Chapter 15
Nested Case-Control Risk Score Prediction

Sherri Rose, Bruce Fireman, Mark J. van der Laan

Risk scores are calculated to identify those patients at the highest level of risk for an outcome. In some cases, interventions are implemented for patients at high risk. Standard practice for risk score prediction relies heavily on parametric regression. Generating a good estimator of the function of interest using parametric regression can be a significant challenge. As discussed in Chap. 3, high-dimensional data are increasingly common in epidemiology, and researchers may have dozens, hundreds, or thousands of potential predictors that are possibly related to the outcome.

The analysis of full cohort data for risk prediction is frequently not feasible, often due to the cost associated with purchasing access to large comprehensive databases, storage and memory limitations in computer hardware, or other practical considerations. Thus, researchers frequently conduct nested case-control studies instead of analyzing the full cohort, particularly when their prediction research question involves a rare outcome. This type of two-stage design introduces bias since the proportion of cases in the sample is not the same as the population. This complication may have contributed to the relative lack of prediction studies for rare diseases.

We consider a two-stage sampling design in which one takes a random sample from a target population and measures Y, the outcome, on each subject in the first stage. The second stage involves drawing a subsample from the original sample, collecting additional data on the subsample. The decision regarding selection into the subsample is influenced by Y. This data structure can be viewed as a missing-data structure on the full data structure X collected in the second stage of the study. Using nested case-control data from a Kaiser Permanente database, we generate a function for mortality risk score prediction using super learner and inverse probability of missingness weights to correct the bias introduced by the sampling design.

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_15,

15.1 Data, Model, and Parameter

Kaiser Permanente Northern California provided medical services to approximately 3 million members during the study period. They served 345,191 persons over the age of 65 in the 2003 calendar year, and 13,506 of these subjects died the subsequent year. The death outcome was ascertained from California death certificate filings. Disease and diagnosis variables, which we refer to in this paper simply as medical flags, were obtained from Kaiser Permanente clinical and claims databases. There are 184 medical flags covering a variety of diseases, treatments, conditions, and other reasons for visits. Gender and age variables were obtained from Kaiser Permanente administrative databases.

A nested case-control sample was extracted from the Kaiser Permanente database for computational ease. All 13,506 cases from the 2003–2004 data were sampled with probability 1, and an equal number of controls were sampled from the full database with probability 0.041 for a total of 27,012 subjects. Approval from the institutional review board at Kaiser Permanente Northern California for the protection of human subjects was obtained.

Formally, we define the full data structure as $X = (W, Y) \sim P_{X,0}$, with covariate vector $W = \{W_1, \ldots W_{186}\}$ and binary outcome Y, indicating death in 2004. The observed data structure for a randomly sampled subject is $O = (Y, \Delta, \Delta X) \sim P_0$, where Y is included in X and Δ denotes the indicator of inclusion in the second-stage sample (nested case-control sample). The parameter of the full-data distribution of X is given by $\bar{Q}_0 = E_{X,0}(Y \mid W)$ and the full-data statistical model $\mathcal{M}^F$ is nonparametric.

15.2 Loss Function

Had our sample been comprised of n i.i.d. observations X_i, we would have estimated $\bar{Q}_0 = E_{X,0}(Y \mid W)$ with loss-based learning using loss function $L^F(X, \bar{Q})$. Given the actual observed data, we can estimate $\bar{Q}_0$ with super learning and weights $\Delta_i / P_{X,n}(\Delta_i = 1 \mid Y_i)$ for observations $i = 1, \ldots, n$, which corresponds with the same super learner, but now based on the inverse probability of missingness (censoring) weighted loss function:

$$L(O, \bar{Q}) = \frac{\Delta}{P_{X,n}(\Delta = 1 \mid Y)} L^F(X, \bar{Q}).$$

We define our parameter of interest as: $\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L(O, \bar{Q})$, where $\bar{Q}$ is a possible function in the parameter space of functions that map an input W into a predicted value for Y. $E_0 L(O, \bar{Q})$, the expected loss, evaluates the candidate $\bar{Q}$, and it is minimized at the optimal choice of $\bar{Q}_0$.

15.3 Data Analysis

We implemented super learning with observation weighting in R to obtain our estimate of $\bar{Q}_0$ using our observed data. Observation weights within the super learner were assigned based on the inverse probability of missingness, $w_i = \Delta_i / P_{X,n}(\Delta_i = 1 \mid Y_i)$ thus cases were given observation weights equal to 1 and controls were given observation weights of $1/0.041 = 24$. One could further stabilize the weights by standardizing them to sum to 1: in other words, we would divide the above w_i by $\sum_{i=1}^{n} \Delta_i / P_{X,n}(\Delta = 1 \mid Y_i)$. Any algorithm that allows observation weighting can be used with super learner in nested case-control data.

The collection of 16 algorithms included in this analysis can be found in Table 15.1. We implemented dimension reduction among the covariates as part of each algorithm, retaining only those covariates associated with Y in a univariate regression (p< 0.10). After screening, 135 covariates remained. Algorithms with different options (e.g., degree, size, etc.) were considered distinct algorithms. The selection of these algorithms was based on investigator knowledge, the ability to take observation weights, and computational speed. The super learner algorithm is explained in detail in Chap. 3, and we refer readers to this chapter for an intuitive understanding of the procedure. Demonstrations of the super learner's superior finite sample performance in simulations and publicly available data sets, as well as asymptotic results, are also discussed in Chap. 3.

A summary of the nested case-control variables can be found in Table 15.2. All 187 variables, except death, were evaluated from 2003 records. The majority of the sample is female, with 45.2% male. The age category with the largest num-

Table 15.1 Collection of algorithms

Algorithm	Description
glm.1	Main terms logistic regression
glm.2	Main terms logistic regression with gender × age interaction
glm.3	Main terms logistic regression with gender × age^2 interaction
glm.4	Main terms logistic regression with gender × age^3 interaction
glm.5	Main terms logistic regression with age^2 term
glm.6	Main terms logistic regression with age^3 term
glm.7	Main terms logistic regression with age × covariate interaction for remaining main terms
glm.8	Main terms logistic regression with gender × covariate interaction for remaining main terms
glm.9	Main terms logistic regression with age × covariate and gender × covariate interaction
bayesglm	Bayesian main terms logistic regression
glmnet.1	Elastic net, $\alpha = 1.00$
glmnet.5	Elastic net, $\alpha = 0.50$
gam.2	Generalized additive regression, degree = 2
gam.3	Generalized additive regression, degree = 3
nnet.2	Neural network, size = 2
nnet.4	Neural network, size = 4

ber of members was 70 to 79, with 41.0%. (For presentation, age is summarized categorically in Table 15.2, although the variable is continuous and was analyzed as a continuous variable. All other variables are binary.) The top ten most prevalent medical flags in the sample were: screening/observation/special exams, other endocrine/metabolic/nutritional, hypertension, minor symptoms, postsurgical status/aftercare, major symptoms, history of disease, other musculoskeletal/connective tissue, cataract, and other dermatological disorders. The majority of medical flags (47.2%) had a prevalence of less than 1%. Twenty medical flags had a prevalence of 0%. These variables were excluded from our analysis as they provide no information. We remind the reader that these percentages do not reflect estimates of prevalence in the *population* given the biased sampling design.

The super learning algorithm for predicting death (risk score) in the nested case-control sample performed as well as or outperformed all single algorithms in the collection of algorithms. With a cross-validated MSE (i.e., the cross-validated risk, not to be confused with *risk score*) of 3.336e-2, super learner improved upon the

Table 15.2 Characteristics of Northern California Kaiser Permanente members aged 65 years and older in nested case-control sample, 2003

Variables	No.	%
Death (in 2004)	13,506	50.0
Male	12,213	45.2
Age, years[a]		
65 to <70	5,193	19.2
70 to <80	11,077	41.0
80 to <90	8,525	31.6
≥ 90	2,217	8.2
Most prevalent medical flags	No.	%
Screening/observation/special exams	23,597	87.4
Other endocrine/metabolic/nutritional	10,633	39.4
Hypertension	10,612	39.3
Minor symptoms, signs, findings	9,748	36.1
Postsurgical status/aftercare	9,447	35.0
Major symptoms, abnormalities	8,251	30.5
History of disease	7,376	27.3
Other musculoskeletal/connective tissue	7,359	27.2
Cataract	5,976	22.1
Other dermatological disorders	5,692	21.1
Medical flag prevalence	No.	%
Zero	20	10.8
$0 < x < 1\%$	67	36.4
$1 \leq x < 10\%$	72	39.1
$\geq 10\%$	25	13.6

[a] Age is summarized categorically although the variable is continuous.

Table 15.3 Results from super learner analysis

Algorithm	CV MSE	RE	R^2
SuperLearner	3.336e-2	–	0.113
glm.1	3.350e-2	1.004	0.109
glm.2	3.350e-2	1.004	0.109
glm.3	3.349e-2	1.004	0.109
glm.4	3.348e-2	1.004	0.109
glm.5	3.348e-2	1.004	0.109
glm.6	3.348e-2	1.004	0.109
glm.7	3.458e-2	1.037	0.080
glm.8	3.443e-2	1.032	0.084
glm.9	3.533e-2	1.059	0.060
bayesglm	3.778e-2	1.132	-0.005
glmnet.1	3.337e-2	1.000	0.112
glmnet.5	3.336e-2	1.000	0.112
gam.2	3.349e-2	1.004	0.109
gam.3	3.349e-2	1.004	0.109
nnet.2	3.913e-2	1.173	-0.041
nnet.4	3.913e-2	1.173	-0.041

worst algorithms by 17% with respect to estimated cross-validated MSE. MSEs in the collection of algorithms ranged from 3.336e-2 to 3.913e-2. While the collection of algorithms was somewhat limited, which isn't optimal from a theoretical perspective, we see some benefits in relative efficiency. Results are presented in Table 15.3 where relative efficiency for each of the k algorithms is defined as RE=cross validated MSE(k)/cross validated MSE(*super learner*).

When examining R^2 values, the super learner had the largest R^2 compared to the collection of algorithms with an $R^2 = 0.113$, although ten of the algorithms approached this value. Super learner had an 11.3% gain relative to using the marginal probability (i.e., assigning probability of death 0.039 to each observation). The algorithms in the collection had R^2 values ranging from 0.112 to -0.041. (Negative R^2 values indicate that the marginal prevalence probability is a better predictor of mortality than the algorithm. Values for R^2 can fall outside the range [0,1] when calculated in cross-validated data.) See Table 15.3. While the performance of the super learner improved upon the collection of algorithms with respect to R^2 values, it should be noted that the overall prediction power of this data set is somewhat limited with the best $R^2 = 0.113$.

15.4 Discussion

Alternatives to parametric approaches to risk score prediction include the flexible approach super learning. The algorithm provides a system to combine many estimators into an improved estimator and returns a function we can use for prediction in new data sets. Cross-validation of the individual algorithms and the super learner prevents overfitting and the selection of a fit that is too biased. Our criterion for estimator selection is based on an a priori established benchmark (e.g., cross-validated MSE).

Super learning allows for the use of observation weighting in order to generate prediction functions with nested case-control data, as well as data from other two-stage sampling designs, case-control designs, and general biased sampling designs. In our nested case-control Kaiser Permanente data, super learner performed as well as or outperformed all algorithms in the collection of algorithms. While the overall predictive power of this data set was limited ($R^2 = 0.113$), the utility of super learning is still apparent. In Chap. 3, larger improvements in cross-validated MSE were seen in other real data sets. The minimal improvement of the super learner in this analysis is not unexpected since the outcome is rare in the population of interest. This can be understood intuitively since any large improvement in predicting death by an algorithm among "case" subjects is averaged over the entire sample.

It is not possible to know with certainty a priori which single algorithm will perform the best in any given data set. Even when the result is a negligible improvement relative to the best algorithms in the collection, the super learner provides a tool for researchers to run many algorithms and return a prediction function with the best cross-validated MSE, avoiding the need to commit to a single algorithm.

For example, even in this analysis, had the logistic regression with main terms and age covariate and gender covariate interactions for each covariate (glm.9) been the a priori selected single algorithm, with $R^2 = 0.060$, its performance is poor compared to that of the super learner. Several other algorithms were considerably worse thanglm.9 and also could have been the single a priori selected algorithm. In other words, the use of the super learner prevents poor a priori algorithm choices.

15.5 Notes and Further Reading

Prediction has been used most notably to generate tables for risk of heart disease (Kannel et al. 1976; Anderson et al. 1991; Ramsay et al. 1995, 1996; Wilson et al. 1998; Jackson 2000) and breast cancer (Gail et al. 1989; Costantino et al. 1999; Tyrer et al. 2004; Barlow et al. 2006). An existing method for prediction in parametric statistical models with nested case-control samples is intercept adjustment. The addition of $\log(P_{X,0}(\Delta = 1 \mid Y = 1)/P_{X,0}(\Delta = 1 \mid Y = 0))$, or equivalently $\log(q_0/(1 - q_0))$, to the intercept in a logistic regression yields the true logistic regression function $P_{X,0}(Y = 1 \mid W)$, assuming the statistical model is correctly specified. Here Δ denotes the indicator of inclusion in the nested case-control sample, and the value q_0 is the prevalence probability $P_{X,0}(Y = 1) = q_0$ (Anderson 1972;

Prentice and Breslow 1978; Greenland 1981; Wacholder 1996; Morise et al. 1996; Greenland 2004).

We introduced a more flexible method for prediction in two-stage nested case-control data. This method is an application of the general loss-based super learner, the appropriate loss function is selected. It corresponds with an inverse probability of missingness full-data loss function. The method involves observation weights $w_i = \Delta_i / P_n(\Delta_i = 1 \mid Y_i)$ to eliminate the bias of the sampling design, where these weights are determined by the inverse probability of missingness. For nested case-control studies, this is equalivalent to using case-control weights, with cases assigned the weight q_n (an estimate of q_0 obtained from the full cohort) and controls assigned a weight of $(1 - q_n)/J$, where J is the average number of controls per case. Thus the choice of loss function can also be presented as the case-control-weighted loss function presented in the preceding two chapters, van der Laan (2008a), and Rose and van der Laan (2008, 2009).

One might also be interested in the effect of each medical flag on mortality, controlling for all other medical flags. This is a variable-importance research question, one where we can use a TMLE. In a recent paper, Rose and van der Laan (2011) describe the TMLE for two-stage designs. We also refer readers to Appendix A.