Model Assisted Statistics and Applications 12 (2017) 227–237
DOI 10.3233/MAS-170397
IOS Press

A new framework for prediction and variable selection for uncommon events in a large prospective cohort study

Hye-Seung Lee* and Jeffrey P. Krischer
Health Informatics Institute, University of South Florida, Tampa, FL, USA

Abstract. When prediction is a goal, validation utilizing data outside of the prediction effort is desirable. Typically, data is split into two parts: one for a development and one for validation. But this approach becomes less attractive when predicting uncommon events, as it substantially reduces power. When predicting uncommon events within a large prospective cohort study, we propose the use of a nested case-control design, which is an alternative to the full cohort analysis. By including all cases but only a subset of the non-cases, this design is expected to produce a result similar to the full cohort analysis. In our framework, variable selection is conducted and a prediction model is fit on those selected variables in the case-control cohort. Then, the fraction of true negative predictions (specificity) of the fitted prediction model in the case-control cohort is compared to that in the rest of the cohort (non-cases) for validation. In addition, we propose an iterative variable selection using random forest for missing data imputation, as well as a strategy for a valid classification. Our framework is illustrated with an application featuring high-dimensional variable selection in a large prospective cohort study.

Keywords: Nested case-control, high dimensional variable selection, validation, penalized regression, random forest imputation

1. Introduction

In a large prospective cohort study, prediction and variable selection important to an outcome are frequently investigated. Variables are often selected because of their higher predictive power based on some measure of prediction performance. However, the predictive power needs to be assessed in data that was not used for the technical development. In epidemiological studies, prediction and variable selection are often criticized for lacking proper validation (Collins et al., 2014; Bleeker et al., 2003). To make an out-of-data validation available, one could split the full cohort data proportionally into two: a training data set to perform a prediction and variable selection development and a validation data set to validate the development. However, when the event of interest is uncommon, this approach is not attractive since it limits statistical power in both training and validation data sets.

In this paper, instead of splitting the cohort, we propose to use a nested case-control design that includes all the cases of interest but selects controls for a case among those subjects who were event free at the case's event time in the full cohort (risk-set matched case-control design). This design is expected to produce a similar result that the full cohort analysis would have produced (Samet & Munoz, 1998; Wacholder, 1991). We use the case-control cohort as the training data set without losing statistical power in the development and the rest of the cohort as the validation data set. Then, since all cases are included in the case-control cohort, only partial validation is available based on the fraction of true negative prediction (specificity). However, when predicting uncommon clinical events, limiting false positive prediction (1-specificity) is often more of interest than limiting false negative prediction. In the nested

*Corresponding author: Hye-Seung Lee, Health Informatics Institute, 3650 Spectrum Blvd., Suite 100, University of South Florida, Tampa, FL 33612, USA. E-mail: leeh@epi.usf.edu.

ISSN 1574-1699/17/$35.00

case-control cohort, we conduct variable selection and fit a prediction model on those selected variables. Then, we validate the selection by comparing the specificity of the fitted prediction model in the case-control subjects (internal) to that in the subjects who were not selected as controls within the cohort (external).

On the other hand, a prospective cohort study often gathers extensive data to investigate the nature of exposures and their relations to a clinical outcome. Variable selection is commonly used as a data driven search to find exposures relevant to an outcome. One may look for significant variables through exhaustive analyses of one variable at a time, but this inflates false positive findings and ignores correlations between variables. A standard approach analyzing multiple variables together is to do a stepwise selection by fitting a multiple regression model. However, when interactions between exposures are considered, the dimensionality increases exponentially, and such selection becomes unavailable or shows poor performance (Wiegand, 2010). High-dimensional data is common in genetic research and imaging studies, and machine learning methods have been extensively used for variable selection in those fields. They have also attracted increasing attention in epidemiologic studies, but interpretation has been difficult (Strobl et al., 2009; Speiser et al., 2015; Lu & Petkov, 2014).

For high dimensional variable selection, utilizing a nested case-control design not only makes external validation available without losing statistical power but also provides ways to control confounders and interpret the selection through a fitted prediction model. Featuring high-dimensional variable selection, we propose a new variable selection strategy to address missing data issues/problems. By repeating the variable selection process in data imputed using a random forest technique, we select the variables consistently included in those repetitions. Through comparisons between the internal and external specificities, prediction and variable selection are directly assessed, and a cut-point for a valid classification is determined where internal and external specificities are equivalent at a desired specificity. Our framework is illustrated through an example from a large prospective cohort study.

2. Materials and methods

In this section, we review existing methods to conduct prediction and variable selection in a matched case-control design.

2.1. Design and analysis

A nested case-control design is a matched case-control design including all cases. When matching is used, statistical analysis should take into account the condition matched per case-control pair. In unmatched case-control studies, logistic regression analysis is well accepted because it provides consistent estimates and standard errors of all coefficients, except for the intercept term (Prentice & Pyke, 1979; Farewell, 1979; Carroll et al., 1995). However, in matched case-control studies, the standard logistic regression analysis can produce severely biased estimates from a mis-specified function of the matching factors that appears in the model (Levin & Paik, 2001). Thus, for matched case-control studies, conditional likelihood methods are the analyses of choice, as they cancel out the effect of the matching factors in estimating the parameters of interest (Fleiss et al., 2003). Although matched case-control designs are commonly used in epidemiologic studies, it is relatively recent that studies dealing with high dimensional data have started acknowledging that the analysis should reflect the design (Rundle et al., 2012; Lee et al., 2014). There are only five published papers that considered the matched design in high dimensional variable selection (Balasubramanian et al., 2014; Qian et al., 2014; Tan et al., 2007; Adewale et al., 2010; Reid & Tibshirani, 2014). This is because variable selection is motivated to select factors with the best prediction of an outcome, but a matched case-control design is not developed for prediction.

2.2. Prediction model for a matched case-control design

Prediction is usually not a goal when a matched case-control design is used. But as indicated in (Qian et al., 2014), a prediction model can be fit when a matched case-control design is set up in a prospective cohort study.

When the odds is considered in a matched case-control design, it can be written as follows:

$$\frac{P(Y=1|X,Z,V=1)}{P(Y=0|X,Z,V=1)} = \frac{P(V=1|Y=1,X,Z)}{P(V=1|Y=0,X,Z)} \frac{P(Y=1|X,Z)}{P(Y=0|X,Z)}$$

where Y is an outcome ($Y=1$ for cases and $Y=0$ for controls), X is the vector of the selected variables, Z is the vector of matching variables, and $V=1$ indicates that a subject is sampled into the matched case-control study and $V=0$ indicates otherwise.

Here,

$$\frac{P(Y=1|X,Z)}{P(Y=0|X,Z)}$$

is the odds if the full cohort is analyzed, which can be fit using unconditional logistic regression $logit P(Y=1|X,Z) = \alpha + \beta^T X + \gamma^T Z$. As for the term

$$\frac{P(V=1|Y=1,X,Z)}{P(V=1|Y=0,X,Z)},$$

if we assume the sampling of cases and controls is independent of X, it can be written as

$$\frac{P(V=1|Y=1,Z)}{P(V=1|Y=0,Z)},$$

which is the same as

$$\frac{P(Y=0|Z)}{P(Y=1|Z)} \frac{P(Y=1|V=1,Z)}{P(Y=0|V=1,Z)}.$$

Then, depending on the size of design (say 1 to m matched design),

$$\frac{P(Y=1|V=1,Z)}{P(Y=0|V=1,Z)} = \frac{1}{m}.$$

Hence,

$$logit P(Y=1|X,Z,V=1) = \alpha + \beta^T X + \gamma^T Z + f(Z)$$

where

$$f(Z) = \log \frac{P(Y=0|Z)}{P(Y=1|Z)} - \log m.$$

Generally,

$$\frac{P(Y=0|Z)}{P(Y=1|Z)}$$

is unknown in matched case-control studies, but it is estimable if the matched case-control study was constructed from a prospective cohort study. Then, unconditional logistic regression model with $\hat{f}(Z)$ as offset can be fit for prediction, where $\hat{A}$ denotes an estimate of A.

Prediction performances must be measured using new data that was not used for developing a prediction model (Steyerberg et al., 2010). Based on the property of discrimination, the predictive power can be assessed by the fraction of true positive prediction (sensitivity) and the fraction of true negative prediction (specificity) at a given cut-point for classification. The area of under the Receiver Operating Characteristic curve (AUC) is commonly used, which plots sensitivity against 1-specificity for consecutive cut-points for the probability of an outcome. From the AUC, the Youden Index determines the cut-point for classification where the sum of sensitivity and specificity -1 is the highest (Youden 1950). The sensitivity and specificity at the cut-point determined by the Youden Index is often reported with the AUC to assess the predictability. These performance measures are expected to be the best when they are obtained from the data that was used for the prediction model development.

2.3. Variable selection for a matched case-control design

In a matched case-control design, variable selection can be done using a conditional logistic regression model. When multiple variables are analyzed together, a stepwise selection can be used by fitting a multiple conditional logistic regression model. But this approach becomes unstable or infeasible with a large number of variables. For high-dimensional variable selection, the penalized likelihood method with the lasso penalty (LASSO) is frequently used, since this method produces a subset of variables with nonzero regression coefficients (Tibshirani, 1996). Another common approach is random forest (RF) since it produces the so-called "variable importance (VIMP) score" for each input variable (Breiman, 2001). Recently, two approaches for a matched case-control design have been published based on a penalized conditional likelihood. One is a penalized conditional logistic regression model proposed by (Reid & Tibshirani, 2014). This approach uses a cyclic coordinate descent algorithm (Friedman et al., 2010; Wu & Lange, 2008) and allows the choice of the lasso, ridge, or elastic net penalty. The other is an algorithm proposed by (Balasubramanian et al., 2014) to obtain VIMP scores in a matched case-control design, using the ridge penalty for a penalized conditional logistic regression. Applying the RF technique, the average VIMP score is obtained for each variable over all bootstrap replicates in which the variable was selected as one of the selected variables.

3. Proposed framework

In our framework, in order to make the out-of-data validation available within a cohort, we set up a nested case-control design for developing a prediction model, and then the rest of the cohort is used for validating the model. In the case-control cohort, a set of variables are first selected using a method appropriate for a matched case-control design. For the selected variables, the parameter of interest β is estimated by fitting a multiple conditional logistic regression. This independent estimation of β makes the estimation robust to other parameter estimation. Then, we fit the prediction model using unconditional logistic regression with $\hat{\beta}^T X + f(\hat{Z})$ as offset. Finally, the selection is evaluated by assessing the performance of the prediction model internally using the subjects included in the case-control study and externally using the subjects in the rest of the cohort.

3.1. Validation

In order to assess whether internal predictive power remains the same in external evaluation, we propose to use a graphical approach, overlapping the internal and external specificities at consecutive cut-points for the predicted probability of the event. Better overlap will indicate better overall validity of prediction and variable selection. Also, greater area under the curve will indicate better specificity of the prediction fitted on the selected variables. We obtain the area under the curve for specificity plot using the method by (Gagnon & Peterson, 1998). For a valid classification, we propose to choose the cut-point for classification where the difference between internal and external specificities can be ignored at the maximum of desired specificity, through an exhaustive search.

3.2. Missing data

When conducting variable selection, subjects with a missing data are generally removed. The loss is greater with more input variables possibly including missing data. In a matched design, the loss is even greater because a subject with a missing variable leads to removing pairs including the subject. RF can be used for imputation of missing values and performs well in mixed types of variables where complex interactions and nonlinear relationships are suspected (Ishwaran et al., 2008; Stekhoven & Breiman, 2012). Here, RF is applied to data roughly imputed by the mean or frequency and produces a proximity matrix, which is a symmetric matrix whose (i, j) entry is the frequency that two subjects i and j occur within the same terminal node. Then the missing values are imputed by the proximity weighted average of the non-missing data. RF is again applied to the updated data, and the imputation is repeated to get a stable solution (RF imputed data). The imputed data can be used for further analysis. Implementing RF imputation for missing data, we conduct a multiple imputation that repeats variable selection in imputed data sets and choose variables jointly selected from repeated variable selections. In each repetition, we use unsupervised RF to generate the RF imputed data, and then variables are selected by an available method. After a desired number of repetitions, among those variables selected contingent on imputation, we order the variables from the most frequently selected to the least. Our approach selects variables from the highest rank until the frequency drops at the maximum.

Table 1
Design and missing data

Design	Planned			Without a missing variables among 136 input variables	
	% of controls in case-control cohort	Number of subjects	Number of cases/number of pairs	Number of subjects	Number of cases/number of pairs
1 to 1	3%	218	109/109	199	98/91
1 to 3	8%	436	109/327	401	98/276
1 to 10	25%	1199	109/1090	1110	98/915

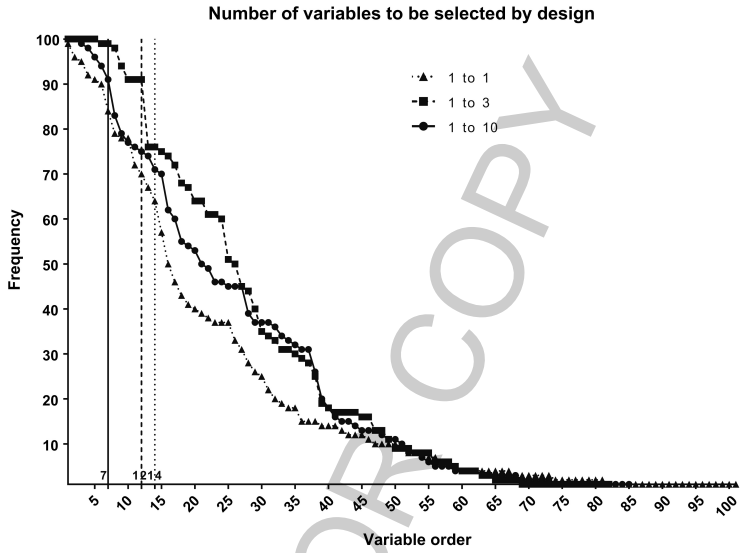


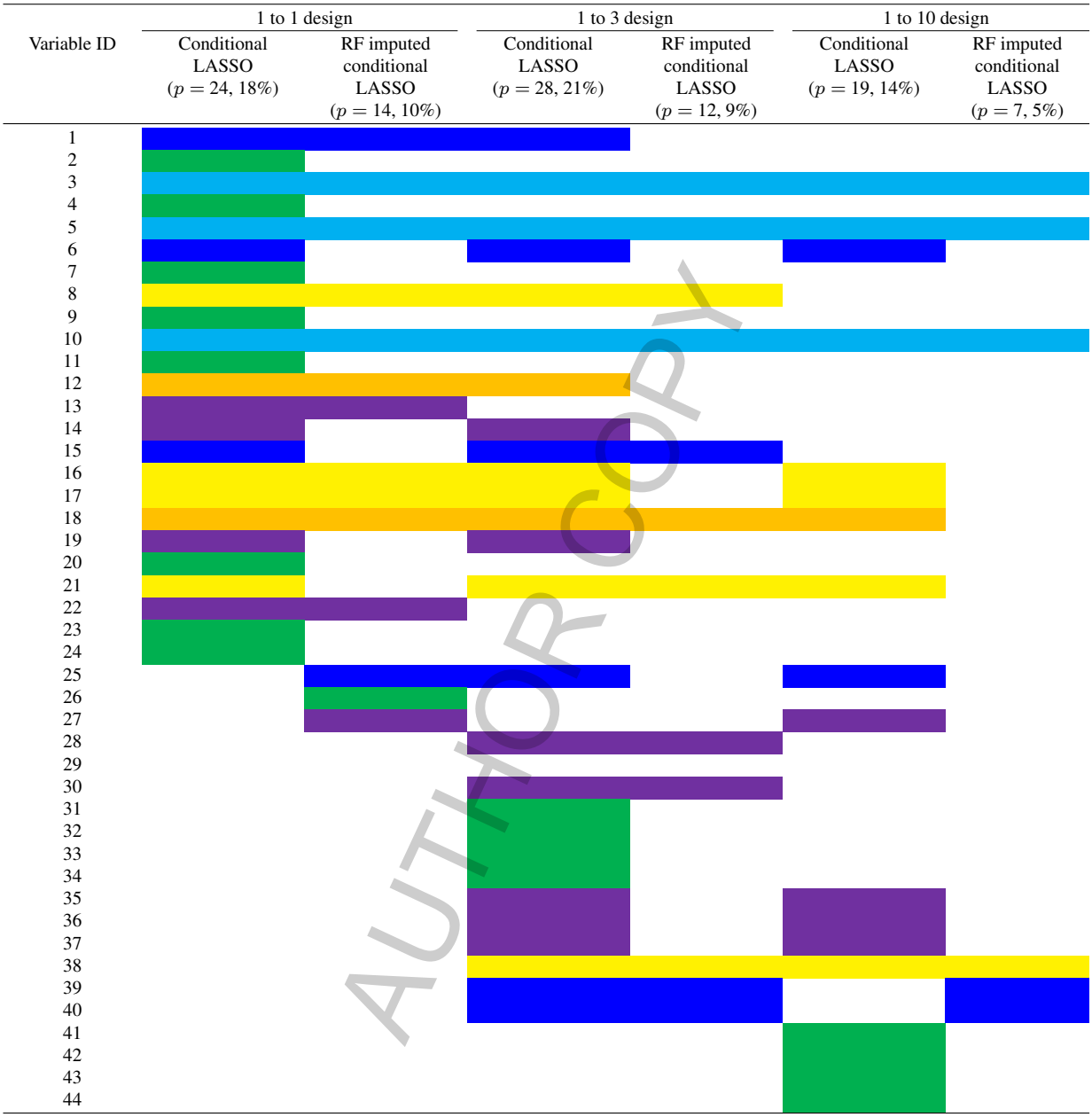
Fig. 1. Variable frequency in RF imputed conditional LASSO.

4. Application

To illustrate our proposed framework, an example was used from a prospective cohort study including 109 cases and 4,326 non-cases. Total 136 variables were used as input variables. After matching, we considered three study designs by varying the number of controls per case to be randomly selected from non-cases: 1 to 1 matched (3% of control selection); 1 to 3 matched (8%); and 1 to 10 matched (25%). Of the 109 cases, 98 had complete data for all 136 variables. The 1 to 1 matched design included 109 pairs, but the complete data for all 136 variables were available only in 91 pairs (83%) from 199 subjects (91%). That is, the proportion of missing was about 9% at the subject level but became 17% at the pair level. In this example, the data completeness in the input variables was similar between designs; 84% of pairs and 92% of subjects in the 1 to 3 design and 84% of pairs and 93% of subjects in the 1 to 10 design (Table 1).

For variable selection, we used the approach for a matched case-control design proposed by (Reid & Tibshirani, 2014): Conditional LASSO. Conditional LASSO selected 24 variables (18%) of the 136 input variables in the 1 to 1 design, 28 variables (21%) in the 1 to 3 design, and 19 variables (14%) in the 1 to 10 design by using the complete data only. Next, our proposed framework was applied to handle missing data. Conditional LASSO was repeated in 100 RF imputed data sets: RF imputed conditional LASSO. The 100 repetitions selected 101 variables at least once in the 1 to 1 design, 82 in the 1 to 3 design, and 85 in the 1 to 10 design. By ordering those variables from the most frequent selection to the least, the maximum separation occurred after the 14^{th} variable in the 1 to 1 design, the 12^{th} variable in the 1 to 3 design, and the 7^{th} variable in the 1 to 10 design (Fig. 1). Hence, the most frequently appearing 14 variables in the 1 to 1 design (10%), 12 variables (9%) in the 1 to 3 design, and 7 variables (5%) in the 1 to 10 design were selected by RF imputed conditional LASSO.

Table 2
Variable selection: Green – once, Purple – twice, Blue – three times; Yellow – four times, Orange – five times, Light blue- six times; (p = number of selected variables, percentage of variable selection)



As shown in Table 2, of the 136 input variables, 44 variables were selected at least once out of the 6 variable selections attempted. Irrespective of the design, a random selection appearing only once (green) happened more in Conditional LASSO, compared to RF imputed conditional LASSO (33% vs. 7% in the 1 to 1 design, 14% vs. 0% in the 1 to 3 design, and 21% vs. 0%). Three variables were selected in all of the 6 selections (light blue). Two variables appeared in 5 of the 6, 5 variables did in 4 of the 6, 6 variables did in 3 of the 6, and 10 variables did in 2 of the 6. The RF imputed conditional LASSO selection showed a good overlap with the Conditional LASSO selection: 79% in the 1 to 1 design, 100% in the 1 to 3 design and 71% in the 1 to 10 design.

The proposed prediction model was built on those variables selected by each approach. Internal validation within

Table 3
Prediction and validation

Design	Validation	Conditional LASSO		RF imputed conditional LASSO	
		Internal	External	Internal	External
1 to 1	Number of variables	24		14	
	AUC	0.774	n.a.	0.759	n.a.
	Sensitivity by the Youden index	0.684	n.a.	0.818	n.a.
	Specificity by the Youden index**	0.723	0.649	0.624	0.563
	Area under the specificity curve	0.686	0.610	0.641	0.617
1 to 3	Number of variables	28		12	
	AUC	0.822	n.a.	0.770	n.a.
	Sensitivity by the Youden index	0.827	n.a.	0.776	n.a.
	Specificity by the Youden index	0.726	0.621*	0.686	0.614*
	Specificity by the proposed approach to find a valid classification**	0.574	0.517	0.637	0.581
	Area under the specificity curve	0.828	0.755	0.805	0.760
1 to 10	Number of variables	19		7	
	AUC	0.752	n.a.	0.713	n.a.
	Sensitivity by the Youden index	0.857	n.a.	0.816	n.a.
	Specificity by the Youden index**	0.559	0.554	0.546	0.523
	Area under the specificity curve	0.916	0.907	0.914	0.909

Internal, from case-control cohort; External, from the rest of cohort; AUC, Area under the ROC curve; Sensitivity, fraction of true positive prediction; Specificity, fraction of true negative prediction; *Significant difference between internal and external; **Valid classification.

Table 4
Prediction and validation in the full cohort

Design	Validation	Variable selection	
		Conditional LASSO	RF imputed conditional LASSO
1 to 1	AUC	0.705	0.728
	Valid sensitivity	0.684	0.818
	Valid specificity	0.651	0.566
	Number of variables whose univariate p-value < 0.01	9 (38%)	8 (57%)
1 to 3	AUC	0.754	0.721
	Valid sensitivity	0.867	0.786
	Valid specificity	0.521	0.585
	Number of variables whose univariate p-value < 0.01	11 (39%)	6 (50%)
1 to 10	AUC	0.733	0.701
	Valid sensitivity	0.857	0.816
	Valid specificity	0.525	0.529
	Number of variables whose univariate p-value < 0.01	10 (53%)	5 (71%)

the case-control cohort was conducted and compared to external validation in the cohort but not in the case-control study (Table 3). In the internal validation, the AUC was higher in Conditional LASSO than in RF imputed conditional LASSO in all designs, suggesting worse classification in RF imputed conditional LASSO. Although the AUC was generally higher with more variables, RF imputed conditional LASSO in the 1 to 3 design produced a higher AUC value with 12 variables (77%), compared to RF imputed conditional LASSO in the 1 to 1 design (76% with 14 variables) or Conditional LASSO in 1 to 10 design (75% with 19 variables). We obtained the sensitivity and specificity at the cut-point determined by the Youden Index. The specificity was lower in RF imputed conditional LASSO than Conditional LASSO in all designs.

Except for the 1 to 3 design, the external specificity was not significantly different from the internal specificity, implying that the classification by the Youden Index is valid. For overall validation, we plotted those internal and external specificities against consecutive cut-points. The internal and external specificities overlapped better in RF imputed conditional LASSO. In the 1 to 1 design, although the internal and external specificities were similar, the area under the specificity curve was 61 to 69%, requiring a higher cut-point to reach an acceptable specificity. The 1 to 10 design showed a desirable shape with higher area under the specificity curve (91 to 92%), but selecting 10 controls per case (25% of control selection) may not be necessary or feasible in most studies. The 1 to 3 design also showed a good shape with relatively high area under the specificity curve (76 to 83%), but the internal and external

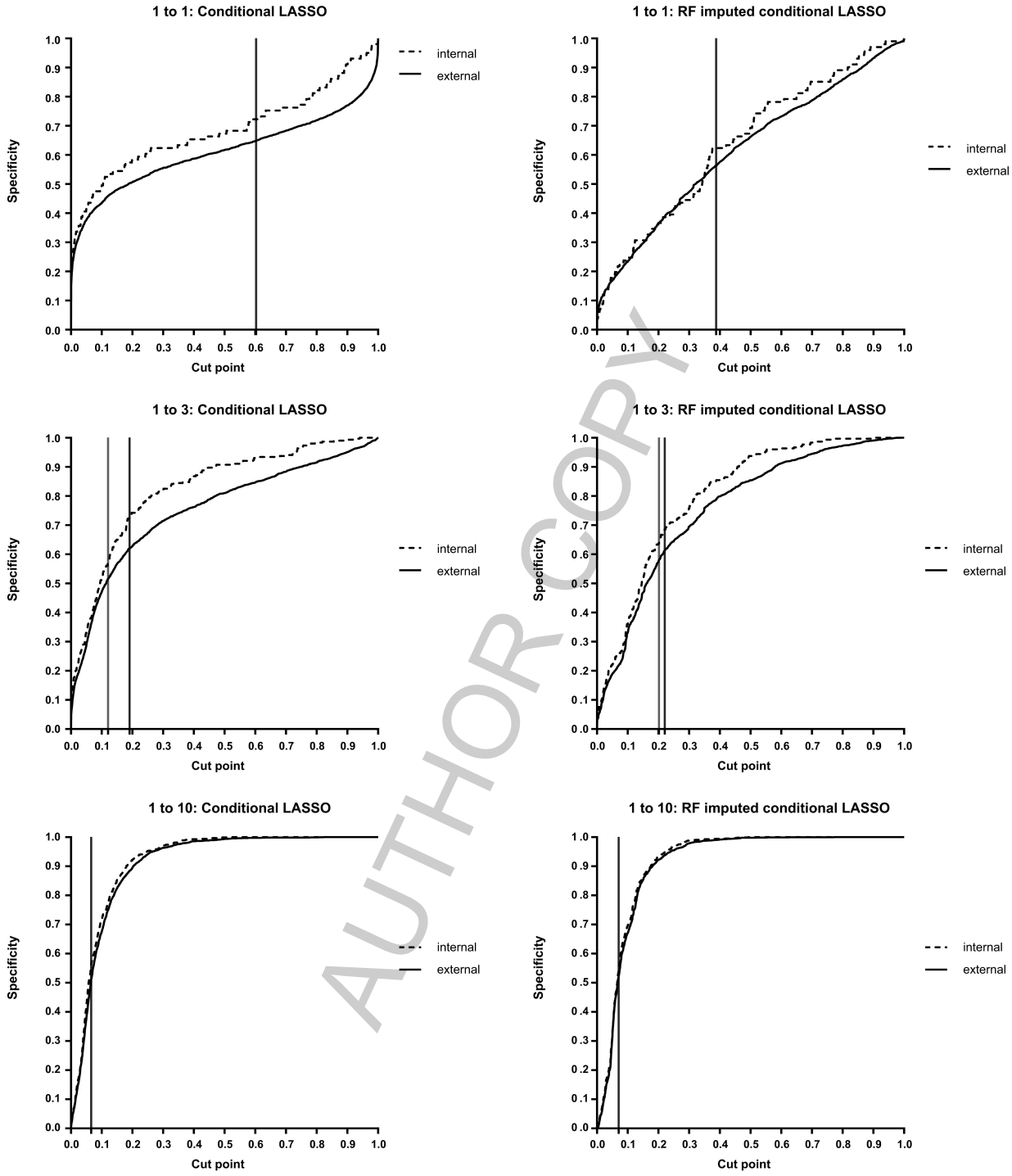


Fig. 2. Internal and external specificities: Black reference line indicates the cut-point by the Youden index, and grey reference line indicates the proposed cut-point.

specificities by the Youden index were significantly different. Hence, our proposed approach was used to find the cut-point for a valid classification (reference line in grey in Fig. 2).

Finally, we applied each prediction model in the full cohort (Table 4). As expected, the AUC decreased in both

approaches, but the change in RF imputed conditional LASSO was less in all designs. Since the sensitivity is the same between the case-control cohort and the full cohort, this AUC reduction is due to the specificity change. The trend of higher AUC with more variables remained. But when the lowest AUC was 70% with 7 variables by RF imputed conditional LASSO in the 1 to 10 design, the 1 to 1 design by Conditional LASSO produced 71% of the AUC with 24 variables. The sensitivity and specificity were obtained at the cut-point determined to be a valid classification in each design. The valid specificity was similar across designs in RF imputed conditional LASSO (57% in the 1 to 1 design, 59% in the 1 to 3 design and 53% in the 1 to 10 design). Additionally, the independent effect of the selected variables was examined using the standard unconditional logistic regression after adjusting for the matching factor in the full cohort. When those variables with p-value < 0.01 were counted, higher proportion was shown in RF imputed conditional LASSO in all designs.

Internal validation suggested better performance in Conditional LASSO, but external validation suggested otherwise. In both variable selection approaches, the area under the specificity curve suggested that selecting 3 or more controls is reasonable for overall validity. Variable selection and prediction by RF imputed conditional LASSO were more consistent in external evaluation by including fewer variables with a higher proportion of a significant independent effect. Hence, in this example, the RF imputed conditional LASSO prediction and variable selection in the 1 to 3 design seem to be an efficient choice when a simpler prediction model or variable selection is desired. In our application, the prediction developed in the 1 to 3 design is expected to be 59% specific and 79% sensitive.

5. Discussion

Instead of the full cohort analysis, our framework constructs a nested case-control design to make an appropriate validation available. By including all cases, this design preserves statistical power for prediction and variable selection, but the validation is restricted to specificity only. When the outcome is common, our framework may not help much, compared to the typical approach of splitting data into a training set and a validation set. However, when the outcome is uncommon, the possibility of misclassifying negative subjects as positive is more of a concern since the error can affect more subjects. In this context, our framework can be preferred as it controls specificity in the development.

The impact of missing data can be greater in a matched case-control study, due to the exclusion of pairs. In high-dimensional data analysis, the algorithms use a cross-validation to determine the amount of regularization by repeating a sub-sampling approach. This process produces generally unstable selection contingent on the sub-sampling (Meinshausen & Buhlman, 2010). Also, more limited sample size compared to the number of input variables tends to include more variables, leading to over-fitting the prediction model. By including variables repeatedly selected through iterations, our proposed approach selected fewer variables with a higher proportion of variables with independent effects, irrespective of the design. While this could be an indication of improving stability in variable selection and prediction, a simpler model not only helps interpretation, but also reduces the impact of missing data.

Prediction performance is expected to be worse when they were measured in data that was not used for the development, compared to the internal assessment. Our proposed framework also showed that the internal AUC in the nested case-control cohort was an over-estimate when it was validated in the full cohort. However, in RF imputed conditional LASSO, the AUC reduction was smaller (about 7% vs. 5%) and internal specificity was similar to external specificity, indicating better validity. The sensitivity and specificity were also similar across different designs when the proposed classification was applied for validity.

In this paper, our framework was illustrated in the use of high-dimensional data where the number of variables is close to or moderately higher than the number of subjects. But this approach also helps with low-dimensional data with appropriate variable selection. The LASSO selection was arbitrarily chosen, so any variable selection procedure can replace the LASSO selection. In a matched case-control design, the methods for high dimensional variable selection are currently limited. Since our framework separates variable selection from fitting the prediction model for the matched design, variable selection ignoring the matched design may not have a severe impact on the prediction model performance. A further investigation can be pursued. On the other hand, a nested case-control design is mostly used in the need of costly data collection. In this context, data is only available in the case-control

cohort, so our framework using external specificity is not directly available. However, one may use our framework for available data in the full cohort to guide in validating the findings from data only available in the case-control cohort. In future work, we will investigate the use of our framework when data are available only in the case-control cohort.

Acknowledgements

The work described in this article was 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, and Contract No. HHSN267200700014C 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), Juvenile Diabetes Research Foundation (JDRF), and Centers for Disease Control and Prevention (CDC).

References

Adewale, A. J., Dinu, I., & Yasui, Y. (2010). Boosting for correlated binary classification. *Journal of Computational and Graphical Statistics*, *19*, 140-153.

Balasubramanian, R., Houseman, E. A., Coull, B., Lev, M. H., Schwamm, L. H., & Betensky, R. A. (2014). Variable importance in matched case-control studies in settings of high dimensional data. *Journal of the Royal Statistical Society: Series C*, *63*, 639-655.

Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R., Derksen-Lubsen, G., Grobbee, D. E., & Moons, K. G. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, *56*(9), 826-832.

Breiman L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Carroll, R. J., Wang, S., & Wang, C.-Y. (1995). Prospective analysis of logistic case-control studies. *J Am Statist Assoc*, *90*, 157-169.

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M., Moons, K. G., & Altman, D. G. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*, 40.

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, *66*, 27-32.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions. 3rd edition. Wiley.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear model via coordinate descent. *Journal of Statistical Software*, *33*(1), 1-22.

Gagnon, R. C., & Peterson, J. J. (1998). Estimation of confidence intervals for area under the surve from destructively obtained pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, *26*(1), 87-102.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*, 841-860.

Lee, H-S., Burkhard, B. R., McLeod, W., Smith, S., Eberhard, C., Lynch, K., Hadley, D., Rewers, M. J., Simell, O. G., She, J-X., Hagopian, B., Lernmark, A., Akolkar, B., Ziegler, A. G., Krischer, J. P., & The TEDDY study group. (2014). Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes/Metabolism Research and Reviews*, *30*, 424-434.

Levin, B., & Paik, M. C. (2001). The unreasonable effectiveness of a biased logistic regression procedure in the analysis of pair-matched case-control studies. *Journal of Statistical Planning and Inference*, *96*, 371-385.

Lu, F., & Petkova, E. (2014). A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in Medicine*, *33*, 401-421.

Meinshausen, N., & Buhlman, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, *72*, 417-473.

Prentice, R., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *65*, 153-158.

Qian, J., Payabvash, S., Kemmling, A., Lev, M. H., Schwamm, L. H., & Betensky, R. A. (2014). Variable selection and prediction using a nested, matched case-control study: Application to hospital acquired pneumonia in stroke patients. *Biometrics*, *70*(1), 153-163.

Reid, S., & Tibshirani, R. (2014). Regularization paths for conditional logistic regression: The clogitL1 package. *Journal of Statistical Software*, *58*, 1-23.

Rundle, A., Ahsan, H., & Vineis, P. (2012). Better cancer biomarker discovery through better study design. *European Journal of Clinical Investigation*, *42*(12), 1350-9.

Samet, J. M., & Munoz, A. (1998). Evolution of the cohort study. *Epidemiologic Reviews*, *20*(1), 1-14.

Speiser, J. L., Durkalski, V. L., & Lee, W. M. (2015). Random forest classification of etiologies for an orphan disease. *Statistics in Medicine*, *34*, 887-899.

Stekhoven, D. J., & Breiman, L. (2012). MissForest-nonparametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, *21*(1), 128-138.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*(4), 323-348.
Tan, Q., Thomassen, M., & Kruse, T. A. (2007). Feature selection for predicting tumor metastases in microarray experiments using paired design. *Cancer Information*, *3*, 213-218.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*, 267-288.
Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, *2*(2), 155-8.
Wiegand, R. E. (2010). Performance of using multiple stepwise algorithms for variable selection. *Statistics in Medicine*, *29*, 1647-1659.
Wu, T., & Lange, K. (2008). Coordinate descent procedures for lasso penalized regression. *The Annals of Applied Statistics*, *2*(1), 224-244.
Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32-35.