University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

A Note on Risk Prediction for Case-Control Studies

Sherri Rose*   Mark J. van der Laan†

*Division of Biostatistics, University of California, Berkeley, sherrirosephd@gmail.com

†University of California - Berkeley, laan@berkeley.edu

A Note on Risk Prediction for Case-Control Studies

Sherri Rose and Mark J. van der Laan

Abstract

We introduce a new method for prediction in case-control study designs, which is a simple extension of the work by van der Laan (2008). Case-control samples are biased since the proportion of cases in the sample is not the same as the population of interest. The case-control weighting for prediction proposed in this paper relies on knowledge of the true incidence probability P(Y=1) to eliminate the bias of the sampling design. In many practical settings, case-control weighting will outperform an existing method for prediction, intercept adjustment.

1 Introduction

The use of modeling for prediction has been well established in the literature. Risk prediction models have been used most notably to generate tables for risk of heart disease (Ramsay et al., 1995, 1996; Jackson, 2000). Recently, Whiteman and Green (2005) discussed the lack of studies for diseases such as cancer, that have any informative value for prediction. Models do exist, but many of these instruments have not been validated, or they may perform poorly for prediction at an individual level versus a population level. Comprehensive references for risk prediction models in certain types of cancers can be found at `http://riskfactor.cancer.gov/cancer_risk_prediction`.

What is common among all forms of cancer is a low incidence probability, and, as such, case-control studies are frequently performed. Many traditional risk modeling approaches for prediction (e.g. traditional logistic regression) are not effective when based on case-control study data since the study design produces a biased sample. The bias is due to the fact that the proportion of cases in the sample is not the same as the population of interest. This complication may have contributed to the relative lack of predictive modeling for rare diseases. Many of the published findings for prediction of rare diseases are based on the stratification of case-control samples.

We introduce a new method for developing predictive models with case-control study data, which is an extension of theories originally developed for causal inference in case-control study designs presented by van der Laan (2008). Our new prediction method involves implementing a simple weighted model to eliminate the bias of the sampling design, where the weights are determined by the prevalence probability $P_0^*(Y = 1) \equiv q_0$. We will compare our weighted models to the use of intercept adjusted logistic maximum likelihood estimation that also relies on knowledge of q_0.

2 Intercept Adjusted Maximum Likelihood Estimation

First presented by Anderson (1972), the addition of $\log \frac{q_0}{1-q_0}$ to a logistic regression model intercept yields the true logistic regression function $P_0^*(Y = 1 \mid W)$. The method has also been discussed elsewhere throughout the literature (Prentice and Breslow, 1978; Prentice and Pyke, 1979; Greenland, 1981; Benichou and Wacholder, 1994; Morise et al., 1996; Wacholder,

1

1 Introduction

The use of modeling for prediction has been well established in the literature. Risk prediction models have been used most notably to generate tables for risk of heart disease (Ramsay et al., 1995, 1996; Jackson, 2000). Recently, Whiteman and Green (2005) discussed the lack of studies for diseases such as cancer, that have any informative value for prediction. Models do exist, but many of these instruments have not been validated, or they may perform poorly for prediction at an individual level versus a population level. Comprehensive references for risk prediction models in certain types of cancers can be found at `http://riskfactor.cancer.gov/cancer_risk_prediction`.

What is common among all forms of cancer is a low incidence probability, and, as such, case-control studies are frequently performed. Many traditional risk modeling approaches for prediction (e.g. traditional logistic regression) are not effective when based on case-control study data since the study design produces a biased sample. The bias is due to the fact that the proportion of cases in the sample is not the same as the population of interest. This complication may have contributed to the relative lack of predictive modeling for rare diseases. Many of the published findings for prediction of rare diseases are based on the stratification of case-control samples.

We introduce a new method for developing predictive models with case-control study data, which is an extension of theories originally developed for causal inference in case-control study designs presented by van der Laan (2008). Our new prediction method involves implementing a simple weighted model to eliminate the bias of the sampling design, where the weights are determined by the prevalence probability $P_0^*(Y = 1) \equiv q_0$. We will compare our weighted models to the use of intercept adjusted logistic maximum likelihood estimation that also relies on knowledge of q_0.

2 Intercept Adjusted Maximum Likelihood Estimation

First presented by Anderson (1972), the addition of $\log \frac{q_0}{1-q_0}$ to a logistic regression model intercept yields the true logistic regression function $P_0^*(Y = 1 \mid W)$. The method has also been discussed elsewhere throughout the literature (Prentice and Breslow, 1978; Prentice and Pyke, 1979; Greenland, 1981; Benichou and Wacholder, 1994; Morise et al., 1996; Wacholder,

1

1996; Greenland, 2004). Thus, this method, intercept adjusted maximum likelihood estimation, can be used to ascertain the predicted probability of disease Y given covariates W with case-control study data.

3 Case-Control Weighting for Prediction

Sampling Design. Let us define $O^* = (W, Y) \sim P_0^*$ as the experimental unit and corresponding distribution P_0^* of interest. The experimental unit O^* consists of covariates W and a binary outcome Y that defines case or control status. P_0^* represents the population from which all cases and controls will be sampled. We define independent case-control sampling as sampling nC cases from the conditional distribution of W, given $Y = 1$, and sampling nCo controls from W, given $Y = 0$. However, extensions of our method to other types of case-control study designs, such as matched case-control sampling and incidence-density sampling, will be addressed in the discussion.

Weighting. Weights q_0 and $(1 - q_0)\frac{1}{J}$ are assigned to cases and controls, respectively. The value of J used to weight each control is nCo/nC, the average number of controls per case.

Case-Control Weighted Maximum Likelihood Estimation. We want to estimate the conditional probability of Y given W, $P_0^*(Y \mid W) \equiv Q_0^*(W)$, for each experimental unit in the case-control study. This estimate of $Q_0^*(W)$ is denoted $\hat{Q}^*(W)$. Maximum likelihood estimation for prospective sampling can then be performed, using the assigned weights, for prediction with case-control study data. Consider a nonparametric model for the marginal distribution of W and a model $\{Q_\theta^* : \theta\}$ for $Q_0^*(W)$. The case-control weighted maximum likelihood estimator for $Q_0^*(W)$ is then:

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^n q_0 \log \hat{Q}_\theta^*(W_{1i}) + (1 - q_0)\frac{1}{J} \sum_{j=1}^J \log(1 - \hat{Q}_\theta^*(W_{2i}^j)).$$

The subscripts 1 and 2 indicate values for cases and controls, respectively. Case-control weighted maximum likelihood estimation can be implemented as a weighted logistic regression. One can use existing software, including SAS, STATA, and R, to perform weighted logistic regression.

2

Model Selection. In many applied settings, the form of the prediction model may be unknown. The case-control weighting scheme described in this paper can also be used in any data-adaptive model selection procedure, as long as the procedure allows for the experimental units to be weighted. Similarly, intercept adjusted maximum likelihood estimation can be performed with data-adaptive model selection procedures. However, in this case, the predicted values returned by the estimation procedure are then updated in a separate step. The Deletion/Substitution/Addition (DSA) algorithm is a data-adaptive model selection procedure based on cross-validation and uses polynomial basis functions to search through a parameter space of potential regression functions (Sinisi and van der Laan, 2004). It has an option for the experimental units to be weighted. We will use DSA in our simulations to demonstrate the use of case-control weighting for prediction in case-control studies.

4 Simulation Studies

4.1 Simulation Study 1

Population. Our first simulation study was designed to illustrate the use of the case-control weighting scheme for prediction in case-control designs in a simple population. It was based on a population of $N = 55,000$ individuals, and we simulated 2 covariates $W = \{W_1, W_2\}$ and an indicator Y, which was 1 for cases and 0 for controls. These variables were generated according to the following rules:

$$W_1 \sim U(0,1)$$

$$W_2 = \frac{1}{1+\exp(-(2W_1-1))}$$

$$P_0^*(Y=1|W) = \frac{1}{1+\exp(-(-\sin(W_1^2)+9\log(W_1)+1.2W_2-1))}.$$

The resulting population had a prevalence probability $q_0 = 0.037$. We sampled the population at several sample sizes, each with equal numbers of cases and controls, and for each sample size we ran 1000 simulations.

**Prediction Methods.** We estimated $P_0^*(Y=1|W)$ using:

3

1. Case-Control Weighted DSA (CCW DSA): Implementing case-control weighted logistic regression, discussed in Section 3, using the data-adaptive model selection procedure DSA (Sinisi and van der Laan, 2004).

2. Intercept Adjusted DSA (IA DSA): Implementing an intercept adjusted logistic regression (Section 2) using DSA.

For the DSA algorithm, we did not force any variables into the model and thus DSA selected all terms. The maximum size of the model was set to 6 terms, the model was limited to two-way interactions, and the maximum sum of powers for any term was set to 3. The default setting for cross-validation was maintained, thus five-fold cross validation was performed to prevent overfitting. The case-control weighted DSA included weights q_0 and $(1 - q_0)\frac{1}{J}$ for cases and controls, respectively. Intercept adjusted DSA was performed without weighting and the predicted values generated for each model were updated with $\log \frac{q_0}{1-q_0}$.

Results. We illustrated each of the predictors using the predicted probabilities of being a case. The means of the true values for each sample size, among cases only, controls only, and in the total sample, are listed in Table 1. Results for average bias and mean squared error can be seen in Figure 1 and Table 2, respectively. The case-control weighted DSA and intercept adjusted DSA performed similarly with respect to bias and mean squared error. There were slight differences in bias, but the magnitudes of these differences were small, such that they may not be true differences. Similarly, the relative efficiency of case-control weighted DSA compared to intercept adjusted DSA hovered around 1.00.

Table 1: **Simulation 1 True Mean Probabilities.** N is total sample size. Equal numbers of cases and controls were used at each sample size.
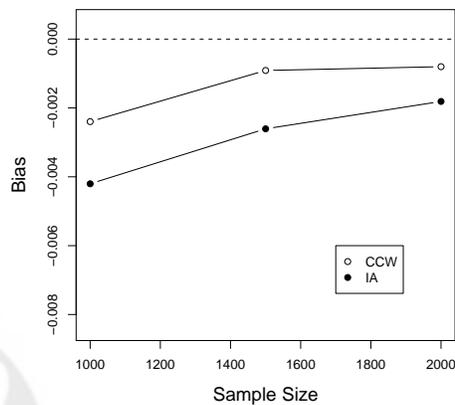
N	Cases Only	Controls Only	Total Sample
1000	0.1829	0.0313	0.1071
1500	0.1828	0.0315	0.1071
2000	0.1829	0.0314	0.1071

4
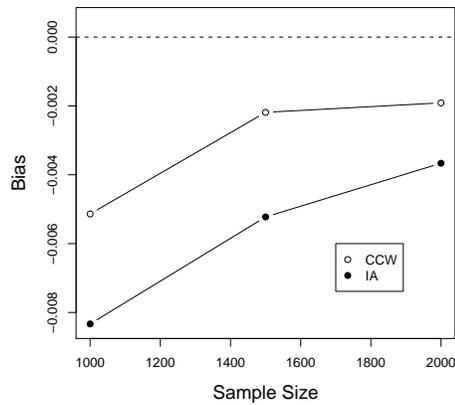
(a) Cases Only
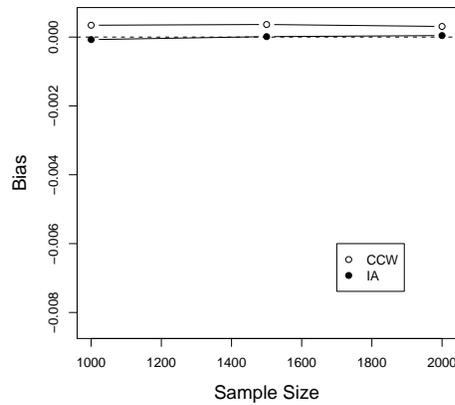
(b) Controls Only

(c) Total Sample

Figure 1: **Simulation 1 Bias Results.** CCW is case-control weighted DSA procedure and IA is intercept adjusted DSA procedure.

5

Table 2: **Simulation 1 MSE.** N is total sample size, MSE is Mean Squared Error and RE is the Relative Efficiency of CCW DSA compared to IA DSA. Equal numbers of cases and controls were used at each sample size.

N	Method	Cases Only	Controls Only	Total Sample
1000	IA MSE	1.39E-03	1.31E-04	7.63E-04
	CCW RE	0.96	0.97	0.96
1500	IA MSE	8.65E-04	8.84E-05	4.77E-04
	CCW RE	1.08	1.04	1.07
2000	IA MSE	6.69E-04	6.81E-05	3.68E-04
	CCW RE	1.06	1.03	1.06

4.2   Simulation Study 2

Population.   Our second simulation study illustrated the use of the case-control weighting scheme for prediction in case-control designs involving many covariates, such as a list of candidate SNPs. It was also designed to demonstrate, in a machine learning setting, that case-control weighting often outperforms intercept adjustment in practical settings. The simulation was based on a population of $N = 19,500$ individuals, and we simulated a 25-dimensional covariate W and indicator Y, which was 1 for cases and 0 for controls. The 25-dimensional covariate W was comprised of dichotomous values, generated according to $W_i \sim \text{Binomial}(p_i)$. The values of p_i ranged from 0.02 to 0.90. The remaining variable Y was generated using the first 10 covariates, with 8 interaction terms and 1 main term. The resulting population had a prevalence probability $q_0 = 0.054$. We sampled 1000 cases and 1000 controls from the population and ran 100 simulations.

Prediction Methods.   The same methods described in the previous simulation were used here: case-control weighted DSA and intercept adjusted DSA. However, for the DSA algorithm, the maximum size of the model was set to 10 and the model was limited to main terms. All other settings were as described in Section 4.1. Due to the complexity of the underlying population model, the limiting of the model to a maximum of 10 main terms led to misspecified models.

Results.   Again, we illustrated each of the predictors using the predicted probabilities of being a case. The means of the true probabilities were 0.2267,

6

Table 3: **Simulation 2 Results.** CCW is Case-Control Weighted DSA, IA is Intercept Adjusted DSA, MSE is Mean Squared Error, and RE is the Relative Efficiency of CCW DSA compared to IA DSA. Sample size was 2000 with 1000 cases and 1000 controls, with 1000 samples taken.

MSE	Cases Only	Controls Only	Total Sample
IA MSE	1.60E-02	3.35E-03	9.68E-03
CCW RE	1.50	1.42	1.49
Bias			
IA MSE	5.46E-02	-3.73E-03	2.54E-02
CCW RE	4.98E-02	-2.52E-03	2.37E-02

0.0443, and 0.1355 among cases only, controls only, and in the total sample. Mean squared error and bias results are displayed in Table 3. The case-control weighted DSA estimator produced less biased predicted probabilities among cases only, controls only, and in the total sample. The magnitudes of the differences were larger than those seen in Simulation 1, although still small. The more important distinction here is that the case-control DSA procedure was substantially more efficient. These results were not unexpected based on our findings published in Rose and van der Laan (2008a), where we discussed the sensitivity of intercept adjusted maximum likelihood estimation to model misspecification.

5   Discussion

This paper was designed to introduce the use of case-control weighted models for prediction with case-control study data. This extension follows from the case-control methodology developed for causal inference described by van der Laan (2008). Our simulations demonstrated the use of case-control weighted maximum likelihood estimation in a data-adaptive model selection procedure. Case-control weighting performed similarly to a previously known method for prediction in case-control study designs, intercept adjustment, in our simulations with few covariates and allowances for interactions and higher powered terms. When the simulation included a larger number of covariates and was limited to main terms, case-control weighting outperformed intercept adjustment. This result coincided with our conclusions from Rose and van der Laan (2008a). There, we demonstrated that intercept adjusted maximum

7

likelihood estimation was very sensitive to model misspecification, whereas case-control weighted maximum likelihood estimation was not. Therefore, the use of case-control weighting may outperform intercept adjustment in many practical settings, including situations where a priori specified models are used *and* when data-adaptive model selection procedures are used.

We hope that case-control weighting will become a useful device for creating predictive models for rare diseases. It is an opportunity for us, as part of the research community, to translate our work into a tool of immediate use for clinicians and other researchers. We presented case-control weighting for prediction with independent case-control study data. However, the case-control weighting scheme easily extends to other study designs, such as matched case-control sampling and incidence-density sampling. For matched case-control study data, the weights are q_0 for cases and $\bar{q}_0(M) \equiv q_0 \frac{P_0^*(Y=0|M)}{P_0^*(Y=1|M)}$ for controls, where M is the matching variable. In incidence-density sampling, q_0 is defined as the incidence probability, and the case-control weights depend on the time points the cases and controls were sampled. See van der Laan (2008), Rose and van der Laan (2008a), and Rose and van der Laan (2008b) for details on these additional weighting schemes. For further reading on advances in prediction for case-control study design data, we also refer readers to Huang and Pepe (2008) who discuss a 'predictiveness curve' (Pepe et al., 2007) that incorporates risk prediction with classification performance measures for case-control studies. Their methodology also assumes knowledge of $P_0^*(Y = 1) \equiv q_0$.

References

J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59: 19–35, 1972.

J. Benichou and S. Wacholder. A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. *Statistics in Medicine*, 13:651–661, 1994.

S. Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160(4):301–305, 2004.

8

S. Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chron Dis*, 34:445–453, 1981.

Y. Huang and M. Pepe. Semiparametric and nonparametric methods for evaluating risk prediction markers in case-control studies. *Technical Report 333, Department of Biostatistics, University of Washington*, 2008.

R. Jackson. Updated new zealand cardiovascular disease risk-benefit prediction guide. *Br Med J*, 320(7236):709–710, 2000.

A.P. Morise, G.A. Diamon, R. Detrano, M. Bobbio, and Erdogan Gunel. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Med Decis Making*, 16:133–142, 1996.

M.S. Pepe, Z. Feng, Y. Huang, G. Longton, R. Prentice, I.M. Thompson, and Y. Zheng. Integrating the predictiveness of a marker with its perfomance as a classifier. *Am J Epidemiol*, 167(3):362–368, 2007.

R.L. Prentice and N.E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

L.E. Ramsay, I.U. Haq, P.R. Jackson, and W.W. Yeo. Sheffield risk and treatment table for cholesterol lowering for primary prevention of coronary heart disease. *Lancet*, 346(8988):1467–1471, 1995.

L.E. Ramsay, I.U. Haq, P.R. Jackson, and W.W. Yeo. The Sheffield table for primary prevention of coronary heart disease: corrected. *Lancet*, 348 (9036):1251, 1996.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, 4(1):Article 18, 2008a.

S. Rose and M.J. van der Laan. Why match? Investigating matched case-control study designs with causal effect estimation. *Technical Report 240, Division of Biostatistics, University of California, Berkeley*, 2008b.

9

S. Sinisi and M.J. van der Laan. Deletion/substitution/addition algorithm in loss function based estimation. *Journal of Statistical Methods in Molecular Biology*, 3(1):Article 18, 2004.

M.J. van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, 4(1): Article 17, 2008.

S. Wacholder. The case-control study as data missing by design: Estimating risk differences. *Epidemiology*, 7(2):144–150, 1996.

D.C. Whiteman and A.C. Green. A risk prediction tool for melanoma? *Cancer Epidemiol Biomarkers Prev*, 14(4):761–763, 2005.