Lifetime Data Anal (2013) 19:170–201
DOI 10.1007/s10985-012-9237-1

Estimating improvement in prediction with matched case–control designs

Aasthaa Bansal · Margaret Sullivan Pepe

Received: 23 March 2012 / Accepted: 26 November 2012 / Published online: 29 January 2013
© Springer Science+Business Media New York 2013

Abstract When an existing risk prediction model is not sufficiently predictive, additional variables are sought for inclusion in the model. This paper addresses study designs to evaluate the improvement in prediction performance that is gained by adding a new predictor to a risk prediction model. We consider studies that measure the new predictor in a case–control subset of the study cohort, a practice that is common in biomarker research. We ask if matching controls to cases in regards to baseline predictors improves efficiency. A variety of measures of prediction performance are studied. We find through simulation studies that matching improves the efficiency with which most measures are estimated, but can reduce efficiency for some. Efficiency gains are less when more controls per case are included in the study. A method that models the distribution of the new predictor in controls appears to improve estimation efficiency considerably.

Keywords Classification · Diagnosis · Medical decision making · Receiver operating characteristic curve

1 Introduction

Medical decisions are often based on an individual's calculated risk of having or developing a condition. For example, decisions to prescribe long-term cholesterol lowering

A. Bansal (✉) · M. S. Pepe
Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
e-mail: abansal@uw.edu

M. S. Pepe
Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500,
Seattle, WA 98109, USA
e-mail: mspepe@u.washington.edu

statin therapy are often made with use of the Framingham risk of a cardiovascular event (Truett et al. 1967; Kannel et al. 1976; Gordon and Kannel 1982; Anderson et al. 1991) that uses as input information the individual's sex, age, blood pressure, total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, smoking behavior and diabetes status. The Breast Cancer Risk Assessment Tool (BCRAT) is used to calculate 10 year risk of breast cancer for individuals, using information on age, personal medical history (number of previous breast biopsies and the presence of atypical hyperplasia in any previous breast biopsy specimen), reproductive history (age at the start of menstruation and age at the first live birth of a child) and family history of breast cancer. If a woman's risk exceeds an age-specific threshold, she may be recommended for hormone therapy that reduces the risk at least in some women. Risk prediction models can also be used to determine if a person's risk is low enough to forgo certain unpleasant or costly medical interventions (Gail et al. 1989, 1999).

Our ability to predict risk with currently available clinical predictors is often very poor. For example the BCRAT model has a very modest capacity to discriminate women who develop breast cancer within 10 years from those who do not. The area under the age-specific receiver operating characteristic curve is approximately 0.56 (Mealiffe et al. 2010). Therefore new predictors are sought for their capacity to improve upon its prediction performance. Recent advances in and wider availability of molecular and imaging biotechnologies offer the potential for new powerful predictors. Recent studies have examined the use of data on genetic polymorphisms and breast density to improve the performance of BCRAT.

This paper concerns study designs to estimate the improvement in prediction performance that is gained by adding a new predictor Y to a set of baseline predictors X, to predict the risk of an outcome D ($D = 1$ for a bad outcome and $D = 0$ for a good outcome). When resources are limited and Y is difficult to ascertain, it may not be feasible to measure it on all subjects in a study cohort. Consider, for example, if the new predictor is a biomarker measured on biological samples obtained and stored while women were healthy at enrollment in the Women's Health Initiative. The preciousness of such biological samples dictates that they be used with maximum efficiency. Typically therefore a case–control study design is employed wherein Y is measured on a random subset of cases (denoted by $D = 1$) and a selected subset of controls ($D = 0$).

Our specific interest concerns whether or not the controls on whom Y is measured should be selected to frequency match the cases with regard to the baseline predictors X. Matching is in fact routinely done in practice in order to avoid observing associations between Y and D that are solely due to associations of X with both Y and D. However, the effect of this practice on estimation of performance improvement is not fully understood. We have raised concerns about matching with regards to bias, emphasizing that naïve analyses typically employed are misleading, as they underestimate performance (Pepe et al. 2012). The effect of matching on the estimation of incremental value with regards to efficiency has not been examined. Nevertheless, the practice is entrenched in the field of biomarker research. Here, we propose a two-stage estimator that accounts for matching to produce unbiased estimates. Using this estimator, we look to address the question of whether matching can improve the efficiency of estimating the increment in performance. This is an important question given that matching also necessitates a somewhat more complicated analysis algorithm than is required for an unmatched study. We ask whether there is a large enough (or any) efficiency gain that justifies the common practice of matching and a more complicated analysis.

Matching is known to improve efficiency for estimating the odds ratio for Y in a risk model that includes X (Breslow and Day 1980). However, the odds ratio, $\frac{P(D=1|X,Y=y+1)/P(D=0|X,Y=y+1)}{P(D=1|X,Y=y)/P(D=0|X,Y=y)}$, does not characterize prediction performance or improvement in prediction performance gained by including Y in the risk model over and above use of X alone. The distribution of (X, Y) in the population is an additional component that enters into the calculation of prediction performance. Janes and Pepe (2009) showed that matching on X is also optimal for estimating the covariate adjusted ROC curve, which is a measure of prediction performance. However, Janes and Pepe (2008) show that the covariate adjusted ROC curve that characterizes the ROC performance of Y within populations where X is fixed, does not quantify the improvement in the ROC curve gained by including Y in the risk model. It is currently unknown if matching leads to gains in efficiency for estimating performance improvement.

There are many metrics available for gauging improvement in prediction performance, and there is much confusion in the field about which metrics are most worthy for reporting. In Sect. 2, we review the most popular measures, providing some novel insights about their interpretations and inter-relationships. We provide rationale for the measures we selected to study here. In Sect. 3, we describe how these measures can be estimated from matched and unmatched studies. Simulation studies that were performed to evaluate the properties of the estimators and the efficiencies of matched designs are described in Sect. 4 using a simulated dataset and a real dataset concerning the prediction of renal artery stenosis. In Sect. 5, we propose a bootstrap approach for inference and demonstrate its validity through simulation studies. In Sect. 6, we illustrate our methodology in the context of renal artery stenosis. We close with some recommendations and suggestions for further research.

2 Measures of improvement in prediction performance

We first consider the most popular measures used to quantify improvement in prediction performance. Table 1 presents definitions for these measures. In this section, we review the measures in more detail.

2.1 Notation

Recall our use of D for the outcome variable, $D = 1$ denoting a case with a bad outcome and $D = 0$ denoting a control with a good outcome. We use X for predictors in the baseline risk function, $\text{risk}(X) = P(D = 1|X)$, Y for the novel predictors to be added and we write $\text{risk}(X, Y) = P(D = 1|X, Y)$. All measures of prediction performance involve the distributions of $\text{risk}(X)$ and $\text{risk}(X, Y)$ in cases and controls. We write these distributions as:

$$F_X^D(r) = P(\text{risk}(X) \le r|D = 1)$$

$$F_X^{\bar{D}}(r) = P(\text{risk}(X) \le r|D = 0)$$

$$F_{X,Y}^D(r) = P(\text{risk}(X, Y) \le r|D = 1)$$

$$F_{X,Y}^{\bar{D}}(r) = P(\text{risk}(X, Y) \le r|D = 0)$$

The joint distributions of $(\text{risk}(X), \text{risk}(X, Y))$ in cases and controls will be denoted by $F^D(r, r')$ and $F^{\bar{D}}(r, r')$ respectively.

2.2 Proportions at high risk and net benefit

In some settings a threshold exists for high risk classification and patients designated as 'high risk' receive an intervention. For example, patients whose 10-year risk of a cardiovascular event exceeds 20 % are recommended for cholesterol lowering therapy (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults 2001). A risk model performs well, in the sense of treating people who would have an event in the absence of therapy, i.e. the cases, if a large proportion of those subjects are placed in the high risk category by the model, i.e. if $\text{HR}^D(r) \equiv P[\text{risk} > r|D = 1]$ is large. Conversely, one must consider to what extent subjects that would not have an event in the absence of intervention, i.e. the controls, are inappropriately given intervention. A good model will place few of the controls in the high risk category, i.e. $\text{HR}^{\bar{D}}(r) \equiv P[\text{risk} > r|D = 0]$ is small. The changes in $\text{HR}^D(r)$ and $\text{HR}^{\bar{D}}(r)$ that are gained by adding Y to the risk model are therefore key entities for quantifying improvement in model performance for decision making when a therapeutic threshold for risk exists:

$$\Delta\text{HR}^D(r) \equiv P[\text{risk}(X, Y) > r|D = 1] - P[\text{risk}(X) > r|D = 1]$$

$$\Delta\text{HR}^{\bar{D}}(r) \equiv P[\text{risk}(X) > r|D = 0] - P[\text{risk}(X, Y) > r|D = 0].$$

These measures are also called changes in the true and false positive rates. Note that our goal is to increase $\text{HR}^D(r)$ and reduce $\text{HR}^{\bar{D}}(r)$ by adding Y to the baseline risk model. Therefore positive values of ΔHR^D and $\Delta\text{HR}^{\bar{D}}$ are desirable.

There is a net expected benefit (B) associated with designating a case as high risk and a net expected cost (C) associated with designating a control as high risk. It has been noted that a rational choice of risk threshold is $r = C/(C + B)$ (Pauker and Kassierer 1980; Vickers and Elkin 2006) and that the expected population net benefit associated with use of a risk model and threshold r to assign treatment is $\text{NB}(r) = \{\rho\text{HR}^D(r) - (1 - \rho)\frac{r}{(1-r)}\text{HR}^{\bar{D}}(r)\}B$ where ρ is the population prevalence, $P(D = 1)$. Baker (2009) suggests standardizing $\text{NB}(r)$ by the maximum possible benefit, ρB, achieved when all cases and no controls are designated as high risk. This standardized measure $\mathrm{B}(r) \equiv \mathrm{HR}^D(r) - \frac{(1-\rho)}{\rho}\frac{r}{(1-r)}\,\mathrm{HR}^{\bar{D}}(r)$, the proportion of maximum benefit, can also be viewed as the true positive rate $\mathrm{HR}^D(r)$ discounted (appropriately) for the false positive rate $\mathrm{HR}^{\bar{D}}(r)$. The change in $\mathrm{B}(r)$ that is achieved by adding Y to the risk model is an appropriate summary of its components $\Delta\mathrm{HR}^D(r)$ and $\Delta\mathrm{HR}^{\bar{D}}(r)$:

$$\Delta\mathrm{B}(r) = \Delta\mathrm{HR}^D(r) + \frac{1-\rho}{\rho}\frac{r}{1-r}\Delta\mathrm{HR}^{\bar{D}}(r).$$

In some settings all subjects receive treatment by default and use of a prediction model is to identify low risk subjects that can forego treatment. Parameters analogous to $\Delta\mathrm{HR}^D(r)$, $\Delta\mathrm{HR}^{\bar{D}}(r)$ and $\Delta\mathrm{B}(r)$ can be defined but we do not focus on those here.

2.3 Performance measures related to fixed points on the ROC curve

When risk thresholds or costs and benefits are not available, other approaches to summarizing prediction performance have been proposed. Points on the ROC curve or on its inverse are commonly used in practice because of their use in evaluating diagnostic tests and classifiers. We define

$$\Delta\mathrm{ROC}\left(p^{\bar{D}}\right) = \mathrm{ROC}_{(X,Y)}\left(p^{\bar{D}}\right) - \mathrm{ROC}_X\left(p^{\bar{D}}\right)$$

where $\mathrm{ROC}(p^{\bar{D}})$ is the proportion of cases with risks above the threshold $r(p^{\bar{D}})$ that allows the fraction $p^{\bar{D}}$ of controls to be classified as high risk. Analogously,

$$\Delta\mathrm{ROC}^{-1}\left(p^D\right) = \mathrm{ROC}_X^{-1}\left(p^D\right) - \mathrm{ROC}_{(X,Y)}^{-1}\left(p^D\right)$$

where $\mathrm{ROC}^{-1}(p^D)$ is the proportion of controls with risks above the threshold $r(p^D)$ that is exceeded by the fraction p^D of cases.

Interestingly, the ROC points are closely related to measures proposed by Pfeiffer and Gail (2011) for quantifying prediction performance. They argue for choosing a high risk threshold $r(p^D)$ so that a specified proportion of cases (p^D) are designated as high risk and define the proportion needed to follow, $\mathrm{PNF}(p^D) = P[\text{risk} > r(p^D)]$, as a performance metric. In words, $\mathrm{PNF}(p^D)$ is the proportion of the population designated as high risk in order that p^D of the cases are classified as high risk. A little algebra shows that $\mathrm{PNF}(p^D) = \rho p^D + (1-\rho)\mathrm{ROC}^{-1}(p^D)$. The reduction in the proportion of the population needed to follow in order to identify p^D of the cases (ΔPNF) that is gained by adding Y to the model is

$$\Delta\mathrm{PNF}\left(p^D\right) = (1-\rho)\Delta\mathrm{ROC}^{-1}\left(p^D\right).$$

We choose to study $\Delta\text{ROC}^{-1}(p^D)$ here as it does not depend on the prevalence. Pfeiffer and Gail (2011) also define a performance metric that is the proportion of cases followed, $\text{PCF}(p)$, when a fixed proportion p of the population is designated as highest risk. This measure relates directly to the ROC:

$$\text{PCF}(p) = \text{ROC}\left(p^{\bar{D}}\right)$$

where $p^{\bar{D}}$ is the point on the x-axis of the ROC plot such that $p = \rho\text{ROC}(p^{\bar{D}}) + (1 - \rho)p^{\bar{D}}$. We study $\Delta\text{ROC}(p^{\bar{D}})$ rather than $\Delta\text{PCF}(p)$ here because of its widespread use and its independence from the prevalence.

2.4 Global performance measures that do not specify a risk threshold

The above measures require explicit or implicit choices for risk thresholds. Measures that average over all risk thresholds in some sense are popular in part because they avoid the need to choose a risk threshold. The change in the area under the ROC curve by adding Y to the model, denoted ΔAUC, is the most commonly used measure in practice. The AUC is often written as

$$\text{AUC} = P(\text{risk}_i > \text{risk}_j | D_i = 1, D_j = 0)$$

and

$$\Delta\text{AUC} = \text{AUC}_{(X,Y)} - \text{AUC}_X.$$

A more recently proposed measure, called the integrated discrimination improvement (IDI) index, is the change in the difference in mean risks between cases and controls:

$$\text{IDI} = \Delta\text{MRD} = \text{MRD}_{(X,Y)} - \text{MRD}_X$$

where

$$\text{MRD} = E(\text{risk}|D = 1) - E(\text{risk}|D = 0).$$

Both the AUC and the MRD are measures of distance between the case and control distributions of modeled risks. Another measure of distance between distributions is the above average risk difference:

$$\text{AARD} = P(\text{risk} > \rho | D = 1) - P(\text{risk} > \rho | D = 0),$$

the name deriving from the fact that $E(\text{risk}) = \rho$ regardless of the risk model. We study the AARD because it is related to several other measures of prediction performance. We note in particular that AARD $=$ B(ρ). Youden's index is a measure of diagnostic performance for binary tests and we write $\text{YI}(r) = \text{HR}^D(r) - \text{HR}^{\bar{D}}(r)$. We note that AARD $=$ YI(ρ). Moreover, theory from Gu and Pepe (2009a) implies that $\text{YI}(\rho) = \max(\text{ROC}(\rho) - \rho) = \max(\text{YI}(r))$. Therefore, $\text{AARD} = \max(\text{YI}(r))$. This is also known as the Kolmogorov–Smirnov measure of distance between the case and control risk distributions. Finally, Gu and Pepe (2009a) also showed that this statistic is equal to the standardized total gain statistic (Bura and Gastwirth 2001), a measure derived from the population distribution of risk. The measure of improvement in prediction performance that we consider is the difference in measures calculated with $\text{risk}(X, Y)$ compared with when calculated with $\text{risk}(X)$:

$$\Delta\text{AARD} = \text{AARD}_{(X,Y)} - \text{AARD}_X.$$

2.5 Risk reclassification performance measures

Reclassification measures of performance compare $\text{risk}(X, Y)$ with $\text{risk}(X)$ within individuals and summarize across subjects. The most popular measure is the net reclassification improvement (NRI) index (Pencina et al. 2008). We focus on the continuous NRI (Pencina et al. 2011), written NRI(>0):

$$\begin{aligned}\text{NRI}(>0) \equiv\ & P(\text{risk}(X, Y) > \text{risk}(X)|D = 1) - P(\text{risk}(X, Y) < \text{risk}(X)|D = 1)\\ & +P(\text{risk}(X, Y) < \text{risk}(X)|D = 0) - P(\text{risk}(X, Y) > \text{risk}(X)|D = 0)\\ =\ & 2\{P(\text{risk}(X, Y) > \text{risk}(X)|D = 1) - P(\text{risk}(X, Y) > \text{risk}(X)|D = 0)\}\end{aligned}$$

It is interesting to consider the NRI(>0) statistic when the baseline model contains no covariates, i.e. when all subjects are assigned risk $= \rho$. In this setting it is related to measures mentioned previously:

$$\text{NRI} = 2\left\{\text{HR}^D(\rho) - \text{HR}^{\bar{D}}(\rho)\right\} = 2\text{AARD}(\rho) = 2\text{YI}(\rho) = 2\text{B}(\rho).$$

Originally the NRI was proposed for categories of risk and was defined as the net proportion of cases that moved to a higher risk category plus the net proportion of controls that moved to a lower risk category. When there are two categories, above or below the risk threshold r, the NRI$= \Delta\text{HR}^D(r) + \Delta\text{HR}^{\bar{D}}(r) = \Delta\text{YI}(r)$. Similar to $\Delta\text{B}(r)$, it is a weighted summary of improvements in true and false positive rates but unfortunately it uses inappropriate weights.

Another risk reclassification measure is the IDI, also defined as:

$$\text{IDI} = E\{\text{risk}(X, Y) - \text{risk}(X)|D = 1\} + E\{\text{risk}(X) - \text{risk}(X, Y)|D = 0\}.$$

Table 1 Definitions of performance measures

Name	Definition and notation	Performance improvement measure
High risk cases (r)	$\mathrm{HR}^D(r) = P(\text{risk} > r \mid D = 1)$	$\Delta\mathrm{HR}^D(r) = \mathrm{HR}^D_{(X,Y)}(r) - \mathrm{HR}^D_X(r)$
High risk controls (r)	$\mathrm{HR}^{\bar{D}}(r) = P(\text{risk} > r \mid D = 0)$	$\Delta\mathrm{HR}^{\bar{D}}(r) = \mathrm{HR}^{\bar{D}}_X(r) - \mathrm{HR}^{\bar{D}}_{(X,Y)}(r)$
Standardized benefit (r)	$\mathrm{B}(r) = \mathrm{HR}^D(r) - \frac{(1-\rho)}{\rho}\frac{r}{(1-r)}\mathrm{HR}^{\bar{D}}(r)$	$\Delta\mathrm{B}(r) = \mathrm{B}_{(X,Y)}(r) - \mathrm{B}_X(r)$
Cases above control defined threshold $(p^{\bar{D}})$	$\mathrm{ROC}(p^{\bar{D}}) = P(\text{risk} > r(p^{\bar{D}}) \mid D = 1)$	$\Delta\mathrm{ROC}(p^{\bar{D}}) = \mathrm{ROC}_{(X,Y)}(p^{\bar{D}}) - \mathrm{ROC}_X(p^{\bar{D}})$
Controls above case defined threshold (p^D)	$\mathrm{ROC}^{-1}(p^D) = P(\text{risk} > r(p^D) \mid D = 0)$	$\Delta\mathrm{ROC}^{-1}(p^D) = \mathrm{ROC}^{-1}_X(p^D) - \mathrm{ROC}^{-1}_{(X,Y)}(p^D)$
Area under the ROC curve	$\mathrm{AUC} = P(\text{risk}_i > \text{risk}_j \mid D_i = 1, D_j = 0)$	$\Delta\mathrm{AUC} = \mathrm{AUC}_{(X,Y)} - \mathrm{AUC}_X$
Mean risk difference	$\mathrm{MRD} = E(\text{risk} \mid D = 1) - E(\text{risk} \mid D = 0)$	$\Delta\mathrm{MRD} = \mathrm{MRD}_{(X,Y)} - \mathrm{MRD}_X = \mathrm{IDI}$
Above average risk difference	$\mathrm{AARD} = \{P(\text{risk} > \rho \mid D = 1) - P(\text{risk} > \rho \mid D = 0)\}$	$\Delta\mathrm{AARD} = \mathrm{AARD}_{(X,Y)} - \mathrm{AARD}_X$
Net reclassification improvement	NRI(>0)	$\mathrm{NRI} = 2\{P(\text{risk}(X,Y) > \text{risk}(X) \mid D = 1)$ $-P(\text{risk}(X,Y) > \text{risk}(X) \mid D = 0)\}$

The subscript X or (X, Y) denotes if the measure is calculated with the baseline or expanded risk models

Interestingly, because of the linearity, this measure of individual changes in risk due to adding Y to the model can also be interpreted as a difference of two population performance measures. That is, as noted earlier

$$\Delta\text{MRD} = \text{MRD}_{(X,Y)} - \text{MRD}_X = \text{IDI}.$$

3 Estimation from matched and unmatched designs

We now consider how the measures defined above can be estimated from a cohort study within which a case–control study of a new predictor is nested.

3.1 Data

We assume that data on the outcome and baseline covariates are available on a simple random sample of N independent identically distributed observations: (D_k, X_k), $k = 1\ldots, N$. We select a simple random sample of n_D cases from the cohort to ascertain Y: $Y_i, i = 1, \ldots, n_D$. The controls on whom Y is ascertained $\{Y_j,\ j = 1, \ldots, n_{\bar{D}}\}$ may be obtained as a simple random sample in an unmatched design. Alternatively, in a matched design, a categorical variable W is defined as a function of X, $W = W(X)$, and the number of controls within each level of W is chosen to equal a constant K times the number of cases with that value for W.

As shown in Table 1, all performance improvement measures are defined as functions of the risk distributions (notation in Sect. 2.1). We estimate $\text{risk}(X)$ and $\text{risk}(X, Y)$ first, then estimate their distributions in cases and controls and substitute the estimated distributions into expressions for the performance improvement measures.

3.2 Estimating risk functions

For the baseline model, we fit a regression model to the cohort data $\{(D_k, X_k), k = 1, \ldots, N\}$ and calculate predicted risks, $\widehat{\text{risk}}(X)$, for each individual in the cohort. For the expanded model, $\text{risk}(X, Y)$, we consider two approaches.

Case-control with adjustment We fit a model to data from the case–control subset, yielding fitted values $\widehat{\text{risk}}^{cc}(X, Y)$, and then adjust the intercept to the prevalence in the cohort

$$\text{logit}\ \widehat{\text{risk}}^{adj}(X, Y) = \text{logit}\ \widehat{\text{risk}}^{cc}(X, Y) - \text{logit}\left(\frac{n_D}{n}\right) + \text{logit}\left(\frac{N_D}{N}\right),$$

where $n = n_D + n_{\bar{D}}$ and N_D is the number of cases in the cohort. This is a well-known and standard approach to estimation of absolute risk for epidemiologic case–control studies (Breslow 1996). It draws upon the results of Prentice and Pyke (1979), which suggested that a prospective logistic model can be fit to retrospective data from a case–control study with a slight modification that adds an offset term to the logistic model.

The approach maximizes the pseudo- (or conditional-) likelihood that an observation in the case–control sample is a case or a control (Breslow and Cain 1988; Fears and Brown 1986).

However this approach does not account for matching. Pencina et al. (2011) presented a similar approach that used intercept adjustment to estimate NRI(>0) in the context of simple case-control studies.

Two-stage Two-stage methods acknowledge that selection of subjects for whom Y is measured, i.e. the second stage of sampling, may depend on their values of (D, X) found in the first stage. In particular, they account for matching. We generalize the intercept adjustment idea presented above to account for matching on X. This requires using the cohort to adjust the odds ratio associated with X. The odds ratio associated with Y is correctly estimated using standard logistic regression applied to the case–control dataset. We use the corresponding fitted values but adjust them using fitted values from the baseline model fit to the cohort and to the case–control datasets. Specifically, if we let $\widehat{\text{risk}}^{cohort}(X)$ and $\widehat{\text{risk}}^{cc}(X)$ denote the fitted values for the baseline models, then the two-stage estimator of the absolute risk is:

$$\text{logit } \widehat{\text{risk}}^{2-stage}(X, Y) = \text{logit } \widehat{\text{risk}}^{cc}(X, Y) - \text{logit } \widehat{\text{risk}}^{cc}(X) + \text{logit } \widehat{\text{risk}}^{cohort}(X)$$

Using '$cohort$' and 'cc' to denote sampling in the cohort or in the case–control subset, rationale for $\widehat{\text{risk}}^{2-stage}(X, Y)$ derives from the facts that

$$\text{logit } P(D = 1|X, Y, cohort) = \text{logit } P(D = 1|X, cohort) + \log \text{ DLR}_X(Y)$$

and

$$\text{logit } P(D = 1|X, Y, cc) = \text{logit } P(D = 1|X, cc) + \log \text{ DLR}_X(Y)$$

where the covariate-specific diagnostic likelihood ratio

$$\text{DLR}_X(Y) = P(Y|X, D = 1)/P(Y|X, D = 0)$$

is the same in the (matched or unmatched) case–control and cohort populations. The equations are a simple application of Bayes' theorem (Gu and Pepe 2009b). Substituting the expression for $\log \text{ DLR}_X(Y)$ derived from the case–control equation into that for the cohort equation gives the expression above for logit $\widehat{\text{risk}}^{2-stage}(X, Y)$.

3.3 Estimating distributions of risk

To estimate the risk distributions, we draw upon previously proposed methods for the estimation of risk distributions in simple case–control studies (Gu and Pepe 2009b; Huang et al. 2007; Huang and Pepe 2009). Here, we propose methodology for estimation with matched nested case–control data, which has not been previously considered. We estimate the baseline risk distributions, F_X^D and $F_X^{\bar{D}}$, using the empirical distributions of $\widehat{\text{risk}}(X)$ in the cohort data. Since the cases in the case–control set are drawn as a simple random sample from the cases in the cohort, we use the empirical distribution of $\widehat{\text{risk}}(X, Y)$ in the cases as the estimator of $F^{D}_{X,Y}$. For estimation of the distribution of $\widehat{\text{risk}}(X, Y)$ in the controls, we propose nonparametric and semiparametric approaches.

Nonparametric estimation In unmatched case–controls studies we can also use the empirical distribution of $\widehat{\text{risk}}(X, Y)$ among the controls to estimate $F^{\bar{D}}_{X,Y}$. However in matched designs the controls are not a simple random sample and the distribution of $\widehat{\text{risk}}(X, Y)$ must be reweighted to reflect the distribution in the population. Specifically, letting $c = 1, \ldots, C$ represent the distinct levels of the matching variable we can write

$$\begin{aligned} F^{\bar{D}}_{X,Y}(r) &= P\{\text{risk}(X, Y) \leq r|D = 0\} \\ &= \sum_{c=1}^{C} P\{\text{risk}(X, Y) \leq r|D = 0, W = c\}P(W = c|D = 0). \end{aligned} \tag{1}$$

A nonparametric estimator substitutes the observed proportions in the cohort for $P(W = c|D = 0)$ and the observed empirical stratum specific distributions of $\widehat{\text{risk}}(X, Y)$ for $P\{\text{risk}(X, Y)|D = 0, W = c\}$. We also consider a semiparametric estimator that substitutes semiparametric stratum specific estimates for $P\{\text{risk}(X, Y) \leq r|D = 0, W = c\}$.

Semiparametric estimation Observe that

$$P\{\text{risk}(X, Y) \leq r|D=0, W=c\}=E\{P(\text{risk}(X, Y) \leq r|D=0, X)|D=0, W=c\}. \tag{2}$$

A semiparametric location-scale model for the distribution of Y conditional on $(D = 0, X)$ is written

$$Y = \mu^{\bar{D}}(X) + \sigma^{\bar{D}}(X)\varepsilon$$

where the distribution of ε is unspecified, $\varepsilon \sim F_0$, and $\mu^{\bar{D}}(X)$, and $\sigma^{\bar{D}}(X)$ are parametric functions of X (Heagerty and Pepe 1999). After fitting the regression functions $\mu^{\bar{D}}(X)$ and $\sigma^{\bar{D}}(X)$, the empirical distribution of the residuals $\hat{\varepsilon_j} = (Y_j - \widehat{\mu}^{\bar{D}}(X_j))/\widehat{\sigma}^{\bar{D}}(X_j)$, $j = 1, ..., n_{\bar{D}}$, yields an estimator $\widehat{F}_0$. The semiparametric estimate of the distribution of Y is then

$$\begin{aligned} \widehat{P}(Y \leq y|D = 0, X) &= \widehat{P}\left\{\frac{Y - \widehat{\mu}^{\bar{D}}(X)}{\widehat{\sigma}^{\bar{D}}(X)} \leq \frac{y - \widehat{\mu}^{\bar{D}}(X)}{\widehat{\sigma}^{\bar{D}}(X)}\middle| D = 0, X\right\} \\ &= \widehat{P}\left\{\hat{\varepsilon} \leq \frac{y - \widehat{\mu}^{\bar{D}}(X)}{\widehat{\sigma}^{\bar{D}}(X)}\middle| D = 0, X\right\} \\ &= \widehat{F}_0\left\{\frac{y - \widehat{\mu}^{\bar{D}}(X)}{\widehat{\sigma}^{\bar{D}}(X)}\right\}, \end{aligned} \tag{3}$$

which in turn yields $\widehat{P}\{\text{risk}(X, Y) \le r|D = 0, X\}$. For example, if we use a logistic model for $\text{risk}(X, Y)$ and write logit $\widehat{\text{risk}}(X, Y) = \widehat{\theta}_0 + \widehat{\theta}_1 X + \widehat{\theta}_2 Y$ where $\widehat{\theta}_2 > 0$, then

$$\begin{aligned}\widehat{P}\{\text{risk}(X, Y) \le r|D = 0, X\} &= \widehat{P}\left\{\text{logit } \widehat{\text{risk}}(X, Y) \le \text{logit}(r)|D = 0, X\right\}\\ &= \widehat{P}\left\{\widehat{\theta}_0 + \widehat{\theta}_1 X + \widehat{\theta}_2 Y \le \text{logit}(r)|D = 0, X\right\}\\ &= \widehat{P}\left\{Y \le \frac{\text{logit}(r) - \widehat{\theta}_0 - \widehat{\theta}_1 X}{\widehat{\theta}_2}\middle| D = 0, X\right\}\\ &= \widehat{F}_0\left\{\frac{\frac{\text{logit}(r)-\widehat{\theta}_0-\widehat{\theta}_1 X}{\widehat{\theta}_2} - \widehat{\mu}^{\bar{D}}(X)}{\widehat{\sigma}^{\bar{D}}(X)}\right\},\end{aligned}$$

by substituting into (3). In turn, we estimate (2) as

$$\widehat{P}\{\text{risk}(X, Y) \le r|D=0, W=c\} = \frac{\sum_{j=1}^{N} \widehat{P}\{\text{risk}(X_j, Y) \le r|D_j=0, X_j\}\ I\{W(X_j)=c, D_j=0\}}{N_{\bar{D}}^{c}}$$

where $N_{\bar{D}}^{c}$ is the number of controls in the cohort with matching covariate value $W = c$. This estimator is then substituted into (1) to get $\widehat{F}_{X,Y}^{\bar{D}}(r)$. As noted above, a nonparametric estimator substitutes the observed proportions in the cohort for $P(W = c|D = 0)$, so that $\widehat{P}(W = c|D = 0) = \frac{N_{\bar{D}}^{c}}{N_{\bar{D}}}$. The semiparametric estimator then simplifies to

$$\widehat{F}_{X,Y}^{\bar{D}}(r) = \widehat{P}\{\text{risk}(X, Y) \le r|D=0\} = \frac{\sum_{j=1}^{N} \widehat{P}\{\text{risk}(X_j, Y) \le r|D_j=0, X_j\}\ I\{D_j=0\}}{N_{\bar{D}}}$$

for both matched and unmatched studies.

Both nonparametric and semiparametric estimators of $F_{X,Y}^{\bar{D}}$ are accompanied by a nonparametric estimator of $F_{X,Y}^{D}$.

3.4 Estimates of performance improvement measures

In Table 1, we presented the definitions of all performance improvement measures being studied here. Observe that estimates of $\Delta\text{HR}^{D}(r)$, $\Delta\text{HR}^{\bar{D}}(r)$, $\Delta\text{B}(r)$ and $\Delta\text{AARD}(r)$ follow directly from the estimators described above for the cumulative distributions of $\text{risk}(X)$ and $\text{risk}(X, Y)$ in cases and in controls. Note that since $\Delta\text{HR}^{D}(r)$ relies only on $F_{X,Y}^{D}$, what we refer to as nonparametric and semiparametric estimates of $\Delta\text{HR}^{D}(r)$ are in fact the same empirical estimate.

The pointwise ROC measures are also calculated directly, after noting that $\text{ROC}(p^{\bar{D}}) = 1 - F^{D}(r(p^{\bar{D}}))$ where $r(p^{\bar{D}})$ is such that $1 - F^{\bar{D}}(r(p^{\bar{D}})) = p^{\bar{D}}$ and $\text{ROC}^{-1}(p^{D}) = 1 - F^{\bar{D}}(r(p^{D}))$ where $r(p^{D})$ is such that $1 - F^{D}(r(p^{D})) = p^{D}$.

For ΔAUC, we use the usual empirical estimator with cohort data for the baseline value AUC$_X$, while we use

$$\widehat{\text{AUC}}_{(X,Y)} = \frac{1}{n_D}\sum_{i=1}^{n_D} \widehat{F}_{X,Y}^{\bar{D}} \left\{\widehat{\text{risk}}(X_i, Y_i)\right\},$$

where the summation is over cases, for the enhanced model. Note that this is equal to the usual empirical estimator in an unmatched study but that it also yields an estimate of $P\{\text{risk}(X_j, Y_j) \le \text{risk}(X_i, Y_i)|D_i = 1, D_j = 0\}$ in the matched design setting.

The baseline MRD is calculated empirically from the cohort values of $\widehat{\text{risk}}(X)$ while the enhanced model MRD is calculated as

$$\text{MRD}_{(X,Y)} = \frac{1}{n_D}\sum_{i=1}^{n_D}\widehat{\text{risk}}(X_i, Y_i) - \sum_{c=1}^{C}\widehat{E}\left\{\widehat{\text{risk}}(X, Y)|D = 0, W = c\right\} \\ P(W = c|D = 0).$$

Here $\widehat{E}\{\widehat{\text{risk}}(X, Y)|D = 0, W = c\}$ are the stratum specific sample averages of $\widehat{\text{risk}}(X, Y)$ for controls in the case–control study for the nonparametric estimator. For the semiparametric estimator $\widehat{E}\{\widehat{\text{risk}}(X, Y)|D = 0, W = c\}$ is calculated as the average of

$$\int \widehat{\text{risk}}(X_i, y)d\widehat{F}_0\left\{\frac{y - \mu^{\bar{D}}(X_i)}{\widehat{\sigma}^{\bar{D}}(X_i)}\right\} = \frac{1}{n_{\bar{D}}}\sum_{j=1}^{n_{\bar{D}}}\widehat{\text{risk}}\left\{X_i, \frac{Y_j - \widehat{\mu}^{\bar{D}}(X_j)}{\widehat{\sigma}^{\bar{D}}(X_j)}\widehat{\sigma}^{\bar{D}}(X_i) \\ +\widehat{\mu}^{\bar{D}}(X_i)\right\}$$

over the controls in the cohort stratum with $W = c$.

The NRI(>0) statistic uses the observed proportion of cases with $\widehat{\text{risk}}(X, Y) > \widehat{\text{risk}}(X)$ in the case–control study for the event NRI component, which requires estimation of $P\{\text{risk}(X, Y) > \text{risk}(X)|D = 1\}$. The non-event NRI component requires $P\{\text{risk}(X, Y) < \text{risk}(X)|D = 0\}$, which is estimated as a weighted average of the stratum specific observed proportions for the nonparametric estimator and as $\frac{1}{N_{\bar{D}}}\sum_{i=1}^{N_{\bar{D}}}\widehat{P}\{\widehat{\text{risk}}(X_i, Y) < \widehat{\text{risk}}(X_i)|D_i = 0, X_i\}$ for the semiparametric estimator.

Further details of the performance measure estimators obtained in each scenario are presented in Appendix Tables 8, 9 and 10.

3.5 Summary of estimation approaches

In Table 1, we showed that all performance improvement measures are functions of the risk distributions. Therefore, regardless of which measure is used, estimation of performance improvement is a two-fold task that requires estimating: (1) the risk functions risk(X) and risk(X, Y), and (2) the distributions of the risk functions in cases and in controls. We then substitute the estimated distributions into expressions for the performance improvement measures.

We estimated both risk functions parametrically using simple logistic models with linear terms. Other more flexible forms may be used in practice. In Sect. 3.2, we presented two different modeling approaches for estimating risk(X, Y) under the logistic regression framework. The first method (M_{adj}) is a commonly used approach which utilizes only the data in the case–control subset and is valid only for an unmatched design. The second method ($M_{2-stage}$) is a two-stage estimator which utilizes additional data from the cohort and is valid for both matched and unmatched designs. By comparing these two approaches to modeling the risk function, we aim to demonstrate that matching invalidates commonly used naïve analysis. Additionally, we investigate whether utilizing the parent cohort data for X improves the efficiency of risk function estimation.

In Sect. 3.3, we turned our attention to the estimation of the risk distributions in cases and in controls. We estimated the distributions of risk(X) using the empirical distributions estimated from the cohort. We also estimated the distribution of risk(X, Y) in cases empirically. For the estimation of the risk distribution in controls, we proposed nonparametric and semiparametric approaches for matched and unmatched case–control designs. The nonparametric approach has the advantage of making no modeling assumptions for the distribution of Y given X in controls. On the other hand, the semiparametric approach does make modeling assumptions and borrows information across strata of controls, and is therefore expected to be more efficient. One would therefore use the nonparametric approach in situations where there was uncertainty about how to model the distribution of Y given X in controls. The semiparametric approach would be preferable in situations with sparse controls. Using these two approaches for estimating the risk distribution, we aim to compare the efficiency of semiparametric estimation to that of nonparametric estimation.

Finally, using the above methods, we aim to answer the question of whether matching in the nested case–control subset improves efficiency in the estimation of performance improvement measures.

4 Simulation studies

We investigated the performances of the estimators and the merits of matched study designs using two small simulation studies—in the first study, we generated the data from a bivariate binormal model and in the second study, we used a real dataset.

4.1 Simulation study 1: bivariate binormal data

4.1.1 Data generation

We generated bivariate binormal cohort data of size $N = 5{,}000$ for cases ($D = 1$) and controls ($D = 0$) with population prevalence $\rho = P(D = 1) = 0.10$, so that the cohort contained $N_D = 500$ cases and $N_{\bar{D}} = 4{,}500$ controls:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} \mu_X(D) \\ \mu_Y(D) \end{pmatrix}, \begin{pmatrix} 1 & corr(X, Y|D) \\ corr(X, Y|D) & 1 \end{pmatrix}\right)$$

where $\mu_X(0) = \mu_Y(0) = 0$ and $\mu_X(1) = \mu_Y(1) = 0.742$. The corresponding AUC values associated with X and Y alone are $\text{AUC}_X = \text{AUC}_Y = \Phi(0.742/\sqrt{2}) = 0.7$. Data for $N = 5{,}000$ subjects were generated, so that $\{(D_i, X_i), i = 1, \dots N\}$ constitutes the study cohort data. A random sample of $n_D = 250$ cases were selected from the cohort and their Y values added to the dataset. For the unmatched design, Y values for a random sample of $n_{\bar{D}} = 500$ controls were also added to the dataset. For the matched design, we generated the matching variable W using quartiles of X in the control population and selected 2 controls randomly for each case in each of the four W strata.

4.1.2 Results

Using the notation M for a generic performance improvement measure, Table 2 shows mean values for estimates derived from 5,000 simulations. Estimates calculated using the adjusted case–control modeling approach for risk(X, Y) are denoted by M_{adj}, while estimates calculated using the two-stage modeling approach are denoted by $M_{2-stage}$. Bias estimates are calculated by subtracting the mean values from the true value for each measure. We see that the M_{adj} estimators are valid in unmatched designs, in the sense that mean values are close to the true values. However, M_{adj} estimators are biased in matched designs because they do not account for matching. Note that the direction and size of the bias is such that performance appears to decrease rather than increase with addition of Y to the model. In contrast the $M_{2-stage}$ estimators provide estimates that are centered around the true values in matched and unmatched designs.

The relative efficiencies of estimators are considered in Table 3 using ratios of standard deviations, with the standard deviation of the nonparametric M_{adj} estimator in the unmatched studies as the reference.

In the unmatched design, we found that the nonparametric $M_{2-stage}$ estimator is more efficient than M_{adj} for estimating $\Delta\text{HR}^D(0.20)$, ΔMRD and NRI(> 0). Interestingly, $M_{2-stage}$ performs slightly worse than M_{adj} for $\Delta\text{HR}^{\bar{D}}(0.20)$, but has similar performance to M_{adj} for all other performance measures.

To evaluate the impact of matching on efficiency we only consider $M_{2-stage}$ because M_{adj} estimators are biased. Comparing $M_{2-stage}$ in matched versus unmatched designs, we see that matching improves precision with which performance improvement is estimated for most measures. For example, with nonparametric estimation of the ROC related measures, the standard deviations in matched studies are 80–90 % the size of those in unmatched studies.

Interestingly, the improvement observed from matching can often be achieved in unmatched data by using the semiparametric estimator. In fact, for many of the measures, the efficiency is improved more by modeling $P(Y|X, D = 0)$ in an unmatched study than by matching controls to cases in the design and using the nonparametric estimator. For example, the standard deviation of the nonparametric estimate of $\Delta\text{HR}^{\bar{D}}(0.20)$ in matched studies is 74.0 % of the reference, while the semiparametric

Table 2 Mean estimates of improvement in prediction performance for measures defined in Table 1. Results are from 5,000 simulations of nested case–control studies ($n_D = 250$, $n_{\bar{D}} = 500$) with a cohort of 5,000 subjects. Data were generated from the bivariate binormal model described in the text with $corr(X, Y|D) = 0.5$. Estimates calculated with $\widehat{\text{risk}}^{adj}(X, Y)$ are denoted by M_{adj} and those calculated with $\widehat{\text{risk}}^{2-stage}(X, Y)$ are denoted by $M_{2-stage}$. (a) Nonparametric and (b) semiparametric estimates are presented

Measure	True value	Unmatched design				Matched design			
		M_{adj}		$M_{2-stage}$		M_{adj}		$M_{2-stage}$	
		Estimate	Bias	Estimate	Bias	Estimate	Bias	Estimate	Bias
(a) Nonparametric estimates									
$\Delta\text{HR}^D(0.20)$	0.067	0.069	0.002	0.069	0.002	−0.169	−0.236	0.069	0.002
$\Delta\text{HR}^{\bar{D}}(0.20)$	−0.013	−0.013	0.000	−0.013	0.000	0.056	0.069	−0.013	0.000
$\Delta\text{B}(0.20)$	0.038	0.040	0.002	0.039	0.001	−0.043	−0.081	0.039	0.001
$\Delta\text{ROC}^{-1}(0.80)$	0.046	0.048	0.002	0.048	0.002	−0.030	−0.076	0.046	0.000
$\Delta\text{ROC}(0.10)$	0.041	0.046	0.005	0.045	0.004	−0.029	−0.070	0.041	0.000
ΔAUC	0.028	0.028	0.000	0.028	0.000	−0.020	−0.048	0.028	0.000
$\Delta\text{MRD} = \text{IDI}$	0.020	0.021	0.001	0.020	0.000	−0.027	−0.047	0.020	0.000
ΔAARD	0.042	0.043	0.001	0.043	0.001	−0.030	−0.072	0.043	0.001
NRI(>0)	0.337	0.339	0.002	0.337	0.000	−0.270	−0.607	0.336	−0.001
(b) Semiparametric estimates									
$\Delta\text{HR}^D(0.20)$	0.067	0.069	0.002	0.069	0.002	−0.169	−0.236	0.069	0.002
$\Delta\text{HR}^{\bar{D}}(0.20)$	−0.013	−0.013	0.000	−0.013	0.000	0.056	0.069	−0.013	0.000
$\Delta\text{B}(0.20)$	0.038	0.039	0.001	0.040	0.002	−0.043	−0.081	0.040	0.002
$\Delta\text{ROC}^{-1}(0.80)$	0.046	0.046	0.000	0.046	0.000	−0.030	−0.076	0.046	0.000
$\Delta\text{ROC}(0.10)$	0.041	0.042	0.001	0.042	0.001	−0.030	−0.071	0.042	0.001
ΔAUC	0.028	0.028	0.000	0.028	0.000	−0.020	−0.048	0.028	0.000
$\Delta\text{MRD} = \text{IDI}$	0.020	0.021	0.001	0.021	0.001	−0.027	−0.047	0.020	0.000
ΔAARD	0.042	0.043	0.001	0.043	0.001	−0.030	−0.072	0.043	0.001
NRI(>0)	0.337	0.336	−0.001	0.337	0.000	−0.270	−0.607	0.338	0.001

Table 3 Efficiency of $M_{2-stage}$ in matched and unmatched designs relative to the nonparametric M_{adj} estimator from the unmatched design. Shown are the ratios of the standard deviations of estimates found in simulation studies divided by standard deviations (M_{adj}-NP; unmatched), so smaller values show more efficiency. NP and SP represent nonparametric and semiparametric estimation, respectively, of the distribution of risk(X, Y) in controls

Measure	Unmatched design		Matched design	
	$M_{2-stage}$-NP (%)	$M_{2-stage}$-SP (%)	$M_{2-stage}$-NP (%)	$M_{2-stage}$-SP (%)
$\Delta\text{HR}^{D}(0.20)$	75.3	75.3	74.3	74.3
$\Delta\text{HR}^{\bar{D}}(0.20)$	109.1	53.4	74.0	47.5
ΔB(0.20)	99.4	77.8	82.8	75.1
$\Delta\text{ROC}^{-1}(0.80)$	99.8	87.0	95.7	88.5
ΔROC(0.10)	98.9	77.7	83.1	75.3
ΔAUC	100.0	84.1	86.1	84.0
ΔMRD = IDI	71.1	69.3	65.3	64.7
ΔAARD	99.5	83.4	91.2	83.7
NRI(>0)	61.6	61.3	62.5	59.3

estimate in unmatched studies has a standard deviation that is 53.4 % of the reference. Some intuition for this result is provided by the fact that semiparametric estimation borrows information across strata of controls. While matching enriches strata with larger numbers of cases, it also makes those strata with fewer cases more sparse with respect to the number of controls. Therefore, both matched and unmatched data are prone to sparseness of controls in certain strata and nonparametric estimation suffers in such scenarios. The semiparametric approach, however, is less affected as it borrows information across strata.

4.2 Simulation study 2: renal artery stenosis data

4.2.1 Study description

The kidneys play several major regulatory roles in the human body, including regulation of blood pressure. The renal arteries aid in the proper functioning of the kidneys by supplying them with blood. Narrowing of the renal arteries is a condition termed *renal artery stenosis* (RAS); it inhibits blood flow to the kidneys and can lead to treatment-resistant hypertension.

The gold standard diagnostic test for RAS is an invasive and expensive procedure called renal angiography. In order to avoid unnecessarily performing angiography on individuals with a low likelihood of having disease, a clinical decision rule was developed to predict RAS based on patient characteristics and thus identify high-risk patients as candidates for the procedure (Krijnen et al. 1998).

We illustrate the proposed methodology using data from a RAS study (Janssens et al. 2005). For 426 patients, information is available on disease diagnosis from angiography, as well as age (10-year units), BMI, gender, recent onset of hypertension, presence of atherosclerotic vascular disease and serum creatinine (SCr) concentration. We model baseline risk using the first five characteristics and look to estimate the incremental value gained from adding SCr concentration to the model. Age and BMI were mean-centered. SCr concentration was log-transformed and standardized to have mean 0 and standard deviation 1. The study cohort includes 98 cases and 328 controls.

4.2.2 Methods

We simulated nested case–control studies using this dataset. Specifically, we resampled 426 observations with replacement from the cohort, selected all the cases and twice the number of controls, and disregarded SCr concentration data for patients who were not in the selected case–control subset. In one set of analyses the controls were selected unmatched as a simple random sample from all controls. In a second set of analyses the controls were selected to match the cases in regards to estimated baseline risk category. In particular, we created a three-level risk category variable, W, defined as: low if $\widehat{\text{risk}}(X) < 0.10$, medium if $0.10 < \widehat{\text{risk}}(X) < 0.20$ and high if $\widehat{\text{risk}}(X) > 0.20$. We selected two controls per case at random without replacement within each baseline risk category for the matched controls datasets. We also evaluated settings with 1:1 case–control ratios.

4.2.3 Results from renal artery stenosis dataset

Tables 4 and 5 summarize results of 1,000 nested case–control studies based on the renal artery stenosis dataset. We see that the M_{adj} estimators are only valid in unmatched case–control studies. Interestingly, the bias in M_{adj} in matched studies is such that prediction performance appears to disimprove considerably with addition of Y when the IDI, NRI(>0) or ΔHR^D performance measures are employed. This is very similar to results in Table 2 for the simulated bivariate normal distributions. Also as in Table 2, we see that $M_{2-stage}$ is valid in matched and unmatched designs.

Comparing the efficiency of $M_{2-stage}$ to M_{adj} in unmatched designs where both are valid, we see trends in the top panel of Table 5 that are similar to those observed in Table 3. For a case–control ratio of 1:1, $M_{2-stage}$-NP is more efficient than M_{adj}-NP, but only for ΔHR^D, ΔMRD and NRI(>0). For a larger number of controls (case–control ratio = 1:2), $M_{2-stage}$ loses some of its efficiency advantage. As before, $M_{2-stage}$ has worse performance than M_{adj} for the estimation of $\Delta\text{HR}^{\bar{D}}$, although again, this effect is lessened with the larger case–control ratio of 1:2.

Turning to the main question concerning efficiency due to matching, we again see some trends in the top panel of Table 5 that are similar to observations made for the bivariate binormal simulations in Table 3. Comparing $M_{2-stage}$-NP in matched versus unmatched designs, matching appears to improve the efficiency with which $\Delta\text{HR}^{\bar{D}}$ is estimated. However, ΔHR^D is not affected by matching and estimation of NRI(> 0) may be worse in matched studies. With larger numbers of controls, we see in the bottom panel of Table 5 that there is no gain from matching with regards to efficiency of $M_{2-stage}$-NP.

Table 4 Nonparametric estimates of improvement in prediction performance from the complete renal artery stenosis dataset and from simulated nested case–control datasets derived from it using a 1:2 case–control ratio. Shown are mean (a) nonparametric and (b) semiparametric estimates. Estimates calculated with $\widehat{\text{risk}}^{adj}(X, Y)$ are denoted by M_{adj} and those calculated with $\widehat{\text{risk}}^{2-stage}(X, Y)$ are denoted by $M_{2-stage}$. True values are obtained using the original renal artery stenosis dataset of all 426 subjects

Measure	True value	Unmatched design				Matched design			
		M_{adj}		$M_{2-stage}$		M_{adj}		$M_{2-stage}$	
		Estimate	Bias	Estimate	Bias	Estimate	Bias	Estimate	Bias
(a) Nonparametric estimates									
$\Delta\text{HR}^{D}(0.40)$	0.051	0.054	0.003	0.055	0.004	−0.170	−0.221	0.065	0.014
$\Delta\text{HR}^{\bar{D}}(0.40)$	−0.003	0.013	0.016	0.010	0.013	0.077	0.080	0.005	0.008
$\Delta\text{B}(0.40)$	0.045	0.084	0.039	0.079	0.034	0.002	−0.043	0.077	0.032
$\Delta\text{ROC}^{-1}(0.80)$	0.027	0.045	0.018	0.045	0.018	0.014	−0.013	0.050	0.023
$\Delta\text{ROC}(0.10)$	0.081	0.084	0.003	0.082	0.001	0.051	−0.030	0.081	0.000
ΔAUC	0.027	0.028	0.001	0.027	0.000	0.014	−0.013	0.028	0.001
$\Delta\text{MRD} = \text{IDI}$	0.069	0.068	−0.001	0.068	−0.001	−0.039	−0.108	0.075	0.006
ΔAARD	−0.032	0.034	0.066	0.032	0.064	−0.008	0.024	0.036	0.068
NRI(>0)	0.501	0.438	−0.063	0.467	−0.034	−0.290	−0.791	0.465	−0.036
(b) Semiparametric estimates									
$\Delta\text{HR}^{D}(0.40)$	0.051	0.054	0.003	0.055	0.004	−0.170	−0.221	0.065	0.014
$\Delta\text{HR}^{\bar{D}}(0.40)$	−0.003	0.020	0.023	0.021	0.024	0.074	0.077	0.015	0.018
$\Delta\text{B}(0.40)$	0.045	0.097	0.052	0.102	0.057	−0.005	−0.050	0.098	0.053
$\Delta\text{ROC}^{-1}(0.80)$	0.027	0.044	0.017	0.047	0.020	−0.014	−0.041	0.042	0.015
$\Delta\text{ROC}(0.10)$	0.081	0.094	0.013	0.099	0.018	0.047	−0.034	0.097	0.016
ΔAUC	0.027	0.027	0.000	0.028	0.001	0.005	−0.022	0.026	−0.001
$\Delta\text{MRD} = \text{IDI}$	0.069	0.066	−0.003	0.068	−0.001	−0.043	−0.112	0.074	0.005
ΔAARD	−0.032	0.038	0.070	0.040	0.072	−0.016	0.016	0.035	0.067
NRI(>0)	0.501	0.425	−0.076	0.467	−0.034	−0.251	−0.752	0.452	−0.049

Table 5 Efficiency of estimates of improvement in prediction performance in studies simulated from the renal artery stenosis dataset. Shown are standard deviations (SD) and the ratios of the standard deviations relative to that for nonparametric M_{adj} in the unmatched studies. NP and SP represent nonparametric and semiparametric estimation of the distribution of risk(X, Y) in controls, respectively

Measure	Unmatched design	Unmatched design				Matched design			
	M_{adj}-NP	$M_{2-stage}$-NP		$M_{2-stage}$-SP		$M_{2-stage}$-NP		$M_{2-stage}$-SP	
	SD	SD	Ratio (%)	SD	Ratio (%)	SD	Ratio (%)	SD	Ratio S(%)
Case–control ratio = 1:1									
$\Delta\mathrm{HR}^{D}(0.40)$	0.060	0.051	85.0	0.051	85.0	0.053	88.3	0.053	88.3
$\Delta\mathrm{HR}^{\bar{D}}(0.40)$	0.025	0.030	120.0	0.026	104.0	0.025	100.0	0.025	100.0
$\Delta\mathrm{B}(0.40)$	0.084	0.088	104.8	0.084	100.0	0.078	92.9	0.074	88.1
$\Delta\mathrm{ROC}^{-1}(0.80)$	0.069	0.066	95.7	0.056	81.2	0.065	94.2	0.058	84.1
$\Delta\mathrm{ROC}(0.10)$	0.078	0.072	92.3	0.072	92.3	0.071	91.0	0.063	80.8
$\Delta\mathrm{AUC}$	0.018	0.019	105.6	0.022	122.2	0.017	94.4	0.021	116.7
$\Delta\mathrm{MRD} = \mathrm{IDI}$	0.040	0.031	77.5	0.031	77.5	0.029	72.5	0.027	67.5
$\Delta\mathrm{AARD}$	0.058	0.059	101.7	0.047	81.0	0.057	98.3	0.048	82.8
NRI(>0)	0.237	0.178	75.1	0.170	71.7	0.203	85.7	0.166	70.0
Case–control ratio = 1:2									
$\Delta\mathrm{HR}^{D}(0.40)$	0.052	0.049	94.2	0.049	94.2	0.053	101.9	0.053	101.9
$\Delta\mathrm{HR}^{\bar{D}}(0.40)$	0.018	0.020	111.1	0.015	83.3	0.020	111.1	0.019	105.6
$\Delta\mathrm{B}(0.40)$	0.067	0.069	103.0	0.059	88.1	0.069	103.0	0.066	98.5
$\Delta\mathrm{ROC}^{-1}(0.80)$	0.059	0.058	98.3	0.055	93.2	0.061	103.4	0.057	96.6
$\Delta\mathrm{ROC}(0.10)$	0.064	0.063	98.4	0.057	89.1	0.064	100.0	0.060	93.8
$\Delta\mathrm{AUC}$	0.015	0.014	93.3	0.013	86.7	0.015	100.0	0.018	120.0
$\Delta\mathrm{MRD} = \mathrm{IDI}$	0.030	0.026	86.7	0.026	86.7	0.026	86.7	0.026	86.7
$\Delta\mathrm{AARD}$	0.049	0.049	100.0	0.044	89.8	0.052	106.1	0.046	93.9
NRI(>0)	0.197	0.160	81.2	0.156	79.2	0.179	90.9	0.150	76.1

Semiparametric estimation improves efficiency much more than matching does in these simulations. Again, this is consistent with the earlier simulation results.

5 Bootstrap method for inference

Performance improvement estimates obtained from nested case–control data incorporate variability from both the cohort and the nested case–control subset. However, simple bootstrap resampling from observed data cannot be implemented in this setting, as data on Y are observed only for subjects selected in the original case–control subset. Below we discuss our proposed strategy for bootstrapping with nested case–control data.

5.1 Proposed approach

We propose a parametric bootstrap method that combines resampling observations in the cohort and resampling residuals in the case–control subset (Efron and Tibshirani 1993). To begin, we have the original study cohort for which X and disease status are available and a nested case–control subsample on which Y is measured. We first bootstrap a cohort (say, cohort*) from the original cohort and proceed to generate the matching variable W^* based on quartiles of X^* in the bootstrapped cohort*. A matched or unmatched case–control subsample* is then constructed in the same fashion as before. However, note that in this bootstrapped case–control subsample*, the only subjects that have Y data are those who were selected to be in the original case–control subsample. We generate Y^* values for all subjects in the bootstrapped case–control subsample* using a parametric bootstrap method combined with residual resampling.

Specifically, we use the original case–control subsample to model $Y|X, D = 0$ semiparametrically as in Sect. 3.3,

$$Y^{\bar{D}} = \mu(X^{\bar{D}}) + \sigma\epsilon.$$

Fiting this model on the original case–control subsample gives us estimated values $\hat{\mu}$, $\hat{\sigma}$ and residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_{n_{\bar{D}}}$. Then, for each control* in the bootstrapped case–control subsample*, we use that subject's covariate values, X^*, and sample with replacement a residual from among $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_{n_{\bar{D}}}$ to generate a Y^* value using $\hat{\mu}$ and $\hat{\sigma}$:

$$Y_i^* = \hat{\mu}(X_i^*) + \hat{\sigma}\hat{\epsilon}_i^*, i = 1, ..., n_{\bar{D}}^*.$$

We fit a separate model for $Y|X, D = 1$ in the original case–control subsample and take a similar approach to generate $Y_1^*, \ldots, Y_{n_{D*}}^*$ for cases in the bootstrapped case–control subsample*.

5.2 Simulation study

We assessed the performance of the proposed bootstrap method with a simulation study using bivariate binormal data generated as in Sect. 4.1.1. We carried out 1,000

Table 6 Coverage of normality-based 95 % bootstrap confidence intervals

Measure	Nonparametric estimation				Semiparametric estimation			
	Unmatched design		Matched design		Unmatched design		Matched design	
	M_{adj} (%)	$M_{2-stage}$ (%)	M_{adj} (%)	$M_{2-stage}$ (%)	M_{adj} (%)	$M_{2-stage}$ (%)	M_{adj} (%)	$M_{2-stage}$ (%)
$\Delta\mathrm{HR}^{D}(0.20)$	95.1	95.2	1.3	95.2	95.1	95.2	1.3	95.2
$\Delta\mathrm{HR}^{\bar{D}}(0.20)$	95.3	94.8	1.1	96.6	94.7	94.5	95.1	95.3
$\Delta\mathrm{B}(0.20)$	95.9	94.5	27.8	95.0	96.1	95.6	0.6	96.0
$\Delta\mathrm{ROC}^{-1}(0.80)$	94.4	94.7	66.7	94.9	93.9	93.9	78.3	95.0
$\Delta\mathrm{ROC}(0.10)$	94.7	94.6	55.2	95.0	94.5	93.9	0.1	96.4
$\Delta\mathrm{AUC}$	94.5	94.5	33.2	94.9	95.2	95.1	30.2	95.4
$\Delta\mathrm{MRD} = \mathrm{IDI}$	95.4	94.1	0.9	95.6	95.7	94.5	0.9	95.4
$\Delta\mathrm{AARD}$	93.9	94.4	55.6	94.8	94.7	95.2	39.5	95.7
NRI(>0)	95.3	94.5	0.1	96.0	95.3	94.5	7.0	95.7

Results are from 1,000 simulations of nested case–control studies (n_D = 250, $n_{\bar{D}}$ = 500) with a cohort of 5,000 subjects. 200 bootstrap repetitions were carried out in each simulation. Data were generated from the bivariate binormal model described in the text with $corr(X, Y|D)$ = 0.5. Estimates calculated with $\widehat{\mathrm{risk}}^{adj}(X, Y)$ are denoted by M_{adj} and those calculated with $\widehat{\mathrm{risk}}^{2-stage}(X, Y)$ are denoted by $M_{2-stage}$. Nonparametric and semiparametric estimates are presented

simulations, each time generating a new study cohort of size $N = 5{,}000$ and from this study cohort, selecting a nested case–control subsample of size 250 cases and 500 controls. We used both the matched and unmatched designs. Within each simulation, we carried out 200 bootstrap repetitions using the procedure described above. For each performance measure estimate obtained in that simulation, we estimated its standard error as the standard deviation across the 200 bootstrap repetitions and used it to calculate normality-based 95 % confidence intervals. Coverage was averaged over all 1,000 simulations.

Results are presented in Table 6. Not surprisingly, M_{adj} estimators, which are biased in matched designs, also generate confidence intervals with poor coverage. For all other settings, coverage of the 95 % bootstrap confidence intervals is good.

6 Illustration with renal artery stenosis study

We illustrate our methodology on the renal artery stenosis dataset by simulating a single nested case–control dataset using the unmatched design and a single dataset using the matched design with a 1:2 case–control ratio. We include bootstrap standard errors and normality-based 95 % confidence intervals (CIs), obtained from 500 bootstrap repetitions following the approach described in Sect. 5. Instead of repeating numerous simulations as in Sect. 4.2, we have a single study cohort and a single two-phase dataset here that we bootstrap from.

Results are presented in Table 7. We see that the two-phase estimates are quite different from the full-data estimates. We used only a single two-phase sample here

Table 7 Results from a matched and an unmatched two-phase study simulated from the renal artery stenosis dataset. 95 % bootstrap confidence intervals were obtained from 500 bootstrap repetitions, using $M_{2-stage}$ and M_{adj} for estimation of risk(X, Y) and (a) nonparametric and (b) semiparametric estimation of the distribution of risk(X, Y) in controls

Measure	Full data estimate	$M_{2-stage}$			M_{adj}		
		Estimate	Std Err	95 % CI	Estimate	Std Err	95 % CI
(a) Nonparametric estimates							
Unmatched study design							
$\Delta\text{HR}^{D}(0.20)$	−0.010	−0.020	0.058	(−0.135, 0.094)	−0.020	0.057	(−0.131, 0.091)
$\Delta\text{HR}^{\bar{D}}(0.20)$	0.043	0.082	0.063	(−0.042, 0.206)	0.077	0.062	(−0.045, 0.199)
ΔB (0.20)	0.026	0.048	0.082	(−0.113, 0.210)	0.044	0.081	(−0.114, 0.203)
ΔROC^{-1} (0.80)	0.027	0.067	0.098	(−0.124, 0.258)	0.057	0.097	(−0.134, 0.248)
ΔROC (0.10)	0.081	0.071	0.089	(−0.103, 0.245)	0.081	0.089	(−0.094, 0.256)
ΔAUC	0.027	0.037	0.034	(−0.029, 0.103)	0.039	0.034	(−0.027, 0.105)
ΔMRD = IDI	0.069	0.081	0.026	(0.031, 0.132)	0.087	0.030	(0.028, 0.146)
ΔAARD	−0.032	0.006	0.048	(−0.089, 0.101)	0.037	0.047	(−0.055, 0.129)
NRI (>0)	0.501	0.531	0.155	(0.226, 0.835)	0.510	0.195	(0.129, 0.892)
Matched study design							
$\Delta\text{HR}^{D}(0.20)$	−0.010	−0.020	0.034	(−0.087, 0.046)	−0.112	0.049	(−0.209, −0.015)
$\Delta\text{HR}^{\bar{D}}(0.20)$	0.043	0.043	0.038	(−0.031, 0.117)	0.108	0.040	(0.030, 0.185)
ΔB (0.20)	0.026	0.016	0.048	(−0.078, 0.109)	−0.022	0.059	(−0.137, 0.093)
ΔROC^{-1} (0.80)	0.027	0.028	0.063	(−0.095, 0.150)	0.018	0.073	(−0.125, 0.161)
ΔROC (0.10)	0.081	0.051	0.067	(−0.081, 0.183)	0.020	0.079	(−0.134, 0.174)
ΔAUC	0.027	0.030	0.015	(0.001, 0.060)	0.021	0.019	(−0.016, 0.059)
ΔMRD = IDI	0.069	0.069	0.027	(0.015, 0.122)	−0.041	0.036	(−0.112, 0.029)
ΔAARD	−0.032	−0.024	0.052	(−0.126, 0.078)	−0.046	0.063	(−0.169, 0.077)
NRI (>0)	0.501	0.596	0.181	(0.241, 0.951)	−0.359	0.226	(−0.801, 0.084)

Table 7 continued

Measure	Full data estimate	$M_{2-stage}$			M_{adj}		
		Estimate	Std Err	95 % CI	Estimate	Std Err	95 % CI
(b) Semiparametric estimates							
Unmatched study design							
$\Delta\mathrm{HR}^{D}(0.20)$	−0.010	−0.020	0.058	(−0.135, 0.094)	−0.020	0.057	(−0.131, 0.091)
$\Delta\mathrm{HR}^{\bar{D}}(0.20)$	0.043	0.059	0.063	(−0.066, 0.183)	0.051	0.064	(−0.074, 0.176)
ΔB (0.20)	0.026	0.029	0.080	(−0.129, 0.186)	0.022	0.080	(−0.134, 0.178)
$\Delta\mathrm{ROC}^{-1}(0.80)$	0.027	0.039	0.101	(−0.159, 0.236)	0.015	0.101	(−0.183, 0.212)
ΔROC (0.10)	0.081	0.102	0.087	(−0.068, 0.272)	0.112	0.087	(−0.059, 0.283)
ΔAUC	0.027	0.034	0.034	(−0.032, 0.100)	0.033	0.034	(−0.033, 0.099)
ΔMRD = IDI	0.069	0.077	0.028	(0.023, 0.131)	0.080	0.029	(0.022, 0.138)
ΔAARD	−0.032	0.000	0.043	(−0.084, 0.084)	0.023	0.041	(−0.058, 0.103)
NRI (>0)	0.501	0.532	0.150	(0.237, 0.826)	0.463	0.187	(0.095, 0.830)
Matched study design							
$\Delta\mathrm{HR}^{D}(0.20)$	−0.010	−0.020	0.034	(−0.087, 0.046)	−0.112	0.049	(−0.209, −0.015)
$\Delta\mathrm{HR}^{\bar{D}}(0.20)$	0.043	0.035	0.027	(−0.017, 0.087)	0.080	0.034	(0.013, 0.148)
ΔB(0.20)	0.026	0.009	0.042	(−0.074, 0.091)	−0.045	0.055	(−0.152, 0.062)
$\Delta\mathrm{ROC}^{-1}(0.80)$	0.027	0.013	0.059	(−0.102, 0.128)	−0.046	0.069	(−0.181, 0.089)
ΔROC(0.10)	0.081	0.081	0.062	(−0.041, 0.203)	0.010	0.070	(−0.127, 0.147)
ΔAUC	0.027	0.027	0.015	(−0.002, 0.057)	0.009	0.019	(−0.027, 0.046)
ΔMRD = IDI	0.069	0.069	0.028	(0.013, 0.124)	−0.044	0.038	(−0.118, 0.030)
ΔAARD	−0.032	−0.014	0.046	(−0.104, 0.076)	−0.044	0.058	(−0.159, 0.071)
NRI(>0)	0.501	0.547	0.147	(0.259, 0.836)	−0.232	0.210	(−0.643, 0.179)

to mimic a real-life two-phase dataset. Repeating the sampling 100 times and averaging estimates across repetitions showed that the estimates are unbiased (data not shown). The observed inconsistency is a result of sampling variability. As before, we see that a standard adjusted analysis (M_{adj}) underestimates performance improvement in a matched design. $M_{2-stage}$ produces valid estimates. Conclusions regarding the incremental value of SCr concentration are similar using any of the valid estimation methods in this setting. We use estimates from $M_{2-stage}$ with semiparametric estimation and a matched design in the following discussion.

The incremental value of SCr concentration appears to be significant using ΔMRD and NRI as the measures of interest. Values of 0 for both measures would indicate no improvement from SCr concentration. $\widehat{\Delta \text{MRD}}$ is 0.069 {95 % CI (0.013,0.124)}, indicating that the change in the difference in mean risks between cases and controls is approximately 0.069. $\widehat{\text{NRI}}$ is 0.547 {95 % CI (0.259,0.836)}; given that NRI has a range of (−2,2), this seems like a moderate level of improvement in risk reclassification. Small improvements that are not statistically significant are seen using all other measures.

7 Discussion

Matching controls to cases on baseline risk factors is a common practice in epidemiologic studies of risk. It has also become common practice in biomarker research (Pepe et al. 2008). It allows one to evaluate from simple two-way analyses of Y and D if there is any association between Y and D and to be assured that the association is not explained by the matching factors. Matching also allows for efficient estimation of the relative risk associated with Y controlling for baseline predictors X in a risk model for $\text{risk}(X, Y)$. However, the impact of matching on estimates of prediction performance measures has not been explored previously.

We demonstrated the intuitive result that matching invalidates standard estimates of performance improvement. Our estimators that simply adjust for population prevalence but not for matching, M_{adj}, substantially underestimated the performance of the risk model $\text{risk}(X, Y)$ and therefore underestimated the increment in performance gained by adding Y to the set of baseline predictors X. Intuitively, this underestimation can be attributed to the fact that matching causes the distribution of X to be more similar to cases in study controls than in population controls and therefore the distribution of $\text{risk}(X, Y)$ is also more similar to cases in study controls than in population controls.

We derived two-stage estimators that are valid in matched or unmatched nested case–control studies. We were unable to derive analytic expressions for the variances of these estimates. Therefore we investigated efficiency in two simple simulation studies. Our results suggest that the impact of two-stage estimation and of matching varies with the performance measure in question. In our simulations two-stage estimation in unmatched studies had little impact on efficiencies of ROC measures but was advantageous for estimating the reclassification measures $\text{NRI}(> 0)$ and IDI = ΔMRD. On the other hand, matching improved efficiency of estimates of ROC related measures but did little to improve estimation of reclassification measures.

Our preferred measures of performance increment are neither ROC measures nor risk reclassification measures. We argue for use of the changes in high risk proportions of cases, $\Delta\text{HR}_D(r)$, high risk proportion of controls, $\Delta\text{HR}_{\bar{D}}(r)$, and the linear combination $\Delta\text{B}(r)$. These measures are favored due to their practical value for quantifying effects on improved medical decisions (Pepe and Janes 2013).

In our simulations we found that two-stage estimation improved efficiency of ΔHR_D but that matching had little to no further impact. Note that matching only affects the two-stage estimator for ΔHR_D through the influence of controls on the estimator of $\text{risk}(X, Y)$. That is, given estimates of $\text{risk}(X, Y)$, the empirical estimator of ΔHR_D is employed in both matched and unmatched designs as the cases are a simple random sample from the cohort. We conclude that the improvement in estimating $\text{risk}(X, Y)$ that is gained with matched data does not carry over to substantially impact on estimation of the distribution of $\text{risk}(X, Y)$ in cases. On the other hand, matching improved estimation of $\Delta\text{HR}_{\bar{D}}(r)$, at least with smaller control to case ratio.

We implemented a semiparametric method that modeled the distribution of Y given X among controls. This had a profound positive influence on efficiency with which most measures were estimated, especially in unmatched designs. If one is comfortable with making necessary assumptions to model Y given X in controls, it seems that little additional efficiency is gained by using a matched design.

We recognize that the simulation scenarios we studied are limited and our conclusions may not apply to other scenarios. There are a number of factors to consider with respect to study design and estimation and changing one of these factors could affect results. In fact, we see this happen in our two simulation studies. For example, in our second simulation study, changing the case–control ratio from 1:1 to 1:2 alone lessens the advantage of matching on results. Moreover, the effect of matching is different on different performance measures. More work is needed to derive analytic results that could generalize our observations. In the meantime our practical suggestion is to use simulation studies based on the application of interest in order to guide decisions about matching and other aspects of study design. Simulation studies may be based on hypothesized joint distributions for biomarkers, as in our first simulation study (Sect. 4.1). If pilot data are available one could base simulation studies on that, as we did with the renal artery stenosis data (Sect. 4.2). Simulation studies can be used to guide the design of another larger study, by simulating both matched and unmatched nested case–control studies by varying factors related to study design and estimation approach and investigating which approaches would maximize efficiency for the performance improvement measures of interest.

Another consideration in the decision to match is that inference is complicated by matching. Asymptotic distribution theory is not available for two-stage estimators of performance measures. The difficulty in deriving analytic expressions comes from the fact that there are multiple sources of variability that must be accounted for, given the complicated analytic approach and study design. Simple bootstrap resampling cannot be implemented in this setting because the nested case–control design implies that Y is only available for the study controls. We proposed a parametric bootstrap approach that generates Y for all cohort subjects using semiparametric models for Y given X fit to the original data. We showed that this method was valid with good coverage in simulation studies. We recommend this approach with the caveat that near the null, estimates tend to be skewed and in turn, inference tends to be problematic near the null for all measures of performance improvement. We and others have noted severe problems with bootstrap methods and inference in general for estimates of performance improvement even in cohort studies and especially with weakly predictive markers (Pepe et al. 2013; Kerr et al. 2011; Vickers et al. 2011). In practice, we recommend doing simulations similar to those suggested above to determine if valid inference is possible with the given data and study design or if the performance improvement is too close to the null. Continued effort is needed to develop methods for inference about performance improvement measures in cohort studies and then to extend them to nested case–control designs.

Acknowledgments The support for this research was provided by RO1-GM054438 and PO1-CA-053996. The authors thank Mr. Jing Fan for his contribution to the simulation studies.

Appendix

Table 8 Estimators of performance measures—nonparametric estimators using the baseline risk model and cohort data

Name	Estimator
$\mathrm{HR}_X^D(r)$	$\frac{1}{N_D}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r,\ D_i = 1\}$
$\mathrm{HR}_X^{\bar{D}}(r)$	$\frac{1}{N_{\bar{D}}}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r,\ D_i = 0\}$
$\mathrm{B}_X(r)$	$\frac{1}{N_D}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r,\ D_i = 1\} - \frac{N_{\bar{D}}}{N_D}\frac{r}{(1-r)}\frac{1}{N_{\bar{D}}}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r,\ D_i = 0\}$
$\mathrm{ROC}_X(p^{\bar{D}})$	$\frac{1}{N_D}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r(p^{\bar{D}}),\ D_i = 1\}$,
	where $r(p^{\bar{D}})$ s.t. $\frac{1}{N_{\bar{D}}}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r(p^{\bar{D}}),\ D_i = 0\} = p^{\bar{D}}$
$\mathrm{ROC}_X^{-1}(p^D)$	$\frac{1}{N_{\bar{D}}}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r(p^D),\ D_i = 0\}$,
	where $r(p^D)$ s.t. $\frac{1}{N_D}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > r(p^D),\ D_i = 1\} = p^D$
AUC_X	$\frac{1}{N_D}\Sigma_{i=1}^N \frac{1}{N_{\bar{D}}}\Sigma_{j=1}^N I\{\widehat{\mathrm{risk}}(X_j) \le \widehat{\mathrm{risk}}(X_i),\ D_i = 1,\ D_j = 0\}$
MRD_X	$\frac{1}{N_D}\Sigma_{i=1}^N \widehat{\mathrm{risk}}(X_i)\ I(D_i = 1) - \frac{1}{N_{\bar{D}}}\Sigma_{i=1}^N \widehat{\mathrm{risk}}(X_i)\ I(D_i = 0)$
AARD_X	$\frac{1}{N_D}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > \frac{N_D}{N},\ D_i = 1\} - \frac{1}{N_{\bar{D}}}\Sigma_{i=1}^N I\{\widehat{\mathrm{risk}}(X_i) > \frac{N_D}{N},\ D_i = 0\}$
NRI($>$0)	N/A

Table 9 Estimators of performance measures—nonparametric estimators using the enhanced risk model and case–control subset data

Name	Unmatched design	Matched design
$\mathrm{HR}^{D}_{(X,Y)}(r)$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 1\}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 1\}$
$\mathrm{HR}^{\bar{D}}_{(X,Y)}(r)$	$\frac{1}{n_{\bar{D}}}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 0\}$	$\Sigma_{c=1}^{C} \frac{N_{\bar{D},c}}{N_{\bar{D}}} \frac{1}{n_{\bar{D},c}} \Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 0,\ W_i = c\}$
$\mathrm{B}_{(X,Y)}(r)$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 1\}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 1\}$
	$-\frac{N_{\bar{D}}}{N_D} \frac{r}{1-r} \frac{1}{n_{\bar{D}}} \Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 0\}$	$-\frac{N_{\bar{D}}}{N_D} \frac{r}{1-r} \Sigma_{c=1}^{C} \frac{N_{\bar{D},c}}{N_{\bar{D}}} \frac{1}{n_{\bar{D},c}} \Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r,\ D_i = 0, W_i = c\}$
$\mathrm{ROC}_{(X,Y)}(p^{\bar{D}})$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r(p^{\bar{D}}),\ D_i = 1\}$, where $r(p^{\bar{D}})$ s.t. $\frac{1}{n_{\bar{D}}}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r(p^{\bar{D}}), D_i = 0\} = p^{\bar{D}}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r(p^{\bar{D}}),\ D_i = 1\}$, where $r(p^{\bar{D}})$ s.t. $\Sigma_{c=1}^{C} \frac{N_{\bar{D},c}}{N_{\bar{D}}} \frac{1}{n_{\bar{D},c}} \Sigma_{i=1}^{n} I\{\mathrm{risk}_i(X, Y) > r(p^{\bar{D}})$, $D_i = 0,\ W_i = c\} = p^{\bar{D}}$
$\mathrm{ROC}^{-1}_{(X,Y)}(p^{D})$	$\frac{1}{n_{\bar{D}}}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r(p^{D}),\ D_i = 0\}$, where $r(p^{D})$ s.t. $\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r(p^{D}),\ D_i = 1\} = p^{D}$	$\Sigma_{c=1}^{C} \frac{N_{\bar{D},c}}{N_{\bar{D}}} \frac{1}{n_{\bar{D},c}} \Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r(p^{D}),\ D_i = 0,\ W_i = c\}$, where $r(p^{D})$ s.t. $\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > r(p^{D}),\ D_i = 1\} = p^{D}$
$\mathrm{AUC}_{(X,Y)}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} \frac{1}{n_{\bar{D}}} \Sigma_{j=1}^{n} I\{\widehat{\mathrm{risk}}(X_j, Y_j) \le \widehat{\mathrm{risk}}(X_i, Y_i),\ D_i = 1,\ D_j = 0\}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} \Big[\Sigma_{c=1}^{C} \frac{N_{\bar{D},c}}{N_{\bar{D}}} \frac{1}{n_{\bar{D},c}} \Sigma_{j=1}^{n} I\{\widehat{\mathrm{risk}}(X_j, Y_j) \le \widehat{\mathrm{risk}}(X_i, Y_i)$, $D_i = 1,\ D_j = 0,\ W_j = c\} \Big]$
$\mathrm{MRD}_{(X,Y)}$	$\frac{1}{n_D} \sum_{i=1}^{n} \widehat{\mathrm{risk}}(X_i, Y_i)\, I(D_i = 1) - \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n} \widehat{\mathrm{risk}}(X_i, Y_i)\, I(D_i = 0)$	$\frac{1}{n_D} \sum_{i=1}^{n} \widehat{\mathrm{risk}}(X_i, Y_i)\, I(D_i = 1) - \sum_{c=1}^{C} \frac{N_{\bar{D},c}}{N} \frac{1}{n_{\bar{D},c}} \sum_{i=1}^{n} \widehat{\mathrm{risk}}(X_i, Y_i)$ $I(D_i = 0, W_i = c)$
$\mathrm{AARD}_{(X,Y)}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \frac{N_D}{N},\ D_i = 1\}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \frac{N_D}{N},\ D_i = 1\}$
	$-\frac{1}{n_{\bar{D}}}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \frac{N_D}{N},\ D_i = 0\}$	$-\Sigma_{c=1}^{C} \frac{N_{\bar{D},c}}{N_{\bar{D}}} \frac{1}{n_{\bar{D},c}} \Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \frac{N_D}{N},\ D_i = 0,\ W_i = c\}$
NRI(>0)	$2\Big[\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \widehat{\mathrm{risk}}(X_i),\ D_i = 1\}$	$2\Big[\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \widehat{\mathrm{risk}}(X_i),\ D_i = 1\}$
	$-\frac{1}{n_{\bar{D}}}\Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \widehat{\mathrm{risk}}(X_i),\ D_i = 0\} \Big]$	$-\Sigma_{c=1}^{C} \frac{N_{\bar{D},c}}{N_{\bar{D}}} \frac{1}{n_{\bar{D},c}} \Sigma_{i=1}^{n} I\{\widehat{\mathrm{risk}}(X_i, Y_i) > \widehat{\mathrm{risk}}(X_i),\ D_i = 0,\ W_i = c\} \Big]$

Table 10 Estimators of performance measures—semiparametric estimators using the enhanced risk model and both cohort and case-control subset data. Superscripts ‘*cohort*’ and ‘*cc*’ denote data from the cohort and the case-control subset, respectively

Name	Estimator
$\text{HR}^{D}_{(X,Y)}(r)$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r,\ D_i^{cc}=1\}$
$\text{HR}^{\bar{D}}_{(X,Y)}(r)$	$1-\frac{1}{N_{\bar{D}}}\sum_{j=1}^{N}\widehat{F}_0\left\{\frac{\frac{\text{logit}(r)-\widehat{\theta}_0-\widehat{\theta}_1 X_j^{cohort}}{\widehat{\theta}_2}-\widehat{\mu}^{\bar{D}}(X_j^{cohort})}{\widehat{\sigma}^{\bar{D}}(X_j^{cohort})}\right\} I\{D_j^{cohort}=0\}$
$\text{B}_{(X,Y)}(r)$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r,\ D_i^{cc}=1\}-\frac{N_{\bar{D}}}{N_D}\frac{r}{1-r}\left[1-\frac{1}{N_{\bar{D}}}\sum_{j=1}^{N}\widehat{F}_0\left\{\frac{\frac{\text{logit}(r)-\widehat{\theta}_0-\widehat{\theta}_1 X_j^{cohort}}{\widehat{\theta}_2}-\widehat{\mu}^{\bar{D}}(X_j^{cohort})}{\widehat{\sigma}^{\bar{D}}(X_j^{cohort})}\right\} I\{D_j^{cohort}=0\}\right]$
$\text{ROC}_{(X,Y)}(p^{\bar{D}})$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r(p^{\bar{D}}),\ D_i^{cc}=1\},$
	where $r(p^{\bar{D}})$ s.t. $\frac{1}{N_{\bar{D}}}\sum_{j=1}^{N}\widehat{F}_0\left\{\frac{\frac{\text{logit}(r(p^{\bar{D}}))-\widehat{\theta}_0-\widehat{\theta}_1 X_j^{cohort}}{\widehat{\theta}_2}-\widehat{\mu}^{\bar{D}}(X_j^{cohort})}{\widehat{\sigma}^{\bar{D}}(X_j^{cohort})}\right\} I\{D_j^{cohort}=0\}=1-p^{\bar{D}}$
$\text{ROC}^{-1}_{(X,Y)}(p^{D})$	$1-\frac{1}{N_{\bar{D}}}\sum_{j=1}^{N}\widehat{F}_0\left\{\frac{\frac{\text{logit}(r(p^{D}))-\widehat{\theta}_0-\widehat{\theta}_1 X_j^{cohort}}{\widehat{\theta}_2}-\widehat{\mu}^{\bar{D}}(X_j^{cohort})}{\widehat{\sigma}^{\bar{D}}(X_j^{cohort})}\right\} I\{D_j^{cohort}=0\},$
	where $r(p^{D})$ s.t. $\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > r(p^{D}),\ D_i^{cc}=1\}=p^{D}$
$\text{AUC}_{(X,Y)}$	$\frac{1}{n_D}\Sigma_{i=1}^{n}\left[I(D_i^{cc}=1)\ \frac{1}{N_{\bar{D}}}\sum_{j=1}^{N}\widehat{F}_0\left\{\frac{\frac{\text{logit}\ \widehat{\text{risk}}(X_i^{cc},Y_i^{cc})-\widehat{\theta}_0-\widehat{\theta}_1 X_j^{cohort}}{\widehat{\theta}_2}-\widehat{\mu}^{\bar{D}}(X_j^{cohort})}{\widehat{\sigma}^{\bar{D}}(X_j^{cohort})}\right\} I\{D_j^{cohort}=0\}\right]$
$\text{MRD}_{(X,Y)}$	$\frac{1}{n_D}\sum_{i=1}^{n}\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc})\ I(D_i^{cc}=1)-\frac{1}{N_{\bar{D}}}\sum_{i=1}^{N}\frac{1}{n_{\bar{D}}}\sum_{j=1}^{n}\widehat{\text{risk}}\Big\{X_i^{cohort},\ \frac{Y_j^{cc}-\widehat{\mu}^{\bar{D}}(X_j^{cc})}{\widehat{\sigma}^{\bar{D}}(X_j^{cc})}\widehat{\sigma}^{\bar{D}}(X_i^{cohort})$
	$+\widehat{\mu}^{\bar{D}}(X_i^{cohort})\Big\}\ I(D_i^{cohort}=0,\ D_j^{cc}=0)$

Table 10 continued

Name	Estimator
AARD$_{(X,Y)}$	$\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > \frac{N_D}{N},\ D_i^{cc} = 1\} -$
	$\left[1 - \frac{1}{N_{\bar{D}}}\sum_{j=1}^{N} \widehat{F}_0\left\{\frac{\frac{\text{logit}(\frac{N_D}{N}) - \widehat{\theta}_0 - \widehat{\theta}_1 X_j^{cohort}}{\widehat{\theta}_2} - \widehat{\mu}^{\bar{D}}(X_j^{cohort})}{\widehat{\sigma}^{\bar{D}}(X_j^{cohort})}\right\} I\{D_j^{cohort} = 0\}\right]$
NRI(>0)	$2\left(\frac{1}{n_D}\Sigma_{i=1}^{n} I\{\widehat{\text{risk}}(X_i^{cc}, Y_i^{cc}) > \widehat{\text{risk}}(X_i^{cc}),\ D_i^{cc} = 1\}\right.$
	$\left. - \left[1 - \frac{1}{N_{\bar{D}}}\sum_{j=1}^{N} \widehat{F}_0\left\{\frac{\frac{\text{logit}\ \widehat{\text{risk}}(X_j^{cohort}) - \widehat{\theta}_0 - \widehat{\theta}_1 X_j^{cohort}}{\widehat{\theta}_2} - \widehat{\mu}^{\bar{D}}(X_j^{cohort})}{\widehat{\sigma}^{\bar{D}}(X_j^{cohort})}\right\} I\{D_j^{cohort} = 0\}\right]\right)$

References

Anderson M, Wilson PW, Odell PM, Kannel WB (1991) An updated coronary risk profile: a statement for health professionals. Circulation 83:356–362

Baker SG (2009) Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst 101:1538–1542

Breslow NE, Day NE (1980) Statistical methods in cancer research, vol 1. International Agency for Research on Cancer, Lyon

Breslow NE, Cain KC (1988) Logistic regression for two-stage case–control data. Biometrika 75(1):11–20

Breslow NE (1996) Statistics in epidemiology: the case–control study. J Am Stat Assoc 91(433):14–27

Bura E, Gastwirth JL (2001) The binary regression quantile plot: assessing the importance of predictors in binary regression visually. Biomet J 43:5–21

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall/CRC, New York

Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001) Executive summary of the Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). J Am Med Assoc 285(19):2486–2497

Fears TR, Brown CC (1986) Logistic regression methods for retrospective case–control studies using complex sampling procedures. Biometrics 42:955–960

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Shairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81(24):1879–1886

Gail MH, Costantino JP, Bryant J, Croyle R, Freedman L, Helzlsouer K, Vogel V (1999) Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 91(21):1829–1846

Gordon T, Kannel WB (1982) Multiple risk functions for predicting coronary heart disease: the concept, accuracy, and application. Am Heart J 103:1031–1039

Gu W, Pepe M (2009a) Measures to summarize and compare the predictive capacity of markers. Int J Biostat 5. doi:10.2202/1557-4679.1188

Gu W, Pepe MS (2009b) Estimating the capacity for improvement in risk prediction. Biostatistics 10(1):172–186

Heagerty PJ, Pepe MS (1999) Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in children. Appl Stat 48(4):533–551

Huang Y, Pepe MS, Feng Z (2007) Evaluating the predictiveness of a continuous marker. Biometrics 63(4):1181–1188

Huang Y, Pepe MS (2009) Semiparametric methods for evaluating risk prediction markers in case–control studies. Biometrika 96(4):991–997

Janes H, Pepe MS (2008) Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. Biometrics 64:1–9

Janes H, Pepe MS (2009) Adjusting for covariate effects on classification accuracy using the covariate adjusted ROC curve. Biometrika 96:371–382

Janssens ACJW, Deng Y, Borsboom GJJM, Eijkemans MJC, Habemma JDF, Steyerberg EW (2005) A new logistic regression approach for the evaluation of diagnostic test results. Ann Intern Med 25(2):168–177

Kannel WB, McGee D, Gordon T (1976) A general cardiovascular risk profile: the Framingham study. Am J Cardiol 38:46–51

Kerr KF, McClelland RL, Brown ER, Lumley T (2011) Evaluating the incremental value of new biomarkers with integrated discrimination improvement. Am J Epidemiol 174(3):364–374

Krijnen P, van Jaarsveld BC, Steyerberg EW, Man in't Veld AJ, Schalekamp MADH, Habbema JDF (1998) A clinical prediction rule for renal artery stenosis. Stat Med 129(9):705–711

Mealiffe ME, Stokowski RP, Rhees BK, Prentice RL, Pettinger M, Hinds DA (2010) Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst 102(21):1618–1627

Pauker SG, Kassierer JP (1980) The threshold approach to clinical decision making. N Engl J Med 302:1109–1117

Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172

Pencina MJ, D'Agostino RB Sr, Steyerberg EW (2011) Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 30:11–21

Pepe MS, Kerr KF, Longton G, Wang Z (2013) Testing for improvement in prediction model performance. Stat Med. doi:10.1002/sim.5727

Pepe MS, Feng Z, Janes H, Bossuyt P, Potter J (2008) Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 100(20):1432–1438

Pepe MS, Fan J, Seymour CW, Li C, Huang Y, Feng Z (2012) Biases introduced by choosing controls to match risk factors of cases in biomarker research. Clin Chem 58(8):1242–1251

Pepe MS, Janes H (2013) Methods for evaluating prediction performance of biomarkers and tests. In MLT Lee, M Gail, G Satten, T Cai, A Gandy, R Pfeiffer (Ed.), Risk assessment and evaluation of predictions. Springer

Pfeiffer RM, Gail MH (2011) Two criteria for evaluating risk prediction models. Biometrics 67:1057–1065

Prentice RL, Pyke R (1979) Logistic disease incidence models and case–control studies. Biometrika 66:403–411

Truett J, Cornfield J, Kannel W (1967) A multivariate analysis of the risk of coronary heart disease in Framingham. J Chron Dis 20:511–524

Vickers AJ, Cronin AM, Begg CM (2011) One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol 11(13). doi:10.1186/1471-2288-11-13

Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26:565–574