors (X s) for the logistic regression models. The first set of predictor was the original measurement of area ratios for each lipid. Second, because many of the candidate lipids in these data were highly correlated with each other, we addressed the issue of collinearity by first performing a principal component analysis (PCA) on the 31 serum lipids and converted them into a set of linearly uncorrelated variables. We computed 31 principle components (PC1 to PC31) based on the factor loadings and used PC1 to PC31 as the second set of predictors in the logistic regression model. A total of six models were considered in the analysis: univariate regression using the original lipid level (model 1), univariate regression using the principal components (PCs; model 2), stepwise regression using the original lipid level (model 3), stepwise regression using the PCs (model 4), LASSO regression using the original lipid level (model 5) and LASSO using the PCs (model 6).

In models 1, 3 and 5, in which the original measure of abundances was used as predictors, we further removed ten (mesothelioma vs control) and eight (asbestos exposure vs control) lipids that had a correlation of 0.95 or above with another lipid and with a lighter m/z ratio. In model 2, 4 and 6, all 31 PCs were used. For each model, we report individual and cumulative AUC and the associated p -values. In addition, the mean of the cross-validated AUC (fivefold) and the associated 95% CI for multiple lipids combined are also reported. Without independent data to perform external validation, cross-validated AUC helps to confirm the model performance and adjust for the over optimism associated with using the same data for model development and evaluation. Due to the limited number of sample sizes in each group, we used five-fold rather than tenfold cross-validation that is typically done for this purpose.

The analysis was performed using R statistical software (version 3.3.3). We used the `glmnet` function from the `glmnet` package to perform the LASSO regression, and the built-in `princomp` and `step` function for PCA and stepwise analysis, respectively.

Results

Figure 1 shows a heat map of the area ratio of each of the 31 lipid biomarkers from the three groups. Most of the area ratios concentrated on the range of 0–0.05. Three lipids (L636.56, L636.57 and L782.57) had the highest area ratios, centering around 1–15. Among all the possible pairwise correlations from 31 candidate lipids, 47% of the pairs had a correlation within -0.2 and 0.2, and 12% of the pairs had a correlation above 0.8, which is considered as a very high correlation in most statistical applications. **Figure 2** shows a scatter and correlation

Figure 1. Heat map of area ratios using 120 subjects.

plot from ten randomly selected lipids. The ellipses in the upper diagonal on **Figure 2** reflect the correlation between any two lipids. The tighter the ellipse is the higher the correlation between lipids is. As shown in **Figure 2**, the correlation varied across different lipids pairs. For example, L374.33 and L379.28, and L610.18 and L612.18 were highly correlated with each other, while the correlation among other lipids were not.

Using PCA, the first three PCs accounted for 65.5, 25.2 and 4.7% variability in the data, respectively, with a cumulative proportion of 95.3%. The first PC (PC1) was heavily loaded on L782.57, while the second PC (PC2) loaded mainly on L638.57 and L710.63 and the third PC (PC3) loaded on L820.36.

Results from the six models for mesothelioma versus control and for asbestos exposure versus control are summarized and reported in **Tables 1 & 2**, respectively.

Mesothelioma versus control

Using the original measure of lipid abundance, univariate regression approach (model 1) selected L379.28 and L372.31, with a cumulative cvAUC of 0.991 (95% CI: 0.976–1). The stepwise model (model 3) selected only one lipid biomarker, L372.31, with a cvAUC of 0.988 (95% CI: 0.967–1). LASSO regression (model 5) identified L372.31 and L782.57, with a cumulative cvAUC of 0.993 (95% CI: 0.981–1). Models 1, 3 and 5 all selected L372.31 as one of the important predictors. L829.68

was selected by both the univariate and LASSO regression. L372.31 alone provides an AUC of 0.988. Correlations among the selected lipids ranged from 0.30 to 0.93, and were below 0.50 under LASSO regression.

Using the computed PCs as predictors, univariate regression (model 2) approach selected PC6, PC3, PC1, PC4 and PC11. PC11 was excluded in the final model because the model with all five predictors resulted in an AUC of 1, indicating an overfitting issue. The final model included the four PCs which had a cumulative

cvAUC of 0.988 (95% CI: 0.973–1). Stepwise regression (model 4) selected PC6, PC1, PC12 and PC2. The final model included the first three selected PCs which had a cvAUC of 0.991 (95% CI: 0.975–1). LASSO regression (model 6) identified PC1, PC6, PC11, PC14 and PC30, with a cumulative cvAUC of 0.99 (95% CI: 0.971–1). PC1, which loaded heavily on L782.57, and PC6, which loaded mostly on L372.31, were selected by all three models. PC14 loaded heavily on L372.31, which was also selected by models 1, 3 and 5. L782.57

Figure 2. Scatter and correlation plot of ten randomly selected lipid biomarkers among mesothelioma versus control.

Table 1. Summarized results from model 1–6 for mesothelioma versus control.												
Statistical methods	Original variable					Principal component						
	Lipid	AUC (p-value)	Cum. AUC	cvAUC (95% CI)	Corr.	PC	Lipid with top loading	AUC (p-value)	Cum. AUC	cvAUC (95% CI)	Corr.	
Univariate	L379.28	0.989 (0.001)	0.989	–	0.50–0.93	PC6	L372.31	0.841 (<0.001)	0.841	–	0.01–0.30	
	L372.31	0.988 (0.002)	0.991	0.991 (0.976–1)	–	PC3	L820.36	0.821 (<0.001)	0.923	–	–	
	L374.33	0.987 (0.001)	–	–	–	PC1	L782.57	0.817 (<0.001)	0.986	–	–	
	L800.58	0.943 (<0.001)	–	–	–	PC4	L636.56	0.720 (0.015)	0.988	0.988 (0.973–1)	–	
	L829.68	0.913 (0.008)	–	–	–	PC11	L548.38	0.713 (0.004)	–	–	–	
Stepwise	L372.31	0.988 (0.002)	0.988	0.988 (0.967–1)	0.50	PC6	L372.31	0.841 (<0.001)	0.817	–	0.01–0.22	
	L829.68	0.913 (0.008)	–	–	–	PC1	L782.57	0.817 (<0.001)	0.977	–	–	
	–	–	–	–	–	PC12	L671.58	0.479 (0.218)	0.991	0.991 (0.975–1)	–	
	–	–	–	–	–	PC2	L638.57	0.611 (0.599)	–	–	–	
	L372.31	0.988 (0.002)	0.988	–	0.02–0.50	PC1	L782.57	0.817 (<0.001)	0.817	–	0.01–0.24	
LASSO	L782.57	0.816 (<0.001)	0.993	0.993 (0.981–1)	–	PC6	L372.31	0.841 (<0.001)	0.977	–	–	
	L829.68	0.913 (0.008)	–	–	–	PC11	L548.38	0.713 (0.004)	0.988	–	–	
	L244.23	0.768 (0.031)	–	–	–	PC14	L372.31	0.671 (0.033)	0.988	–	–	
	L494.32	0.787 (<0.001)	–	–	–	PC30	L357.30	0.640 (0.071)	0.99	0.99 (0.971–1)	–	
	–	–	–	–	–	–	–	–	–	–	–	
AUC: Area under the curve; Corr: Correlation; Cum.: Cumulative; cv: Cross-validated; LASSO: Least absolute shrinkage and selection operator; PC: Principal component.												

was selected by model 5. Because the PCs were a set of linearly uncorrelated variables, the correlation among the selected variables in models 2, 4 and 6 was low as expected (0.01–0.30).

Asbestos exposure versus control

Table 2 summarizes the results from models 1–6 that compared asbestos exposure subjects with controls. Using the original lipid level, model 1 selected L1464.80, L329.21, L379.28, L633.91 and L636.56, with a cumulative cvAUC of 0.914 (95% CI: 0.849–0.98). Model 3 selected L1464.80, L362.29, L710.63, L244.23 and L548.38, with a cumulative cvAUC of 0.999 (95% CI: 0.996–1). The final model from LASSO regression (model 5) selected a larger set of variables but we only reported the first five that were selected: L244.23, L548.38, L362.29, L612.18 and L710.63, with a cumulative cvAUC of 0.944 (95% CI: 0.894–0.944). L329.21 and L1464.80 were both selected by models 1 and 3, and L244.23, L362.29 and L710.63 were selected by both model 3 and 5. The highest AUC from a single predictor was by L1464.80, which gave a cvAUC of 0.877. Similar to the mesothelioma versus control comparison, the univariate and stepwise approach both selected a set of lipids that are highly correlated with each other, while LASSO was able to control the correlations to be below 0.40.

For the models that used PCs, models 2, 4 and 6 all selected PC3, PC8 and PC10, which loaded heavily on L820.36, L244.23 and L329.21, respectively, with cvAUC of 0.999 (95% CI: 0.996–1). Comparing with the results from models 1, 3 and 5, L329.21 appeared to be an important predictor identified by both methods. The correlation among the selected variables in models 2, 4 and 6 ranged from 0.003 to 0.41.

Discussion

In our study, we implemented three commonly used approaches (stepwise, univariate and LASSO regression) to select lipids useful to distinguish mesothelioma and asbestos-exposed subjects from controls. We identified similar set of lipid biomarkers under these three approaches, either using the original area ratio level or the PCs as predictor variables. The main difference between these approaches is how the method deals with correlations among predictors. PCA controls high correlations among the candidate biomarkers by first converting them into a new set of linearly independent variables before subjecting them for selection. Univariate and stepwise regressions do not attempt to control the correlation among the selected predictors, and thus, using the original level of the biomarkers often results in a panel of highly correlated lipids. On the other hand, LASSO regression automatically lim-

its the issue of collinearity by using a penalty during estimation, regardless of the type of predictors used.

Within the penalized regression family, there are other approaches besides LASSO, such as elastic net regression. The main difference between these two approaches is how they handle the penalty term; LASSO uses a penalty function based on the L-1 norm, β , while elastic net also adds an L-2 norm, β^2 penalty term. When there are highly correlated predictors, which is the case in our study, LASSO and elastic net use a similar methodology in choosing a subset of variables out of candidate predictors. Since both LASSO and elastic net are all considered as a type of penalized regression, we only reported the results from LASSO regression in this report, along with the results from other statistical approaches, univariate and stepwise regressions. Although we have shown that stepwise and univariate approaches have the issue of multiplicity and high correlations, they are still very commonly used for biomarker researchers. By comparing the results from different statistical approaches, we showed the advantages and disadvantages of different statistical approaches.

The five lipids with the top five AUCs from the univariate regression approach were highly correlated (0.50–0.93 for mesothelioma vs control and 0.01–0.78 for asbestos exposure vs control) as expected. Stepwise selection approach identified lipids that had a moderate correlation of 0.50 for mesothelioma versus control and 0.001–0.6 for asbestos exposure versus control. Lipids selected by the penalized LASSO regression were only moderately correlated with each other as intended (correlations ranged from 0.02 to 0.50 for mesothelioma vs control). The correlation was somewhat higher among the lipids selected for asbestos exposure versus control, which ranged from 0.01 to 0.69.

Although the different approaches we considered have different theoretical considerations, the consistency in the selected variable we observed in our study has suggested high likelihood that the lipids identified here are valuable and worth for further investigations. For the mesothelioma versus control comparison, lipid L372.31 was selected by all three methods when we used the original level of the lipid biomarker as predictors in the analysis. When PCs were used as the predictors, all three methods selected PC1, which loaded heavily on L782.57, and PC6, which loaded heavily on L374.33. These two lipids were also selected by models that used original level, showing a consistent result between the two different sets of predictors. Similarly, for asbestos exposed versus control comparison, overlapping sets of lipids were selected (L1464.80, L710.63 and L362.29). Using PCs as the predictors, all three models (models 2, 4 and 6) identified the same set, PC3 (loaded heavily on L820.36), PC10 (loaded heav-

Table 2. Summarized results from model 1–6 for asbestos exposure versus control.												
Statistical methods	Original variable					Principal component						
	Lipid	AUC (p-value)	Cum. AUC	cvAUC (95% CI)	Corr.	PC	Lipid with top loading	AUC (p-value)	Cum. AUC	cvAUC (95% CI)	Corr.	
Univariate	L1464.80	0.877 (<0.001)	0.877	–	0.01–0.78	PC3	L820.36	0.889 (<0.001)	0.841	–	0.03–0.25	
	L329.21	0.791 (0.013)	0.879	–	–	PC10	L494.32	0.786 (<0.001)	0.953	–	–	
	L379.28	0.755 (0.011)	0.891	–	–	PC8	L610.18	0.752 (<0.001)	0.999	(0.996–1)	–	
	L633.91	0.750 (0.040)	0.899	–	–	PC6	L372.31	0.726 (0.002)	–	–	–	
	L636.56	0.748 (0.003)	0.914	0.914 (0.849–0.980)	–	PC2	L638.57	0.704 (0.823)	–	–	–	
Stepwise	L1464.80	0.877 (<0.001)	0.877	–	0.001–0.60	PC3	L820.36	0.889 (<0.001)	0.889	–	0.003–0.09	
	L362.29	0.674 (0.006)	0.951	–	–	PC10	L494.32	0.786 (<0.001)	0.953	–	–	
	L710.63	0.563 (0.383)	0.971	–	–	PC8	L610.18	0.752 (<0.001)	0.999	(0.996–1)	–	
	L244.23	0.736 (<0.001)	0.991	–	–	PC17	L538.16	–	–	–	–	
	L548.38	0.640 (0.042)	0.999	0.999 (0.996–1)	–	–	–	–	–	–	–	
LASSO	L244.23	0.736 (<0.001)	0.988	–	0.01–0.69	PC10	L494.32	0.786 (<0.001)	0.786	–	0.30–0.61	
	L548.38	0.640 (0.042)	0.844	–	–	PC8	L610.18	0.752 (<0.001)	0.9	–	–	
	L362.29	0.674 (0.006)	0.891	–	–	PC3	L820.36	0.889 (<0.001)	0.999	(0.996–1)	–	
	L612.18	0.574 (0.273)	0.939	–	–	PC6	L372.31	0.726 (0.002)	–	–	–	
	L710.63	0.563 (0.383)	0.944	0.944 (0.894–0.944)	–	PC4	L636.56	0.659 (0.066)	–	–	–	
AUC: Area under the curve; Corr.: Correlation; Cum.: Cumulative; cv: Cross-validated; LASSO: Least absolute shrinkage and selection operator; PC: Principal component.												

ily on L244.23) and PC8 (loaded heavily on L329.21). L244.23 was selected by model 5 as well, again showing a consistent result between the different approaches.

In general, multiple biomarkers provide better prediction than single biomarker alone and we have observed this in our results. However, some individual lipids in the current study distinguished mesothelioma cases and control extremely well (i.e., very high AUC); therefore the opportunity for further improvement by combining multiple lipids biomarkers was limited.

Conclusion

Identifying biomarkers for diseases has become more and more important in recent years. In this report, we compared several commonly used statistical methods for biomarker identification and hope to offer some insights on the appropriate use of those statistical tools for future exploration. Currently, we only know the m/z ratio (and an associated retention time on the LC which helps provide specificity in further chemical identification) of the selected serum biomarkers; the next step will be to identify compounds corresponding to those m/z ratios. Since this identification step is time consuming, statistical approaches that can maximize the likelihood of a useful biomarker proceeding to this stage are highly valuable. Likewise, confirming these results in other independent samples requiring additional time and effort is warranted and necessary to establish the validity of those lipids as biomarkers.

Future Perspective

Useful biomarker can inform disease etiology and help early detection and treatment. Biomarker discovery will likely be one of the critical tasks in the future of medicine in studying and treating diseases. Given that the massive amount of data on potential biomarkers is expected to be available due to advances in several technologies, it is crucial to use a valid statistical method for biomarker discovery. In the future, we are interested in focusing on penalized regression approach, which has been shown to have advantages over simpler methods, and comparing the results between different models in the penalized regression family, such as LASSO and elastic net regressions.

Financial & competing interests disclosure

The study was supported by NIH award numbers P42ES023720, P30ES013508, K22ES026235 and R03CA211820. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Summary points

- Identifying biomarkers for mesothelioma and asbestos exposure is helpful for early detection.
- Asbestos exposure has been found to be associated with mesothelioma, the estimated relative risks range between 7 and 8 for nonoccupational exposure, and between 3 and 5 for occupational exposure.
- High correlations observed among the serum lipid candidates require to use statistical approaches that can properly handle collinearity to produce valid results.
- Using logistic regression, three variable selection methods were considered: univariate, stepwise and least absolute shrinkage and selection operator regression.
- Two sets of predictors were considered: the original abundance level as the area ratio of the lipids relative to an internal standard; and principal components that are a set of linearly independent predictors constructed as linear combinations of the original highly correlated lipid values.
- Univariate and stepwise regressions did not address the issue of high correlation among the predictors, resulting in a correlation as high as 0.93 among the selected lipids and an overall overfitted model. On the other hand, least absolute shrinkage and selection operator regression controlled the correlation below 0.5. Principal component analysis solves the collinearity issue by using a set of linearly independent variables as the predictors.
- Different variable selection methods differed by how they address the issue of collinearity among predictors and identified similar panel of serum lipids for mesothelioma and asbestos exposure.
- Combining multiple biomarkers can improve predictive performance of the overall model but the room for improvement can be limited if the top-performing biomarkers had very high area under the curves or if the selected biomarkers are highly correlated with each other.

References

Papers of special note have been highlighted as: • of interest;
•• of considerable interest

- 1 Musk AW, Olsen N, Alfonso H *et al.* Predicting survival in malignant mesothelioma. *Eur. Respir. J.* 38, 1420–1424 (2011).
- Describes the 5-year survival rate for malignant mesothelioma patients and shows the importance of biomarker identification for early detection.
- 2 Boutin C, Nussbaum E, Monnet I *et al.* Intrapleural treatment with recombinant gamma-interferon in early stage malignant pleural mesothelioma. *Cancer* 74, 2460–2467 (1994).
- 3 Ferrate D, Mirabelli D, Tunesi S *et al.* Pleural mesothelioma and occupational and non-occupational asbestos exposure: a case-control study with quantitative risk assessment. *Occup. Environ. Med.* 73, 147–153 (2016).
- 4 Lacourt A, Gramond C, Rolland P *et al.* Occupational and non-occupational attributable risk of asbestos exposure for malignant pleural mesothelioma. *Thorax* 69(6), 532–539 (2014).
- 5 Bourdes V, Boffetta P, Pisani P. Environmental exposure to asbestos and risk of pleural mesothelioma: review and meta-analysis. *Eur. J. Epidemiol.* 16(5), 411–417 (2000).
- 6 Creaney J, Dick IM, Robinson BWS. Discovery of new biomarkers for malignant mesothelioma. *Curr. Pulmonol. Rep.* 4, 15–21 (2015).
- 7 Robinson BWS, Creaney J, Lake R *et al.* Mesothelin-family proteins and diagnosis of mesothelioma. *Lancet* 362, 1612–1616 (2003).
- Important paper demonstrating issues associated with fibulin-3 as a biomarker for mesothelioma.
- 8 Pass HI, Levin SM, Harbut MR *et al.* Fibulin-3 as a blood and effusion biomarker for pleural mesothelioma. *N. Engl. J. Med.* 367, 1417–1427 (2012).
- Landmark paper for identifying fibulin-3 as a biomarker for pleural mesothelioma.
- 9 Grigoriu B-D, Scherpereel A, Devos P *et al.* Utility of osteopontin and serum mesothelin in malignant pleural mesothelioma diagnosis and prognosis assessment. *Clin. Cancer. Res.* 13, 2928–2935 (2007).
- 10 Pass HI, Lott D, Lonardo F *et al.* Asbestos exposure, pleural mesothelioma, and serum osteopontin levels. *N. Engl. J. Med.* 353, 1564–1573 (2005).
- 11 Bayram M, Dongel I, Akbaş A *et al.* Serum biomarkers in patients with mesothelioma and pleural plaques and healthy subjects exposed to naturally occurring asbestos. *Lung* 192(1), 197–203 (2014).
- 12 Creaney J, Dick IM, Meniawy TM *et al.* Comparison of fibulin-3 and mesothelin as markers in malignant mesothelioma. *Thorax* 69(10), 895–902 (2014).
- 13 Kirschner MB, Pulford E, Hoda MA *et al.* Fibulin-3 levels in malignant pleural mesothelioma are associated with prognosis but not diagnosis. *Br. J. Cancer* 113–963, (2015).
- 14 Napolitano A, Antoine DJ, Pellegrini L *et al.* HMGB1 and its hyperacetylated isoform are sensitive and specific serum biomarkers to detect asbestos exposure and to identify mesothelioma patients. *Clin. Cancer Res.* 22(12), 3087–3096 (2016).
- Landmark paper identifying HMGB1 as a recently discovered biomarker for detecting mesothelioma and asbestos exposure.
- 15 Snyder NW, Khezam M, Mesaros CA *et al.* Untargeted metabolomics from biological sources using ultraperformance liquid chromatography-high resolution mass spectrometry (UPLC-HRMS). *J. Vis. Exp.* 75, e50433 (2013).
- 16 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36 (1982).
- Describes meaning and use of area under the receiver operating characteristic curve in evaluating model performance.
- 17 Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58(1), 267–288 (1996).
- Landmark paper demonstrating the methodology of least absolute shrinkage and selection operator regression.