ABSTRACT: Classification is a common technique applied to omics data to build predictive models and identify potential markers of biomedical outcomes. Despite the prevalence of case-control studies, the number of classification methods available to analyze data generated by such studies is extremely limited. Conditional logistic regression is the most commonly used technique, but the associated modeling assumptions limit its ability to identify a large class of sufficiently complicated omics signatures. We propose a data preprocessing step which generalizes and makes any linear or nonlinear classification algorithm, even those typically not appropriate for matched design data, available to be used to model case-control data and identify relevant biomarkers in these study designs. We demonstrate on simulated case-control data that both the classification and variable selection accuracy of each method is improved after applying this processing step and that the proposed methods are comparable to or outperform existing variable selection methods. Finally, we demonstrate the impact of conditional classification algorithms on a large cohort study of children with islet autoimmunity.

KEYWORDS: Diabetes, machine learning, support vector machines, biomarker discovery, variable selection

Introduction
Matched case-control (MCC) studies are a common design for epidemiological studies due to their potential gains in efficiency and avoidance of confounders. The general design of MCC studies is to group individuals with the outcome of interest (“cases”) and those without (“controls”) based on features such as age and sex. In omics studies, often the goal of MCC studies is to identify biomarkers that are highly correlated with the case/control labels that can lead to a better understanding of the cause of the disease. Despite the large number of studies using this design, the number of methods that account for the pairing has grown very slowly and almost always assume a linear relationship between sample features and the outcome of interest. Furthermore, most of the popular methods used to analyze MCC studies perform poorly when the covariate space is high dimensional or when the effects are highly nonlinear.

Although other methods have been proposed,1,2 a majority of methods used to analyze MCC studies while controlling for covariate information are based on conditional logistic regression (CLR).3 Conditional logistic regression is similar to standard logistic regression but controls for the matching design by estimating the effects of each covariate conditional on the paired design.4 As such, CLR is able to identify covariates whose linear effects are associated with each patient’s case-control status without flagging spurious relationships due solely to the paired nature of the study.

In its original form, CLR was designed to handle modestly sized covariate datasets and is not well suited to handle the volume or veracity of data present in an omics-type analysis. To remedy this, several variations of standard CLR have been proposed to deal with high-dimensional datasets.5–9 However, based on these published results, CLR and its many variants still struggle to accurately differentiate the cases from the controls, particularly when there are nonlinear effects, the noise is sufficiently large or the number of inputs is too large.7

Modern classification techniques such as random forests (RF),10 support vector machine11 (SVM), and naive Bayes (NB) can successfully model such complicated input/output relationships but do not account for the matched design of MCC studies and require modification to be used in these situations. That is, applying these methods without accounting for the paired nature of the study likely accounts for their poor performance relative to CLR, which does account for the pairing.8,12–15 More recently, Dimou et al16 proposed a paired SVM approach to identify damaged regions of the brain, but the specialized kernel they proposed is not applicable to general classification problems with binary outcomes.

Received: January 21, 2019. Accepted: April 26, 2019.

Type: Technical Advances

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this Article: The TEDDY Study is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, and Contract No. HHSN2720070014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work was supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR001082).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Bobbie-Jo Webb-Robertson, 902 Battelle Blvd J4-18, Pacific Northwest National Laboratory, Richland WA 99352, USA. Email: bjw@pnnl.gov

* Members of the TEDDY Study Group are listed in the online supplemental appendix.

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
In this article, we show that by preprocessing the data, any number of linear and nonlinear classification algorithms can be used to appropriately analyze data generated by MCC studies. This method is a general framework which, for the first time, makes a much larger set of classification algorithms available to researchers analyzing MCC study data. The new group of classification algorithms resulting from this method, called conditional classification algorithms, are designed specifically to analyze MCC studies. Using artificially generated data, we show when and by how much the proposed conditional classification algorithms outperform their standard counterparts. We also identify situations in which they will outperform CLR.

In the next section, we describe the proposed preprocessing technique. We then demonstrate how classification and variable selection accuracies improve in a simulation study. Finally, we employ our methods along with CLR to a large cohort study on The Environmental Determinants of Diabetes in the Young (TEDDY) to understand the ramifications on messy and high-dimensional real-world omics data.

Methods

In this section, we describe the data processing step that defines the proposed set of conditional classification algorithms. The theoretical derivations that prove the validity of the proposed approach are given in the supplemental material. We then describe the methods used to generate and analyze the artificial data. Finally, the TEDDY study is described including the data cleaning steps and how it was analyzed.

Conditional classification algorithms

To make a standard classification algorithm conditional, it must account for the paired structure of the MCC study. We propose centering the within pair data by its mean to address the paired data structure. For example, consider a single protein measured on 4 individuals that are split into 2 case-control pairs. The protein abundance for the case and control in pair 1 is 750 and 500, respectively, while that same protein has abundance 500 and 250, respectively, for pair 2. Because the abundance for the control in pair 1 is the same as the case in pair 2, standard classification algorithms would not identify this protein as significant. After pair correction, however, any classification algorithm would identify this protein as significant because the pair-corrected abundance values are 125 and −125 for the case and control, respectively, in both pairs.

Put mathematically, this is equivalent to the common statistical practice of projecting a feature matrix into the null space generated by the matrix of pair indicators. Let \( n \) denote the cohort size, which is composed of \( P \) disjoint and equally sized strata, i.e., pairs, of MCC subjects. Let \( m \) denote the size of each strata and \( K \) denote the number of features. Define the matrix of strata indicators \( Z = I_{P \times P} \otimes 1_m \), where \( I_{P \times P} \) is a \( P \times P \) identity matrix, \( 1_m \) is an \( m \)-dimensional column vector of ones and \( \otimes \) is the Kronecker product. Then, \( Z \) is a \( n \times P \) matrix where the \((i,j)\)th element is a 1 if subject \( i \) is in strata \( j \) and is 0 otherwise. The projection matrix associated with \( Z \) is \( P_Z = Z(Z^TZ)^{-1}Z^T = I_{P \times P} \otimes (1_{1_m}1_{1_m}^T) / m \), which is a \( n \times n \) block diagonal matrix with \( m \times m \) blocks where every element is \( 1/m \). To project the \( n \times K \) feature matrix \( X \) into the null space of \( Z \), pre-multiply \( X \) by \( I_{n \times n} - P_Z \). Define the case-control corrected feature matrix \( X^* \) as \( X^* = (I_{n \times n} - P_Z)X \).

We define any classification algorithm trained using the pair adjusted features \( X^* \) as a conditional classification algorithm. We refer to training the same classification algorithm on the raw features \( X \) as the standard classification algorithm. In the “Results” section, we empirically compare the conditional and standard classification algorithms. In the context of 1:1 case-control studies, both linear discriminant analysis (LDA) and the Gaussian NB classifier are guaranteed to return one case and one control per strata; no such guarantee exists for the standard versions of these classifiers. In addition, we show that CLR is a special case of the proposed set of paired classification algorithms. We do this by showing the CLR likelihood is unaffected by the pair correction and that the maximizer of the CLR likelihood also maximizes the pair-corrected logistic regression (up to a scaling factor). See the supporting information for details on these two theoretical results.

Figure 1 shows a simulated example that clearly demonstrates the value of the conditional classification approach over CLR and standard machine learning for data with complicated structure. This example includes only two features to allow for easy visualization, but could be extended to large feature spaces. Consider a 1:1 case-control dataset with two features, \( x_{ij} \), where \( i \) indicates the case-control status and the black lines connect individuals in the same pair. Let \( X \) represent the full feature matrix and \( X_i \) be the \( 2 \times 2 \) submatrix of \( X \) that has all the information for case-control strata \( i \). Each submatrix \( X_i \) is created by

\[
X_i = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} = \begin{pmatrix} r_{11} \cos(\phi_i) + \mu_1 & r_{11} \sin(\phi_i) + \mu_2 \\ r_{21} \cos(\phi_i) + \mu_1 & r_{22} \sin(\phi_i) + \mu_2 \end{pmatrix}
\]

where \( r_{11} \sim \text{Unif}[0.4, 0.7] \), \( r_{12} = -r_{11} \), \( \phi_i \sim \text{Unif}(0, 2\pi) \), and \( \mu_j \sim \mathcal{N}(0, 5) \) for \( k = 1, 2 \). The response label for each individual in pair \( i \), \( y_i \), is set to 0 (control) if \( \phi_i \) is in the second, fourth, fifth, or seventh octant of the feature space shown in Figure 1B and 1 (case) otherwise. This design ensures that each pair is composed of one case and one control.

A set of 100 case-control pairs generated in this fashion is plotted in Figure 1A where the dot color indicates case-control status and the black lines connect individuals in the same pair. In its raw form, the data are noisy and difficult to classify. For this dataset, a CLR model with coefficients for \( x_{g1} \), \( x_{g2} \) and \( x_{g1}x_{g2} \) returns a misclassification rate of 43%. A standard SVM with a radial basis function performs slightly worse with a misclassification rate of 47%. A large proportion of the errors
committed by the SVM are due to the fact that both individuals in each strata are given the same label. However, after the pair correction is applied (Figure 1B), the differences between cases and controls are clearly visible through the dividing boundary between the two strata, which is also clearly nonlinear. In fact, the conditional NB classifier and the conditional SVM with a linear kernel both perform worse than the CLR when applied to the corrected data (misclassification rates greater than 43% for this dataset). A conditional SVM with a radial basis function and a conditional random forest (CRF), however, achieve misclassification rates of 3% and 4%, respectively. In addition, the fitted class labels generated by these methods consist of one case and one control per pair. To illustrate the results of the conditional SVM with a radial basis function, Figure 1C is a plot of the decision boundary learned by the conditional SVM with a radial basis function kernel when applied to this dataset.

In terms of variable importance, the $P$-values associated with regression coefficients of the CLR indicate that $X_1$ is an important feature, $P$-value .002, while $X_2$ and the interaction of the two variables are not statistically significant, $P$-values of .59 and .07, respectively. Conversely, removing either variable from the conditional support vector machine (CSVM) or CRF dramatically decreases model performance, indicating both variables are important in differentiating the cases from the controls. This indicates that CLR is not able to identify important features if their relationship with the outcome of interest is highly nonlinear whereas the CSVM and CRF models are.

Although this is a small toy example, it illustrates the potential gains in classifier accuracy and biomarker discovery by incorporating nonlinear classifiers into the domain. In the next section, we describe a much larger simulation study to empirically investigate the potential gains of conditional classifiers.

**Simulation study**

The simulation scenario implemented here is motivated by Balasubramanian et al. Each simulated dataset consists of 200
1:1 case-control pairs and 225 features, thus using the notation from the previous section, \( p = 200, m = 2, n = 400 \) and \( K = 225 \). The first 25 features are significant biomarkers while the remaining 200 features are noise. The features for each pair were drawn from the bivariate normal distribution

\[
\begin{bmatrix}
  x_{1k} \\
  x_{2k}
\end{bmatrix}
\sim N_2\left(\begin{bmatrix} 1 \\ 1 + \delta_k \end{bmatrix}, \begin{bmatrix} 1 & \rho_k \\\n  \rho_k & 1 \end{bmatrix}\right)
\]

where \( \delta_k \) and \( \rho_k \) are the mean shift and within pair correlation, respectively, associated with feature \( k = 1, \ldots, K \). For this study we allowed the magnitude of the mean shift to take 3 possible values \( |\delta_k| \in [0.125, 0.25, 0.5] \) for \( k = 1, \ldots, 20 \) and 0 otherwise. Within each dataset, the sign of each feature was allowed to be positive or negative with equal probability; therefore, biomarkers that are both over-expressed and under-expressed in the case samples relative to their controls are considered. We considered 4 possible values of the biomarker within pair correlations: \( \rho_k \in [0,0.1,0.4,0.8] \) for \( k \leq 20 \) and 0 otherwise. For each combination of \( \delta_k \) and \( \rho_k \), 2000 datasets were created and 7 different classification algorithms were fit to each dataset: logistic regression (LR), NB, SVM with a radial basis function kernel (SVM-RBF), SVM with a linear kernel (SVM-Lin), RF, LDA, and random penalized conditional logistic regression (RPCLR). Random penalized conditional logistic regression is different from the other 6 methods in that only a conditional version exists and it cannot be used to predict an individual’s case/control label. Therefore, we will only compare it to the other 6 methods in terms of variable importance accuracy and not predictive accuracy. Furthermore, because we cannot assess its predictive accuracy in the context of the TEDDY data, it will not be applied to those data.

To assess the impact of the proposed preprocessing step, conditional and standard versions of each method, except RPCLR, were applied to every dataset. See Table 1 for specifics on the implementation of each method. In terms of the tuning parameters associated with each of the algorithms, cross-validation (CV) was used to choose the tuning parameters for the regularized CLR model. The width of the Gaussian kernel used by the SVM-RBF model was set to the median of the squared Euclidean distances between the input features. The number of trees included in the RF model was set to 500 and the number of variables to consider at each node was the largest integer less than the square root of the total number of features. For RPCLR, the number of variables included in each model was set to 7 and the number of bootstrap replicates was set to 2000, as recommended by the authors.

We used predictive accuracy to determine which method should be used to identify important biomarkers. In this context, prediction is the labeling of individuals within a group that was not used to estimate the parameter values. The predictive accuracy of each method was quantified by computing the proportion of individuals whose classes were predicted correctly in a 5-fold CV framework. The exact procedure is summarized as follows:

1. Case-control pairs are randomly split into 5 disjoint groups numbered 1 through 5;
2. For each group \( g = 1, \ldots, 5 \):
   a. Train each classification algorithm using all the data except group \( g \);
   b. Predict the class labels for individuals in the testing set, group \( g \);
3. Compare the true class labels to the predicted class labels in 2b to compute the proportion of individuals that were classified correctly.

The significance of each feature was quantified by a variable importance metric computed on each of the training sets and then averaged across the 5 folds. The metrics used depend on the method applied (last column of Table 1). These metrics were used to quantify the importance of each feature. Receiver operating characteristic (ROC) curves based on the true importance labels and the variable importance metrics were created to assess the accuracy of each method, which was summarized using the area under the curve (AUC). Therefore, each

### Table 1. The 6 classification methods used were implemented in the program R\(^{17}\); the specific functions (and packages) used are given below as well as the different variable importance methods used for each method.

| METHOD      | R FUNCTION (PACKAGE) | VARIABLE IMPORTANCE MEASURE |
|-------------|----------------------|-----------------------------|
| LDA         | lda (MASS\(^{18}\))  | Magnitude of the scalings of the discriminant functions |
| CLR         | clogitL1 (clogitL1\(^{19}\)) | Standardized regression coefficient magnitude |
| Naive Bayes | naiveBayes (e107\(^{20}\)) | KL distance between the conditional distribution of each feature given the target class |
| Both SVMs   | ksvm (kernlab\(^{21}\)) | 1D sensitivity analysis measure based on average absolute deviation from the median\(^{22}\) |
| RF          | ranger (ranger\(^{23}\)) | Gini index |
| RPCLR       | GetVarImp (RPCLR\(^{24}\)) | Average change in out of bag AIC with and without each variable |

Abbreviations: AIC, Akaike information criterion; CLR, conditional logistic regression; KL, Kullback Leibler; LDA, linear discriminant analysis; RF, random forests; RPCLR, random penalized conditional logistic regression; SVM, support vector machines.
method’s variable selection accuracy is measured by this AUC value, which lies in the range [0, 1] with larger values being better, and a value of 0.5 corresponds to an uninformative classifier.

**TEDDY study**

The TEDDY study\textsuperscript{25} is a large prospective study with the goal of discovering factors that initiate the autoimmune response and destruction of the pancreatic beta cells, leading to the development of type 1 diabetes (T1D). TEDDY was formulated into a nested case-control study to enable biomarker studies, pairing on: clinical center, sex, and family history of T1D,\textsuperscript{26} which resulted in 418 case-control pairs for analysis. TEDDY is particularly interested in understanding the environmental factors that trigger islet autoimmunity (IA), thus the metabolomic, lipidomic, and genetic single-nucleotide polymorphism (SNP) data at the time point of autoimmunity are evaluated.

After a log\textsubscript{2} transformation of the’omics data, 5 preprocessing steps were applied to each data source that effectively determines a starting number of features: weighted coefficient of variation (CoV),\textsuperscript{27} percent missingness, near zero variance (NZV), univariate pairwise significance tests, and pairwise correlation. We remove features within a source if the weighted CoV is greater than 200%. We define weighted CoV as

\[
\text{CoV}_{\text{weighted}} = \frac{n_{\text{case}} \times \text{CoV}_{\text{case}} + n_{\text{control}} \times \text{CoV}_{\text{control}}}{n_{\text{case}} + n_{\text{control}}}
\]

where \(n_{\text{case}}\) and \(n_{\text{control}}\) are the number of nonmissing values for cases and control, respectively, within in a time point and

\[
\text{CoV}_{\text{case}} = \left| \frac{100 \times \bar{x}_{\text{case}}}{x_{\text{case}}} \right| \quad \text{and} \quad \text{CoV}_{\text{control}} = \left| \frac{100 \times \bar{x}_{\text{control}}}{x_{\text{control}}} \right|
\]

We also remove any features that were more than 10% missing and use RF imputation\textsuperscript{28} to impute those that were less than 10% missing. Next, we remove features that had very few unique values relative to the number of samples or a much greater frequency of the most common value relative to the second most common value.\textsuperscript{29} Before significance tests, the lipid data are handled specially to remove redundant information by eliminating different adducts for the same lipid. For the negatively ionized lipids, we simply removed all Cl\textsuperscript{–} adducts because they tend to ionize poorly. The positively ionized lipids depend on the lipid class in terms of which adduct is a degradation of the lipid due to source fragmentation. Finally, for the LPC and PC classes, we used the most common adduct (H) as the other (Na) was rarely noted. Next, for all data types but the SNP, univariate paired \(t\)-tests were applied to each feature and all features with \(P\)-value less than .20 were retained.

Four criteria were used to filter the SNPs: missingness, minor allele frequency (MAF), the Hardy-Weinberg test for equilibrium (HWE), and CLR. Missingness, MAF, and HWE testing were performed using PLINK version 1.90.\textsuperscript{30} SNPs with missingness less than 1%, MAF less than 0.2, \(P\)-values from the HWE test less than .001, and a conditional logistic regression \(P\)-values less than .006 were retained for the analysis. The .006 \(P\)-value threshold was chosen to parallel the 0.2 threshold used for the other data sources. That is, because there are roughly 33 times more SNPs than other data types, the SNP threshold was set to 0.2/33 = 0.006. As a final step of the data cleaning procedure, all biomolecules that have pairwise correlations greater than 0.9 were removed one-by-one to minimize redundancy in the final dataset. These steps ensure that each feature does not have an excessive relative variability, does not have an excessive amount of missing data, and contains enough variability and significance to potentially distinguish cases from controls.

The same 5-fold CV approach used in the simulation study was used for the TEDDY data to assess each method’s predictive accuracy. However, instead of implementing 5-fold CV once per dataset, the CV method was repeated 200 times per data source to account for the uncertainty of the CV procedure itself. Feature importance was assessed using a single analysis of the full dataset with each method according to the feature importance metrics reported in Table 1. As mentioned previously, RPCLR will not be applied to these data because we cannot assess its predictive accuracy.

**Results**

**Simulation study**

Scatter plots of the 2000 classification accuracy values for each classification method are shown in Figure 2 for \(|\delta| = 0.125\) and all values of the within pair correlation \(\rho\). Random penalized conditional logistic regression cannot be used as a classification algorithm so it is not included. The accuracy values for the conditional and standard versions of each method are plotted on the \(x\)- and \(y\)-axes, respectively. The red points represent datasets for which the standard version of that method was more accurate than its conditional counterpart while the converse is true for the black points. Within each method, the cloud of points are closest to the identity line when the within correlation is zero or small (top two rows). This indicates the conditional methods behave most similarly to their standard counterparts for small values of \(\rho\). As the within pair correlation grows (move down each column), the cloud of points move to the right indicating the conditional methods get more accurate when \(\rho\) increases. The cloud of points do not move up; however, indicating the performance of the standard methods do not change as a function of \(\rho\).

These conclusions are supported by Table 2, which gives the percent of datasets in which the standard method was preferred to the conditional method for each algorithm and combinations of \(\delta\) and \(\rho\). The number of times the standard method is preferred decreases differently for each of the...
algorithms. This implies the sensitivity of each algorithm to the paired structure varies depending upon how the algorithm performs classification. In particular, the linear discriminative algorithms (LDA, LR, and SVM-Lin) require a stronger within pair correlation for the conditional method to clearly outperform its standard counterpart. The nonlinear discriminative algorithms (RF and SVM-RBF) separate themselves more quickly and by a larger margin as the correlation increases. Finally, the linear generative algorithm (NB) separates itself from the onset and creates the widest gap for large values of $\rho$.

Also apparent from Table 2 is the relationship between $|\delta|$ and the conditional method accuracies. In particular, the fact that the standard method is preferred so infrequently with large $|\delta|$ implies the importance of accounting for the paired structure is more important when the within pair difference is larger, as expected.

An analogous representation of the variable selection accuracy is given in Figure 3. Similar to the classification accuracy results, the conditional and standard methods cluster most closely around the identity line for small values of $\rho$ and then drift to the right, which indicates the conditional method improves in accuracy more quickly than the standard method as a function of $\rho$. Unlike the accuracy results, however, the cloud of points also moves upward as a function of $\rho$.

### Table 2. Percentage of simulated datasets in which the standard version of the classification algorithm outperformed the conditional version in terms of classification accuracy by $\delta$ and $\rho$ combination.

| $|\delta|$ | $\rho$ | LDA | LR | SVM-Lin. | RF | SVM-RBF | NB |
|-------|------|-----|-----|---------|----|---------|----|
| 0.125 | 0    | 8.8 | 53.2 | 14.6    | 3.4 | 2.5     | 0.1 |
| 0.01  | 6.3  | 48.5 | 12.7 | 2.5     | 1.9 | 0.1     |
| 0.4   | 2.8  | 29.6 | 7.4  | 0.6     | 0.5 | 0       |
| 0.8   | 0.1  | 0.8  | 0.2  | 0       | 0   | 0       |
| 0.25  | 0    | 6.5  | 12.9 | 12.4    | 0.5 | 0.5     |
| 0.01  | 3.4  | 7.4  | 8.2  | 0.1     | 0.3 | 0       |
| 0.4   | 0.1  | 0.2  | 0.1  | 0       | 0   | 0       |
| 0.8   | 0    | 0    | 0    | 0       | 0   | 0       |
| 0.5   | 0    | 2.2  | 0    | 3.9     | 0   | 0.1     |
| 0.01  | 0.9  | 0    | 1.6  | 0       | 0   | 0       |
| 0.4   | 0    | 0    | 0    | 0       | 0   | 0       |
| 0.8   | 0    | 0    | 0    | 0       | 0   | 0       |

Abbreviations: LDA, linear discriminant analysis; LR, logistic regression; NB, Naive Bayes; RBF, radial basis function; RF, random forests; SVM, support vector machine.
implying the standard method also becomes more accurate at identifying the important biomarkers as $\rho$ increases. Therefore, the standard methods are better able to identify which features should be used to classify the data as $\rho$ increases, but they are no better at performing the actual classification (Figure 2). Finally, the algorithms appear to improve in accuracy at approximately the same rate as a function of $\rho$. That is, there is no clear distinction between the linear and nonlinear discriminative or generative algorithms with respect to accurate variable selection.

Finally, we compare all of the conditional methods by their average variable selection accuracies in Table 3. As expected, the accuracy of all methods improves when $\delta$ and/or $\rho$ increases. One existing method, RPCLR, was the most accurate method in one instance ($\delta = 0.125$ and $\rho = 0.1$) and was in the top 3 in 9 out of 12 scenarios. The proposed conditional Naive Bayes (CNB) and CRF approaches are most frequently in the top 3, followed by the CLDA and CSVM-RBF methods. We can therefore conclude that the proposed methods are comparable or outperform the existing variable selection methods CLR and RPCLR.

**Figure 3.** Scatter plots of the variable selection accuracies for all conditional ($x$-axis) and standard ($y$-axis) methods and values of $\rho$ when $|\delta| = 0.125$. The color of each point indicates which version, standard (red) or conditional (black), of each algorithm is more accurate for each simulated dataset. LDA indicates linear discriminant analysis; LR, logistic regression; NB, Naive Bayes; RBF, radial basis function; RF, random forests; SVM, support vector machine.

---

**Case study—TEDDY study data**

The sample sizes and descriptive feature statistics for each data type are reported in Table 4. The within pair differences and correlation was computed by taking the average absolute difference between features and the average correlation between feature vectors within each pair, respectively. Because the standard classification algorithms do no account for the pairing, we estimated the same quantities for random pairs in the dataset to determine how influential the pairing may be on algorithm performance. To do this, 10000 random pairs were chosen, the same quantities were computed and then averaged across the chosen pairs. Assuming that the mechanism that differentiates cases and controls in the TEDDY data is similar to the simulation scenario explored in the previous section, then the increased correlation and difference between matched pairs relative to random pairs leads us to believe the conditional methods will outperform their standard counterpart. However, the large correlation in the data as a whole could make the difference between the methods small.

Box plots of the 200 repeated accuracy values are shown in Figure 4 for each method and biomolecule. Overall the
and ρ and |δ| |p = 0.1| 0.742| 0.944| 0.939

CLDA 0.653 0.668 0.742| 0.944

CNB 0.662 0.675 0.743 0.939

CRF 0.125 0.660 0.741 0.934

CSVM-Lin. 0.652 0.667 0.741 0.944

CSVM-RBF 0.654 0.669 0.744 0.952

RPCLR 0.656 0.681 0.738 0.929

CLR 0.881 0.898 0.937 0.977

CLDA 0.869 0.891 0.945 0.993

CNB 0.904 0.925 0.971 1.000

CRF 0.25 0.900 0.968 1.000

CSVM-Lin. 0.866 0.888 0.942 0.992

CSVM-RBF 0.875 0.898 0.951 0.998

RPCLR 0.901 0.911 0.963 1.000

CLR 0.980 0.982 0.987 0.993

CLDA 0.959 0.965 0.980 0.996

CNB 0.999 1.000 1.000 1.000

CRF 0.5 0.999 1.000 1.000

CSVM-Lin. 0.956 0.962 0.978 0.996

CSVM-RBF 0.965 0.969 0.981 0.999

RPCLR 0.999 0.999 1.000 1.000

Abbreviations: CLDA, conditional linear discriminant analysis; CLR, conditional logistic regression; CNB, conditional Naive Bayes; CRF, conditional random forests; CSVM, conditional support vector machine; RBF, radial basis function; RPCLR, random penalized conditional logistic regression.

The top 3 methods for each ρ and |δ| combination are denoted with superscripts 1, 2, and 3.

Table 3. Average variable selection accuracy for each method.

To determine which biomolecules were most predictive of the case/control status of each individual, the features were ranked within each source using a variable importance metric appropriate for each classification algorithm (Table 1). A scree plot of the variable importance measures was used to separate the influential from the noninfluential features according to each model. An abridged list of the influential features chosen by the CSVM–RBF and CLR methods along with their ranks, direction of association (+/-), and published literature connecting each biomolecule to IA are given in Table 5.

Discussion

The conditional NB classifier was the most accurate method in the simulation study and was one of the top two methods for 2 of the 4 data types (Figure 5). We hypothesize this is due to the fact that NB is a generative rather than discriminative algorithm. As described in Ng et al,31 generative algorithms have larger asymptotic classification error limits than discriminative classifiers, but have the potential to reach that error limit sooner than their discriminative counterparts. Thus, the consistent performance of the conditional NB algorithm could be due to the rather limited number of individuals in the TEDDY study relative to the complicated biology associated with IA as being learned by noisy ‘omics feature sets. Similarly, even though the difference between cases and controls in the simulation study was rather simple, the small number of important features relative to unimportant ones made the signal difficult to detect.

Further evidence to support this hypothesis is that the CSVM–RBF was the second most accurate method in the simulation study and one of the top two classifiers for all data types. Because the SVM–RBF is a nonlinear classifier, it is able to model the complicated ‘omics to disease relationship much better than the discriminative methods that rely on linear separators, ie, CLR, CLDA, and CSVM-Lin. This is particularly true for the metabolites in which the difference between the conditional and standard versions of each linear discriminative method is small even though they exhibit the largest between strata discrepancy (Table 4), which is a counterintuitive result given the strong pairing information.
Table 4. Feature set sizes, summary statistics, within mean absolute pair differences, within correlations (cor.), and random pair differences correlations for each data type based on the 504 samples.

| STATISTIC            | METABOLICS | POSITIVE LIPIDOMICS | NEGATIVE LIPIDOMICS | SNPS |
|----------------------|------------|---------------------|---------------------|------|
| # Features (K)       | 170        | 277                 | 252                 | 236  |
| Minimum              | –0.079     | 3.585               | 4.807               | 0.000|
| Maximum              | 23.149     | 23.564              | 21.428              | 2.000|
| Mean                 | 10.416     | 14.766              | 13.097              | 1.296|
| Median               | 9.899      | 14.432              | 12.784              | 1.000|
| Standard deviation   | 3.142      | 2.854               | 2.211               | 0.685|
| Within pair difference| 1.880     | 1.099               | 1.043               | 1.519|
| Random pair difference| 1.497     | 0.549               | 0.430               | 0.687|
| Within pair correlation| 0.938   | 0.984               | 0.984               | –    |
| Random pair correlation| 0.749    | 0.968               | 0.967               | –    |

Abbreviation: SNP, single-nucleotide polymorphism. The random pair distance and correlation was computed by randomly sampling 10,000 pairs of random individuals from the dataset and computing their pairwise correlation and mean absolute distance. Within pair correlations for SNPs are not reported because of their discrete nature.

Figure 4. Box plots of the 200 repeated 5-fold cross-validation accuracies for the 4 different data types and 6 different classification algorithms. LDA indicates linear discriminant analysis; LR, logistic regression; NB, Naive Bayes; RBF, radial basis function; RF, random forests; SVM, support vector machine.
One of the most important components of paired classification methods is the ability to select biomarkers that can lead to a better understanding of the diseases being studied. Unlike standard machine learning, these methods are not designed to predict the class of a single case or control as the model is dependent on the pairing. In terms of the variables selected by the different methods (Table 5), CLR and CSVM-RBF selected the same 2 metabolites as the 2 most important features: α-tocopherol and adipic acid. α-tocopherol is a vitamer of vitamin E whose exact role in the progression of IA is still being studied, but has been shown to protect against its progression. Our results similarly indicate that α-tocopherol is negatively associated with IA progression. The role of adipic acid in type 2 diabetes has previously been studied, but this is the first time it has been shown to play an active role in IA. Our results agree and suggest that large amounts of adipic acid are associated with a higher incidence of IA. Both methods found these 2 metabolites to be highly important and agreed on the direction of association is testament to the proposed methods validity with real data.

Both methods indicate that hydroxybutanoic acid is positively correlated with IA, which agrees with previously published studies that indicate it is an early marker for glucose intolerance. Creatinine, a waste-product created by the breakdown of muscle, was ranked highly by both methods.

**Table 5.** An abridged list of the influential metabolites and lipids identified by the CSVM-RBF and CLR methods along with notes and references for biomolecules that have been previously linked to T1D and/or IA.

| SOURCE | BIOMOLECULE           | CSVM-RBF RANK | CLR RANK | REFERENCES |
|--------|-----------------------|---------------|----------|------------|
| Lipids | Acylcarnitine         | —             | 3(+)     | 32         |
|        | Glccer                | —             | 3(−)     | 33         |
|        | Ceramide              | 13(−)         | —        | 34         |
|        | pc_36_5_4_29_824_54   | 20(−)         | 5(−)     | 35         |
|        | fa_16_1_2_71_253_22   | 6(+)          | 6(+)     | 36,37      |
| Metabolites | α-Tocopherol | 1(−)          | 2(−)     | 38–40      |
|        | Adipic acid           | 2(+)          | 1(+)     | 41,42      |
|        | Glucose               | 3(−)          | 28(−)    | 43         |
|        | Creatinine            | 7(+)          | 3(+)     | 44         |
|        | Heptadecanoic acid    | —             | 4(+1)    | 44         |
|        | Hydroxybutanoic acid  | 8(+)          | 11(+)    | 45,46      |
|        | Leucine               | 10(−)         | —        | 47         |
|        | Isothreonic acid      | 20(−)         | —        | 48         |
|        | Glycerol galactoside  | —             | 23(−)    | 49–51      |
| SNP (Gene) | rs7158663 (Meg3) | 10(−)         | 34(−)    | 52         |
|        | rs17388568 (ADAD1)    | —             | 19(−)    | 53,54      |
|        | rs4580644 (CD38)      | —             | 20(+)    | 55,56      |

Abbreviations: CLR, conditional logistic regression; CSVM, conditional support vector machine; IA, islet autoimmunity; RBF, radial basis function; SNP, single-nucleotide polymorphism; T1D, type 1 diabetes.

The direction of the association is indicated by the “+” symbol for factors that are larger in the case group relative to controls and “−” when the opposite is true.
(CSVM-RBF 7; CLR 3) and is a commonly used marker for kidney function in that higher levels of creatinine in the blood or urine indicate decreased kidney function. Islet autoimmunity and hypoglycemia are closely linked with kidney function; therefore, increased creatinine levels would be expected among individuals with IA relative to healthy control due to compromised kidney function. As such, creatinine could be positively correlated with IA like our analysis suggests.

In terms of lipids, there is less agreement between the 2 methods, but several results consistent with the literature have been identified. Conditional logistic regression found acylcarnitine, identified among the positively ionized lipids, to be highly important to model performance and that it is positively correlated with IA progression. Previous studies also found that C3 and C4 acylcarnitines were significantly more abundant in patients with IA and T2D relative to their controls. A highly significant biomarker identified among the positively ionized lipids was hexosylceramide (annotated as glicer), which is negatively associated with IA progression according to both our results and a previous study that found the activation of natural killer T cells by a variant of α-galactosylceramide prevents the onset and recurrence of autoimmune IA. It is also possible that the hexosylceramide is a glucosylceramide because both molecules are isobaric and indistinguishable by mass spectrometry. Inhibition of glucosylceramide synthesis has been associated with an improvement of insulin tolerance. The CSVM-RBF found ceramide to be negatively associated with IA progression, which is supported by the literature. In general, high levels of some fatty acids, such as PA (16:1), have been found to be risk factors for IA, and a result supported by both methods. Finally, some ω-3 polyunsaturated fatty acids, such as pc_36.5_4_29_824_54, have been found to be negatively associated with IA, which is again confirmed by our results.

The SNPs interrogated in this report are from the ImmunoChip, a custom genotyping array based on robust genome-wide association study (GWAS) results obtained from 12 autoimmune diseases. The SNPs listed in Table 5 consistently point to the insulin component of T1D. The CSVM-RBF method highly ranked rs7158663, a SNP located on the maternal expressed gene 3 (Meg3) on chromosome 14q32.2, while CLR ranked it rather low. You et al showed that the downregulation of Meg3 is associated with impaired glucose tolerance and decreased insulin secretion in mice. Our results are validated in the TEDDY cohort, where decreased Meg3 expression is associated with development of islet autoimmunity or T1D. Conditional logistic regression highly ranked rs17388568 and rs4580644, SNPs that are located in the adenosine deaminase domain containing 1 (ADAD1) gene on chromosome 4q27 and in introns of the cluster of differentiation 38 (CD38) gene on chromosome 4p15.32. The rs17388568 SNP has been identified as a risk factor for T1D in the Wellcome Trust Case Control Consortium as well as in a follow-up study. The rs4580644 SNP is predicted to influence regulation based upon effects on enhancer and histone marks and DNAse hypersensitivity. CD38 plays a key role in insulin secretion and has been shown to differentiate individuals with and without T1D; in particular, anti-CD38 autoantibodies have been suggested as new diagnostic biomarkers in autoimmunity in diabetes.

On the whole, we have demonstrated that a wide range of classification algorithms can be used to correctly analyze and interrogate features of nested case-control studies provided the study design is accounted for with a prior data transformation. Through a simulation study and analysis of the TEDDY data, we have demonstrated that CLR is limited in the types of relationships it can model and can typically be outperformed by more sophisticated classification algorithms like SVM and NB. In particular, CLR and CSVM-RBF agreed on several potential biomarkers for IA, but the CSVM-RBF identified several other potential markers that have not been previously identified. We believe this demonstrates the potential to identify more meaningful biomarkers through the use of more analytical methods than just CLR.

Acknowledgements

The authors would like to thank PNNL scientist Jennifer Kyle for her help with filtering the lipidomic data.

ORCID iDs

Bryan Stanfill https://orcid.org/0000-0003-0612-5333
Lisa Bramer https://orcid.org/0000-0002-8384-1926
Ernesto S. Nakayasu https://orcid.org/0000-0002-4056-2695

Supplemental material

Supplemental material for this article is available online.

REFERENCES

1. Rose S, Laan MJ. Why match? investigating matched case-control study designs with causal effect estimation. Int J Biostat 2009;5:1.
2. Adewale AJ, Dinu I, Yasui Y. Boosting for correlated binary classification. J Comp Graph Stat 2010;19:140–153.
3. Conway A, Rolley JX, Fulbrook P, Page K, Thompson DR. Improving statistical analysis of matched case–control studies. Res Nurs Health 2013;36:320–324.
4. Breslow N, Day N, Halvorsen K, Prentice RL, Sabei C. Estimation of multiple relative risk functions in matched case-control studies. Am J Epidemiol 1978;108:299–307.
5. Hogg T, Perkau J, Zhao Y, Gustafson P, Wijnands JM, Tremlett H. Bayesian analysis of pair-matched case-control studies subject to outcome misclassification. Stat Med 2017;36:4196–4213.
6. Asafu-Adjei J, Mahlet GT, Coull B, et al. Bayesian variable selection methods for matched case-control studies. Int J Biostat 2017;13:0043.
7. Qian J, Payalvash S, Kemmling A, Lev MH, Schwamm LH, Betensky RA. Variable selection and prediction using a nested, matched case-control study: application to hospital acquired pneumonia in stroke patients. Biometrics 2014;70:153–163.
8. Balasubramanian R, , Andres Houseman E, Coull BA, et al. Variable importance in matched case–control studies in settings of high dimensional data. J Royal Stat Soc: Series C 2014;63:639–655.
9. Avalos M, Pouyou H, Grandvalet Y, Otriols L, Lagarde E. Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple algorithm. BMC Bioinform 2015;16:51.
10. Breiman L. Random forests. Mach Learn 2001;45:5–32.
