A concordance index for matched case-control studies with applications in cancer risk

Adam R. Brentnall[a]*, Jack Cuzick[a], John Field[b], Stephen W. Duffy[a]

In unmatched case-control studies the area under the receiver operating characteristic (ROC) curve (AUC) may be used to measure how well a variable discriminates between cases and controls. The AUC is sometimes used in matched case-control studies by ignoring matching, but it lacks interpretation because it is not based on an estimate of the ROC for the population of interest. We introduce an alternative measure of discrimination that is the concordance of risk factors conditional on the matching factors. Parametric and non-parametric estimators are given for different matching scenarios, and applied to real data from breast and lung cancer case-control studies. Diagnostic plots to verify the constancy of discrimination over matching factors are demonstrated. The proposed simple measure is easy to use, interpret, more efficient than unmatched AUC statistics, and may be applied to compare the conditional discrimination performance of risk factors. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: Case-control studies, Classification, Discrimination, Matching

1. Introduction

In matched case-control studies the primary aim is sometimes to assess whether a quantitative variable predicts a binary reponse, and a secondary objective is to measure its discriminatory power on a scale that is not sample dependent. It is recommended that the primary analysis should take account of the matching, commonly via conditional logistic regression [1, 2]. It is less clear what methodology to use in order to measure the rank ordering performance of a marker in a matched design. In this article we develop a simple concordance index for matched data, and show how it may be estimated and used in the analysis of case-control studies.

Determining how well a variable discriminates between cases and controls in a matched study has a lot in common with adjusting for covariates in regression. In that context, it is well known that if the variable of interest x is independent of another covariate z, then including both in the regression against response y can be more efficient than just using x (*e.g.* [3]). Further, when x and z are correlated then the conditional effect of x on y given z will be different than the marginal effect when z is ignored. In a matched context z is used to match cases ($y = 0$) and controls ($y = 1$) and we seek a measure of how well x discriminates between them.

[a]*Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, UK.* [b]*Roy Castle Lung Cancer Research Programme, The University of Liverpool Cancer Research Centre, Institute of Translational Medicine, The University of Liverpool, UK.*
* Correspondence to: Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queeny Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK; a.brentnall@qmul.ac.uk

The utility of an odds ratio, C-index or other performance measures from a case-control study depends on the context. In many circumstances the odds ratio is more relevant than a change in AUC. For example, mutations in the BRCA1 and BRCA2 genes are known to confer substantially higher risks of breast and ovarian cancer [4]. Testing women with high familial risk for these mutations is beneficial because they may then consider preventive options; but the change in the population AUC from a model that incorporates BRCA status will be very small because they are rare. On the other hand, an odds ratio may be less helpful to assess the rank ordering by markers or combination of markers for cancer screening strategies, such as possible triage tests for cervical cancer screening [5]. There is a clear need for methods to estimate both relative risks and measures of discrimination.

The next section introduces our measure of rank ordering performance for matched studies, and then methods for estimation are considered. The approach is demonstrated through three examples from cancer risk, before conclusions are drawn.

2. A matched concordance index

The concordance index and also known as Kendall's τ [6], is the probability $\tau = P(y_1 > y_2, x_1 > x_2) + P(y_1 < y_2, x_1 < x_2)$ that differences between two continuous random variables variables y and x are in the same direction for two randomly sampled individuals 1 and 2; pairs of random variables (x_1, y_1) and (x_2, y_2) are independent and identically distributed (iid). When y is a binary indicator for case-control status and x is a predictor variable, then it is equal to the area under the receiver operating characteristic curve (AUC) [7]. An issue for matched studies is that estimating the C-index by ignoring matching is equivalent to analysing the data by an unconditional logistic regression model. Despite this, the approach is sometimes used, including in evaluating mammographic density as a breast-cancer risk factor in case-control studies (*e.g.* [8, 9]). However, this approach is biased for assessment of variable performance in combination with others, and to assess the incremental increase in AUC, because by design the sample is balanced for the matching risk factors [10]. We propose to side-step these issues by targeting a conditional C-index instead, defined as

$$C_z \quad = \quad P(y_1 > y_2, x_1 > x_2 \mid z_1 = z_2 = z) + P(y_1 < y_2, x_1 < x_2 \mid z_1 = z_2 = z) \tag{1}$$

which is simply the concordance probability for two randomly sampled individuals with the same values for matching factors or sets z_1 and z_2. The matching factors might be about each individual such as their age and gender, or they might be less quantitative so that, for example, individuals are matched 1:1 with a doctor's surgeries and each practice only contributes a single pair. C_z is a generalisation of Kendall's τ to the case of matching with another variable. An overall matched C-index is defined

$$C \quad = \quad \int C_z p(z) \mathrm{d}z \tag{2}$$

where $p(.)$ is a probability mass function. In this article we focus on the case where C_z is constant over z, and so the distribution of z is not required. This generalisation of Kendall's τ only includes pairs with equal values of z; we argue that it is of greater validity and more intuitive for a matched case-control study than an AUC statistic obtained by ignoring the matching.

A hypothetical example is used in Figure 1 to help further introduce the measure, and to show the difference from the classical concordance index. The figure shows a case-control study with two matching groups (young and old), where the mean response differs between the groups. The setup is motivated by experience with biomarkers that are correlated with age, such as mammographic breast density and DNA methylation [11]. Because there is very little overlap between
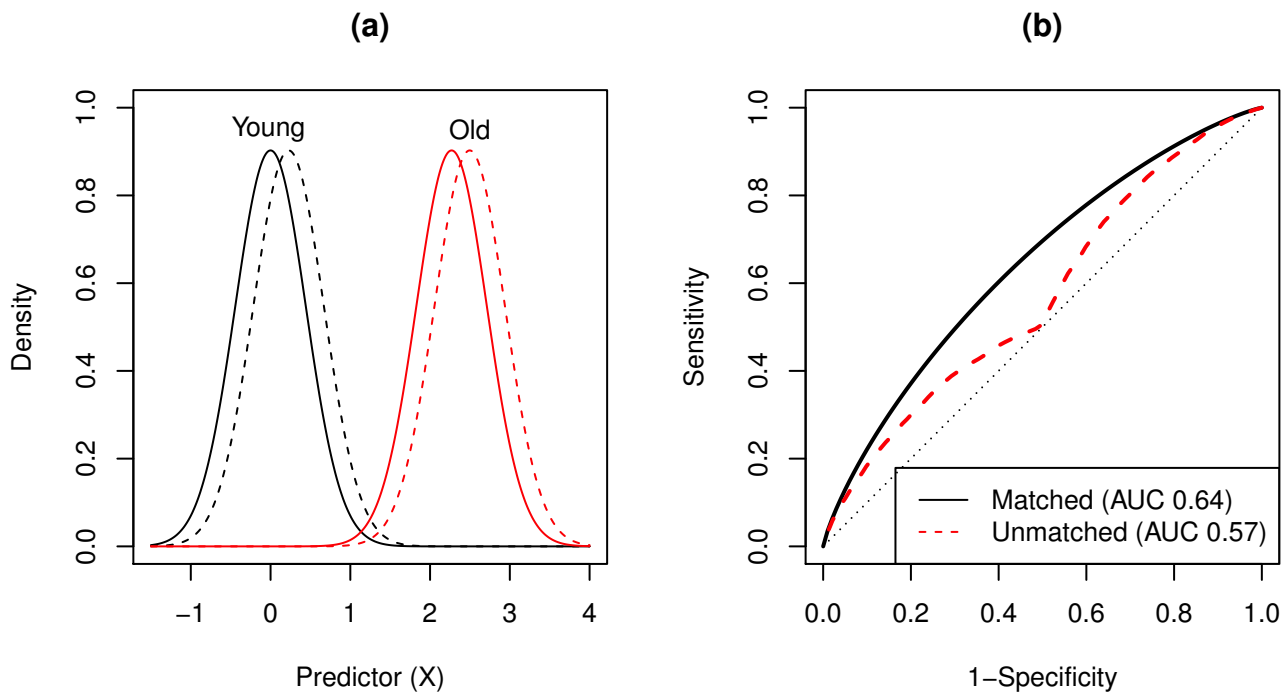
(a) **(b)**

Figure 1. An example to show a difference between matched and unmatched concordance indices. There are two matching groups 'young' and 'old', and a biomarker that is correlated with age. Plot (a) shows the density of the biomarker in young and old groups, split by case (– – –) and control (——) status. Plot (b) shows the corresponding ROCs, where in the example the conditional AUC is the same in young and old groups.

the two matching age groups, the unconditional ROC has a turning point at 0.5 on the x-axis. This makes interpretation of AUC difficult, partly because it depends on the age sampling ratio. The example is designed so that the conditional ROCs for the two groups are identical, and so is the average ROC regardless of the age sampling ratio. Then the matched C-index obtained as the mean of both matching groups is interpreted as the concordance probability for two individuals of the same age class. If there were n groups matched 1:1, then it would be interpreted as the concordance probability for a matched pair of individuals.

Figure 1 may be also be used to show a duality between the matched C-index and logistic regression odds ratios. The figure is based on a normal distribution for $p(x \mid y, z)$, the probability distribution of x given case-control status y and match z, with mean shift δ between cases ($y = 1$) and controls ($y = 0$) that is the same for all matched groups. In the following we take equal unit variance for simplicity. Following Bayes' rule

$$\frac{P(y = 1 \mid x, z)}{P(y = 0 \mid x, z)} = \frac{p(x \mid y = 1, z)P(y = 1 \mid z)}{p(x \mid y = 0, z)P(y = 0 \mid z)}$$
$$= \exp(a + \delta x) \qquad (3)$$

where $a = \ln\{P(y = 1 \mid z)/P(y = 0 \mid z)\} - \delta^2/2$. Equation (3) is of the same form as the odds ratio for x from a conditional logisitic regression, *i.e.* the odds ratio for a unit change in x is $\exp(\delta)$ and the logistic regression model holds, see also [12]. Further, the C-index for each value of the matching variable z is $\Phi(\delta/\sqrt{2})$, where Φ is a standard normal cumulative distribution function [13]. Thus under this model, the odds ratio and matched C-index are dual because both are functions of the same parameter δ. An obvious consequence is that under this simple model the odds ratio $\exp(\delta)$ is constant over matching groups z, and so is the matched C-index $C = \Phi(\delta/\sqrt{2})$.

A disadvantage of the odds ratio is that when the aim is to use a variable or combination of variables to dichotomise a prediction into good *vs* bad, the odds ratio depends on the cutpoint. An advantage of the concordance index is its

interpretation as mean sensitivity over all possible cutpoints, although this may sometimes be a weakness [14].

The proposed measure follows a similar approach to [15, 16], where covariate-adjusted ROC curves were estimated from which AUC statistics may be obtained. Using the example in Figure 1, the covariate-specific ROCs for the two age groups are both the matched ROC, as is the covariate-adjusted ROC (AROC, their vertical average). AROCs do not require that the ROC given covariate z is constant; the AROC is defined as a weighted average of covariate-specific ROCs regardless. However, interpretation is more difficult when the assumption does not hold because the AROC will depend on the sampling distribution of z. An assumption of constant covariate-specific ROCs might be seen to be a stronger assumption than constant C_z, since in theory C_z could be constant even if the covariate-specific ROCs are not. However, the main difference between the approach in this paper and AROC estimation is that we use an assumption about the constancy of C_z to estimate the matched C-index, without modelling the matching factors. This is particularly relavent when it is not possible to model the matching covariates, such as for example 1:1 matching based on individual doctor's surgeries.

The matched C-index for the variable of interest (x) relates theoretically to the AUC from the population ROC as follows. If $C = 1.0$ then x provides perfect discrimination after allowing for the matching factors z, and so in the population the AUC for a model that includes z (*e.g.* age) and x would also be 1.0; but the AUC for an unmatched comparison from a matched study would not necessarily be 1.0. If $C = 0.5$ then the rank ordering of x is no better than random after allowing for the matching factors z, and so the change in AUC in the population when adding x to a model that includes the matching variables z (*e.g.* age) would be zero. In general when $0.5 < C < 1.0$ it is not possible to relate the matched C-index to the incremental benefit of x over z, as described by [15]. However, the matched and unmatched C-indicies for x alone will coincide when C_z is constant and x is not correlated to z, but the matched C-index will be more efficient for statistical inference. Although the population ROC and associated AUC cannot be directly calculated from a matched study, some work has been undertaken to use matched studies to estimate the AUC statistic that would be observed in the population, by using extra information on the distribution of matching risk factors in controls [10].

3. Estimation

We turn next to methods to estimate the matched C-index. Let x_{jik} be measured in matching group $i = 1, \ldots, n$ for controls and cases ($j = 0, 1$) and replicate $k = 1, \ldots, n_{ij}$, and define

$$C_i = P(x_{1i} > x_{0i}) + 1/2P(x_{1i} = x_{0i}), \qquad (4)$$

where the second term in the definition allows for ties, and ensures that $C_i = 0.5$ if the variable is no better than random discrimination. The methods in this section to form an overall matched C-index estimate take that C_i is constant over $i = 1, \ldots, n$. Some diagnostic methods to assess the assumption will be discussed.

3.1. One to m matching

Two methods for when there is one case per matched group are next introduced.

3.1.1. Non-parametric estimator Use

$$T_i = m^{-1} \sum_{k=1}^{m} \{I(x_{1i} > x_{0ik}) + 1/2I(x_{1i} = x_{0ik})\} \qquad (5)$$

to estimate C_i for $i = 1, \ldots, n$, where $I(.)$ is the indicator function and $x_{1i} = x_{1i1}$ to simplify notation. The average of T_i over the sample provides an overall estimate of the matched C-index

$$C_1 \quad = \quad n^{-1} \sum_{i=1}^{n} T_i. \tag{6}$$

3.1.2. A parametric estimator Suppose that x_{jik} has mean μ_{ji} and variance σ_j^2, so that the mean is potentially linked with matching variable i and case-control status j, but the variance is constant over the matching variables i. Define

$$S_i \quad = \quad m^{-1} \sum_{k=1}^{m} x_{1i} - x_{0ik} \tag{7}$$

for $i = 1, \ldots, n$ and

$$S \quad = \quad n^{-1} \sum_{i=1}^{n} S_i. \tag{8}$$

Denoting the variance over S_i by σ_S^2, a z-test statistic is

$$T_Z \quad = \quad \sqrt{n} S / \hat{\sigma}_S \tag{9}$$

where $\hat{\sigma}_S$ is the sample standard deviation. For large enough n the hypothesis that the mean differs between cases and controls may be tested because it is asymptotically standard normal from the central limit theorem. Another reason for considering this approach is that it may be converted to an estimate of the matched C-index under the following assumptions.

- $E(x_{0i} - x_{1i} \mid z_i) = \Delta$, so that the conditional mean difference between cases and controls is constant over the matching variables.
- x_{ji} is normally distributed.

The matched C-index for this so-called binormal model is [13]

$$C_2 \quad = \quad \Phi\{\Delta/\sigma_\Delta\} \tag{10}$$

where $\Phi(.)$ is a standard normal cumulative distribution function and $\sigma_\Delta = (\sigma_0^2 + \sigma_1^2)^{1/2}$ is the total standard deviation in cases and controls. The numerator Δ may be estimated by S, and the denominator from variance σ_Δ^2 as follows. Define $s_m = m(m-1)/2$, then because $\text{cov}(x_{1i} - x_{0ik}, x_{1i} - x_{0il}) = \sigma_0^2$,

$$\sigma_S^2 \quad = \quad m^{-2}\{m(\sigma_0^2 + \sigma_1^2) + 2s_m(\sigma_0^2)\}. \tag{11}$$

If we take $\sigma_0^2 = \sigma_1^2$ then

$$\sigma_\Delta^2 \quad = \quad \sigma_S^2 m^2 (m + s_m)^{-1}, \tag{12}$$

plugging in $\hat{\sigma}_S^2$ to provide an estimator.

3.1.3. Inference With $1 : 1$ matching a binomial sample is obtained for the non-parametric estimate, and standard techniques may be used such as profile likelihood. For $1 : m$ matching in each group, $\text{var}(C_1)$ depends on the distribution of x_{ji}. Under the parametric model introduced a confidence interval may be obtained by estimating a confidence interval

on the term inside (10), and then applying the transformation. A bootstrap may also be applied by resampling with replacement from the matched groups or pairs. Some demonstration code for the statistical software R [17] is provided in an online appendix.

3.2. Matched groups of variable size

Consider where there is $n_{0i} : n_{1i}$ matching, so that each group $i = 1, \ldots, n$ may have $n_{0i} > 0$ controls and $n_{1i} > 0$ cases. Initially in each group the target is the matched C-index for that group alone. We focus on a non-parametric estimate for each matching group $i = 1, \ldots, n$

$$C_{3i} = (n_{0i}n_{1i})^{-1} \sum_{k=1}^{n_{0i}} \sum_{l=1}^{n_{1i}} \psi(x_{1il}, x_{0ik}), \tag{13}$$

where

$$\psi(x_{1il}, x_{0ik}) = \begin{cases} 1 & x_{1il} > x_{0ik} \\ 1/2 & x_{1il} = x_{0ik} \\ 0 & x_{1il} < x_{0ik} \end{cases}$$

which generalises (5). If R_{il} denotes the rank of the x_{jik} in the combined sample for $j = 0, 1$ and $k = 1, \ldots, n_{ji}$ with the ranks of tied observations averaged, then the Wilcoxon ranksum test statistic given by $W_i = \sum_{l=1}^{n_i} R_{il}$ is $n_{0i}n_{1i}C_{3i} + n_{i0}(n_{i0} + 1)/2$, and may be used to calculate C_{3i}.

We are interested in the overall matched C-index over the n groups, where the number of cases and controls may be different in each. If the matching groups are independent, then a weighting factor based on the inverse variance σ_{Ci}^2 from each group will minimise the variance of a weighted average, *i.e.*

$$w_i = \sigma_{Ci}^{-2} / \sum_{i=1}^{n} \sigma_{Ci}^{-2}, \tag{14}$$

so that

$$C_3 = \sum_{i=1}^{n} w_i C_{3i}. \tag{15}$$

One choice is to use the weights under the null hypothesis that $C_{3i} = 0.5$ [18]:

$$\sigma_{1Ci}^2 = \frac{n_{0i} + n_{1i} + 1}{12 n_{0i} n_{1i}}. \tag{16}$$

An alternative is to estimate the variance of C_{3i} using one of the methods described by [19], including [20]. Let $Q_{ijk} = \psi(x_{1ij}, x_{0ik})$, so that the column means are $\bar{Q}_{ij.} = n_{0i}^{-1} \sum_{k=1}^{n_{0i}} Q_{ijk}$ and row means $\bar{Q}_{i.k}$, then

$$\hat{\sigma}_{2Ci}^2 = n_{0i}^{-1} \widehat{\text{var}}(\bar{Q}_{ij.}) + n_{1i}^{-1} \widehat{\text{var}}(\bar{Q}_{i.k}) + (n_{0i}n_{1i})^{-1} C_{3i}(1 - C_{3i}) \tag{17}$$

where $\widehat{\text{var}}(.)$ are sample variance estimators.

3.2.1. Inference
It is easy to see that

$$\text{var}(C_3) = (\sum_{j=1}^{n} \sigma_{Cj}^{-2})^{-2} \sum_{i=1}^{n} \sigma_{Ci}^{-2}, \tag{18}$$

Prepared using **simauth.cls**

substituting σ_{1Ci}^2 or $\hat{\sigma}_{2Ci}^2$ for σ_{Ci}^2, where in theory the former might be used to test the null hypothesis that $C_3 = 0.5$, and the latter to form a confidence interval. In practice, for case-control studies where the number of samples in each group $(n_{0i} + n_{1i})$ is not large, or n_{0i} or $n_{1i} = 1$, then weights from σ_{1Ci}^2 are preferable because $\hat{\sigma}_{2Ci}^2$ is less stable and may be zero. For this reason we also recommend using a bootstrap for confidence intervals, rather than rely on an asymptotic approximation.

3.3. Checking assumption of constancy over matching groups

In the above an overall matched C-index was obtained under an assumption that it is constant over the matched groups. A number of approaches might be used to verify this assumption, based on inspection of T_i from (5) and matching factor z_i. In an example that follows we plot T_i vs z_i, and fit a line using a local regression smoother [21].

3.4. Relative efficiency

It is well established that matched tests are more efficient than unmatched tests, and the same applies to the matched C-index. To demonstrate consider 1:1 matching when the variable of interest is independent of matching, there are no ties and the aim is to reject the null hypothesis that $C = 0.5$. The sample propotion estimator of a binomial parameter q has variance $q(1-q)/n$. Since (6) is of the same form as a binomial estimator, under the null that $C = 0.5$ it has variance $(4n)^{-1}$. The variance of an unmatched estimator of C from a Wilcoxon statistic with $n/2$ in both groups is $(3n)^{-1}$, as seen by plugging in $n_0 = n_1 = n/2$ into (16). Thus the matched approach is more efficient because its variance is $3/4$ of the unmatched estimator. When the matching variable is correlated to the predictor of interest, the matched approach will be even more powerful because the unmatched approach is a conservative estimator of C with asymptotic bias towards 0.5, as shown more generally for ROCs by [15].

4. Examples

In this section real examples from cancer risk studies are used to demonstrate how the matched C-indicies C_1, C_2 and C_3 can be used to help plan and evaluate matched case-control studies.

4.1. Breast density

A case-control study of mammographic density was nested within the International Breast Cancer Intervention Study-I (IBIS-I), a randomised trial of tamoxifen versus placebo in women at high-risk of developing breast cancer [22]. Crudely, breast density is the amount of tissue of white appearance on a mammogram (breast x-ray), which corresponds to the amount of fibroglandular rather than fatty tissue. It is one of the strongest risk factors for breast cancer. Various measures of mammographic density have been developed, and in the following we use a visually-assessed percentage density measurement that ranged between 0 and 100% in 5% intervals, and was described further by [22]. The focus here is discrimination performance, and for this we use participants from the placebo arm in this study (72 cases, 486 controls), where the median follow-up for controls was 11.6 years, and median time to diagnosis for cases was 5.1 years.

It is well known that percentage breast density is negatively correlated with age, but that both factors are positively correlated with breast-cancer risk. In the case-control study the distribution of age was balanced between cases and controls. Table 1 provides summary statistics of the distribution of percent density by age and case-control status. It can be seen that a similar pattern to that introduced by Figure 1 is observed: cases have greater mammographic density than controls by age group, but younger women have a larger percentage mammographic density than older women.

Applying a conditional logistic regression model to 22 matching age groups had a likelihood-ratio χ_1^2 of 7.9. The AUC obtained by calculating a ROC ignoring the matching on age was 0.54 (95% DeLong confidence interval 0.48-0.61). The

Table 1. Summary of percentage mammographic density measurements in three age groups. The percentile points of mammographic density and the AUC for each age group are given.

Age	Status	n	10%	25%	50%	75%	90%	AUC	95% CI
>55	Control	94	0	5	20	45	75	0.63	0.42 - 0.83
	Case	18	0	15	30	75	80		
45-55	Control	332	5	15	45	75	85	0.61	0.52 - 0.70
	Case	44	15	30	70	80	90		
<45	Control	60	5	30	55	75	80	0.58	0.43 - 0.73
	Case	10	20	40	65	85	90		

matched C-index C_3 from (15) with (16) for the weights was 0.60, with a 95% non-parametric bootstrap CI 0.53 - 0.67. Inference from the matched C-index agrees with the conditional logistic regression model; the unmatched AUC does not.

4.2. Lung cancer

In the Liverpool lung project, [23] matched by age and gender 579 lung cancer cases from Liverpool to two controls, and used conditional logistic regression to develop the Liverpool lung project risk score (LLP) based on five factors (smoking, pneumonia, asbestos, previous tumour, family history). The unmatched AUC was estimated to be around 0.7 in the development data. Two validation data sets that were not matched for age had AUCs closer to 0.8 [24]. We next explore the LLP score using the original development data and our matched C-index.

The matched C-index C_1 from (6) was 0.80 (95% non-parametric bootstrap CI 0.78 - 0.83). Figure 2a shows the unmatched ROC in comparison with two AROCs obtained using 5-yr strata either side of age 70, both weighted by the number of matched groups in each strata. The reason that the estimate of C was higher than the unmatched AUC was partly because the matching variables were correlated with the LLP score: the Spearman coefficient with age was 0.59. All the strata-specific ROCs for individuals younger than 70 years were above the unmatched ROC for individuals younger than 70 (342 (65%) cases). Figure 2b shows an estimate of C against age, where the LLP score appears to disciminate better for younger age groups. Figure 2a shows the same pattern, where the AROC for those aged 70 or more are worse than the younger individuals. This is of interest for the intended use of the model, to stratify risk for lung-cancer screening. It also justifies presenting two estimates of C: for those aged 70 or more C_1 was 0.74 (95% CI 0.69 - 0.78); for those younger than 70 it was 0.84 (95% CI 0.81 - 0.87).

Gender was also investigated, but the model appeared to discriminate equally well for both male and female subjects. The matched C-index for men younger than 70 was 0.84 (95% CI 0.81 - 0.89), and 70+ was 0.73 (95% CI 0.66 - 0.78); for women it was respectively 0.82 (95% CI 0.77 - 0.87) and 0.75 (95% CI 0.67 - 0.82).

4.3. Planning a matched study

A further use of the matched C-index is to help plan studies, and to relate primary testing objectives to discrimination performance. The test statistic in (9) may be used to calculate power for different matching scenarios via (11), and tied to the matched C-index C_2 in (10) through the binormal model. To demonstrate, consider a case-control study designed to validate a polygenic single nucletide polymorphism (SNP) score to predict breast cancer risk. Suppose a maximum 400 cases are available, and the problem is to decide the sampling ratio of controls. Note that here the expected matched and unmatched C-indicies will coincide because age is not correlated to the SNP score, but the matched C-index is more efficient. Let the primary aim of the study be to validate the performance of a panel of 49 breast-cancer risk SNPs, that have been reported following the first 18 by [25].
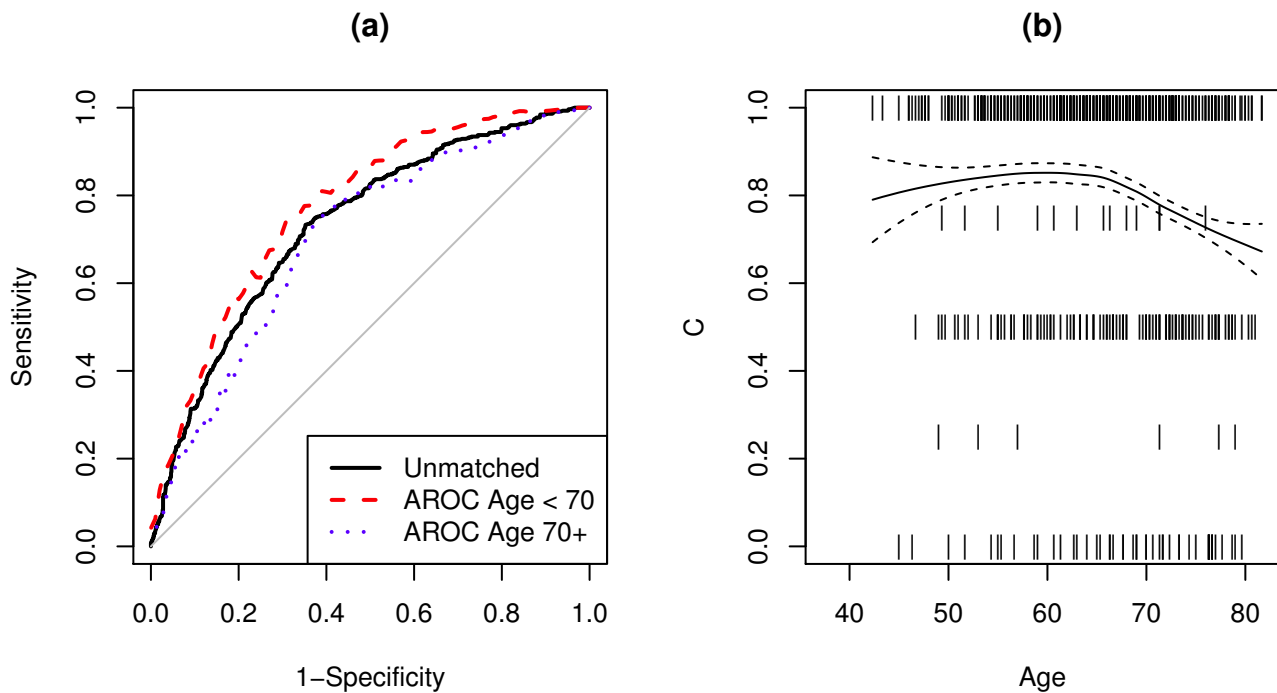
(a)

(b)

Figure 2. Matched C-index performance of the LLP model in the development sample. Panel (a) shows age adjusted ROCs based on 5-yr age-band stratification, and when matching is ignored. Panel (b) shows an estimate of C (—) plus and minus standard error ($-\,-\,-$) against age, where the data points ($|$) are T_i from (5).

The distribution of a relative risk SNP score in the target population was estimated from the relative risk estimates and allele frequencies in [26] via simulation. From the central limit theorem it was approximately log normal with a mean log score difference of 0.088, and standard deviation of 0.295 in both groups. Under the binormal model this corresponds to an AUC of $\Phi(0.088/\sqrt{(2 \times 0.295^2)}) = 0.58$. However, because the relative risks are estimates it is likely that some regression to the mean of the SNP score to risk will be observed, so we consider the alternative hypothesis that AUC will be $\Phi(0.066/(\sqrt{2 \times 0.295^2})) = 0.56$. With 400 cases, the power to reject the null that AUC = 0.5 from 1:1, 2:1 or 3:1 matching via equation (12), would be approximately 89%, 98% and 99% respectively at the 5% level. This suggests that 3:1 matching would not be needed.

5. Conclusion

In this article a concordance index for matched case-control studies was introduced. Non-parametric and parametric estimators were developed for different matching scenarios by building on established methods. The measure was demonstrated in the context of design, analysis and interpretation of case-control studies of cancer.

The estimators use an average over the matched groups because it was assumed that the concordance probability was constant over them. One way to verify the assumption is to compare each matching factor z_i with individual C-index components. If distinct groups are observed, then it might be more useful to report the matched C-index in each stratum, as was seen in a real data set for lung cancer risk.

The measure is straightforward to interpret, and can be used to compare the discriminatory power of different risk factors. For example, the breast-cancer SNPs were expected to confer a matched C-index of 0.58 after allowing for age; visually assessed breast density in the sample of high-risk women was estimated to confer a matched C-index of similar magnitude after allowing for age.

Statist. Med. **2014**, 00 1–11
Prepared using **simauth.cls**

Copyright © 2014 John Wiley & Sons, Ltd.

www.sim.org

9

In conclusion, an important use of matched case-control studies is to develop risk or discrimination scores from analyses allowing for matching factors. It therefore seems desirable to measure the discrimination performance of the risk factors after allowing for matching variables; additional assumptions are required to recover unmatched AUC statistics [10]. Although more advanced analysis such as conditional ROC estimates [15, 16] might be used to estimate the same quantity, the proposed estimators here are quite simple, easy to interpret, offer a good starting point for estimation, and lend themselves to diagnostic tests such as the plot introduced. We recommend that matched C-indices be considered when evaluating the discrimination performance of risk factors in matched case-control studies.

Acknowledgements

References

1. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC scientific publications* 1980; (32):5–338. URL http://view.ncbi.nlm.nih.gov/pubmed/7216345.

2. Clayton D, Hills M. *Statistical Models in Epidemiology*. 1st edn., Oxford University Press, 1993.

3. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* 2008; **27**(23):4658–4677, doi:10.1002/sim.3113. URL http://dx.doi.org/10.1002/sim.3113.

4. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, Risch HA, Eyfjord JE, Hopper JL, Southey MC, *et al.*. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *British Journal of Cancer* 2008; **98**(8):1457–1466, doi: 10.1038/sj.bjc.6604305. URL http://dx.doi.org/10.1038/sj.bjc.6604305.

5. Brentnall AR, Vasiljević N, Scibior-Bentkowska D, Cadman L, Austin J, Szarewski A, Cuzick J, Lorincz AT. A DNA methylation classifier of cervical precancer based on human papillomavirus and human genes. *International Journal of Cancer* 2014; **135**(6):1425–1432, doi:10.1002/ijc.28790. URL http://dx.doi.org/10.1002/ijc.28790.

6. Yule GU, Kendall M. *Introduction to the Theory of Statistics*. Fourteenth edn., Charles Griffin & Company: London, 1968.

7. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**(1):29–36. URL http://radiology.rsnajnls.org/content/143/1/29.abstract.

8. Shepherd JA, Kerlikowske K, Ma L, Duewer F, Fan B, Wang J, Malkov S, Vittinghoff E, Cummings SR. Volume of mammographic density and risk of breast cancer. *Cancer Epidemiology Biomarkers & Prevention* 2011; **20**(7):1473–1482, doi:10.1158/1055-9965.epi-10-1150. URL http://dx.doi.org/10.1158/1055-9965.epi-10-1150.

9. Li J, Szekely L, Eriksson L, Heddson B, Sundbom A, Czene K, Hall P, Humphreys K. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. *Breast Cancer Research* 2012; **14**(4):R114+, doi:10.1186/bcr3238. URL http://dx.doi.org/10.1186/bcr3238.

10. Sullivan M, Fan J, Seymour CW, Li C, Huang Y, Feng Z. Biases introduced by choosing controls to match risk factors of cases in biomarker research. *Clinical chemistry* 2012; **58**(8):1242–1251, doi:10.1373/clinchem.2012.186007. URL http://dx.doi.org/10.1373/clinchem.2012.186007.

11. Vasiljević N, Wu K, Brentnall AR, Kim DC, Thorat MA, Kudahetti SC, Mao X, Xue L, Yu Y, Shaw GL, *et al.*. Absolute quantitation of DNA methylation of 28 candidate genes in prostate cancer using pyrosequencing. *Disease Markers* 2011; **30**(4):151–161, doi:10.3233/dma-2011-0790. URL http://dx.doi.org/10.3233/dma-2011-0790.

12. Kay R, Little S. Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika* 1987; **74**(3):495–501, doi:10.1093/biomet/74.3.495. URL http://dx.doi.org/10.1093/biomet/74.3.495.

13. Demidenko E. Confidence intervals and bands for the binormal ROC curve revisited. *Journal of applied statistics* 2012; **39**(1):67–79, doi:10.1080/02664763.2011.578616. URL http://dx.doi.org/10.1080/02664763.2011.578616.

14. Hand DJ. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. *Statistics in Medicine* 2010; **29**(14):1502–1510, doi:10.1002/sim.3859. URL http://dx.doi.org/10.1002/sim.3859.

15. Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics* 2008; **64**(1):1–9, doi:10.1111/j.1541-0420.2007.00823.x. URL http://dx.doi.org/10.1111/j.1541-0420.2007.00823.x.

16. Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* 2009; **96**(2):371–382, doi:10.1093/biomet/asp002. URL http://dx.doi.org/10.1093/biomet/asp002.

17. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2012. URL http://www.R-project.org/, ISBN 3-900051-07-0.

18. Rice JA. *Mathematical statistics and data analysis*. Second edn., Duxbury Press: Belmont, CA, 1995.

19. Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Academic radiology* 1997; **4**(1):49–58. URL http://view.ncbi.nlm.nih.gov/pubmed/9040870.

20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**(3):837–845. URL http://view.ncbi.nlm.nih.gov/pubmed/3203132.

21. Cleveland WS, Grosse E, Shyu WM. Local regression models. *Statistical Models in S*, Chambers JM, Hastie TJ (eds.). chap. 8, Wadsworth & Brooks/Cole, 1992.

22. Cuzick J, Warwick J, Pinney E, Duffy SW, Cawthorn S, Howell A, Forbes JF, Warren RML. Tamoxifen-Induced reduction in mammographic density and breast cancer risk reduction: A nested CaseControl study. *Journal of the National Cancer Institute* 2011; **103**(9):744–752, doi:10.1093/jnci/djr079. URL http://dx.doi.org/10.1093/jnci/djr079.

23. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, Field JK. The LLP risk model: an individual risk prediction model for lung cancer. *British journal of cancer* 2008; **98**(2):270–276, doi:10.1038/sj.bjc.6604158. URL http://dx.doi.org/10.1038/sj.bjc.6604158.

24. Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, Field JK. Predictive accuracy of the liverpool lung project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. *Annals of internal medicine* 2012; **157**(4):242–250, doi:10.7326/0003-4819-157-4-201208210-00004. URL http://dx.doi.org/10.7326/0003-4819-157-4-201208210-00004.

25. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS, *et al.*. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics* 2010; **42**(6):504–507, doi:10.1038/ng.586. URL http://dx.doi.org/10.1038/ng.586.

26. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK, *et al.*. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics* 2013; **45**(4):353–361, doi:10.1038/ng.2563. URL http://dx.doi.org/10.1038/ng.2563.

Appendix

R code

The methods in this article are demonstrated using code for the statistical software R in an online supplement.