Biases Introduced by Choosing Controls to Match Risk Factors of Cases in Biomarker Research
Margaret Sullivan Pepe,1,2* Jing Fan,1,2 Christopher W. Seymour,3 Christopher Li,4 Ying Huang,1 and Ziding Feng1

BACKGROUND: Selecting controls that match cases on risk factors for the outcome is a pervasive practice in biomarker research studies. Such matching, however, biases estimates of biomarker prediction performance. The magnitudes of these biases are unknown.

METHODS: We examined the prediction performance of biomarkers and improvements in prediction gained by adding biomarkers to risk factor information. Data simulated from bivariate normal statistical models and data from a study to identify critically ill patients were used. We compared true performance with that estimated from case control studies that do or do not use matching. ROC curves were used to quantify performance. We propose a new statistical method to estimate prediction performance from matched studies for which data on the matching factors are available for subjects in the population.

RESULTS: Performance estimated with standard analyses can be grossly biased by matching, especially when biomarkers are highly correlated with matching risk factors. In our studies, the performance of the biomarker alone was underestimated whereas the improvement in performance gained by adding the marker to risk factors was overestimated by 2–10-fold. We found examples for which the relative ranking of 2 biomarkers for prediction was inappropriately reversed by use of a matched design. The new approach to estimation corrected for bias in matched studies.

CONCLUSIONS: To properly gauge prediction performance in the population or the improvement gained by adding a biomarker to known risk factors, matched case control studies must be supplemented with risk factor information from the population and must be analyzed with nonstandard statistical methods.

© 2012 American Association for Clinical Chemistry

In studies to evaluate biomarkers for use in diagnostic and prognostic settings, controls are often chosen that are matched to cases in regard to risk factors for the outcome. For example, because age is a strong risk factor for cancer, studies to evaluate biomarkers for early detection of cancer usually select controls that have the same ages as the cases (1–7). The rationale is straightforward: if the biomarker distribution is different in cases vs controls, it cannot be attributed to age differences in an age-matched study. In contrast, in an unmatched study in which cases are older than controls, one would expect biomarker values to differ between cases and controls if the biomarker simply varies with age. Prostate specific antigen is an example of a biomarker that is typically higher in older subjects. Consequently many studies of biomarkers for prostate cancer have employed age-matched controls (8).

Although matching is commonly used in the design of biomarker studies, it is not widely appreciated that the practice of matching severely limits the questions that can be addressed (9). One cannot assess the classification performance of the biomarker in the general population using only data from a matched study. Neither can one quantify the improvement in performance that can be gained by addition of the biomarker to data on risk factors if the study design has matched on risk factors. For example, relative to use of age alone, the improvement in prediction of prostate cancer risk that is gained by knowledge of prostate-specific antigen concentrations cannot be assessed by a case control study in which controls are age matched to prostate cancer cases.

We emphasize that the problem with matching has to do with quantifying performance. With matched designs one can determine if there is some association between the marker and the outcome but one cannot quantify how well the marker performs as a classifier.

We investigated the degree to which results of a matched study can be misleading regarding the perfor-
mance of the biomarker when applied to the general population. We also describe a modified analysis approach that provides corrected estimates. The method requires data on the distributions of risk factors used for matching in the cohort from which the cases and matched controls were drawn. These data items should be readily available when a well-designed case control study is performed (10).

Methods

We generated biomarker and clinical risk factor data using simulation models. In each simulation scenario the true performances of the biomarker alone, the risk factors alone, and the combination of the biomarker and risk factors were quantified with ROC curves. We then generated a matched case control study drawn from the cohort and evaluated biomarker performance with those data. For most of our investigations we used very large sample sizes so that the distortions caused by matching could be calculated exactly without the sampling variability that is inherent to studies with moderate sample sizes. We also conducted analyses with moderate sample sizes.

DATA FOR BIVARIATE BINORMAL SCENARIOS

Let \( x \) denote the baseline risk factor variables, \( y \) the biomarker, and \( D \) the binary outcome, with \( D = 1 \) for a case and \( D = 0 \) for a control. For a population of \( n = 1 \ 000 \ 000 \) subjects we generated \( D \) with probability \( D = 1 \) being 10%. For controls, \((x,y)\) were generated as bivariate normal with means \((0,0)\), SDs \((1,1)\) and correlation \( \rho = 0.1, 0.5, \) or \(0.9\). For cases, \((x,y)\) were bivariate normal with the same SDs and correlation as for controls but with means \((\mu_x,\mu_y)\) where \( \mu_x > 0 \) and \( \mu_y > 0 \). In this scenario, the magnitude of \( \mu_x \) determines the performance of \( x \) alone and \( \mu_y \) determines the performance of \( y \) alone. For example, the area under the ROC curve (AUC) for \( x \) is \( AUC_x = \Phi(\mu_x/\sqrt{2}) \) where \( \Phi \) is the standard normal cumulative distribution function, and the true-positive rate corresponding to the threshold that allows a false-positive rate \( f \), denoted by \( ROC_x(f) \), is \( ROC_x(f) = \Phi[\mu_x + \Phi^{-1}(f)] \). We fixed \( ROC_x(0.2) \) [and \( ROC_x(0.2) \)] and used the corresponding values of \( \mu_x \) (and \( \mu_y \)) to generate data. For example, in Fig. 1 we wanted a setting for which \( ROC(0.2) = 0.64 \) for \( x \) alone and \( y \) alone, so we used values \( \mu_x,= \mu_y = 1.19 \) because the associated value of \( ROC(0.2) = 0.64 \) for \( y \) alone.

DATA FOR PREDICTING CRITICAL ILLNESS

We investigated the impact of matching controls to cases in the context of a study to identify patients in need of critical care services who undergo care in the prehospital setting (e.g., ambulance). We used a validation data set of 57 647 patients from a population-based cohort study of an emergency medical services system in King County, Washington. Methods and results of this study have been published (11). In brief, in this study we developed a prediction model for critical illness in nontrauma non–cardiac-arrest adult patients transported to a hospital by emergency medical services from 2002 through 2006. Critical illness was defined as severe sepsis, delivery of mechanical ventilation, or death during hospitalization. Predictors of risk in the model included older age, male sex, lower systolic blood pressure, abnormal respiratory rate, lower Glasgow Coma Scale score, lower pulse oximetry, and residence in a nursing home. Studies investigating biomarkers for use in this clinical context are ongoing but data are not yet available. We simulated a predictive biomarker informed by characteristics of whole blood lactate, which is a strong candidate biomarker (12–14). Specifically, we generated the marker as a log-normal variable with (mean, SD) equal to (4.0, 2.6) and uncorrelated with risk factors in the 3121 patients with critical illness and equal to (2.5, 2.0) and uncorrelated with risk factors in the 54 526 controls without critical illness. Because of their common association with out-
come, the biomarker and risk factors are correlated in the population that combines cases and controls. The log-transformed marker was used in data analysis (C.W. Seymour, Colin R. Cooke, Zheyu Wang, Kathleen Kerr, Thomas D. Rea, Jeremy M. Kahn, M.S. Pepe, unpublished data).

All 3121 cases who developed critical illness were included in data analyses. We generated a data set of unmatched controls in which 3121 noncritical illness encounters were selected at random from the 54,526 available controls. For the matched controls data set, we selected controls who, on the basis of the clinical predictors alone, would be considered to be at similar risk of critical illness as the cases. Specifically, we adopted the risk score categories defined previously for these data\footnote{11}, collapsing categories $\geq 5$ into a single category. For each case, we selected a control at random from the same risk category as the case. Sampling was done without replacement in all categories, but in the highest category, where the number of controls available ($n = 247$) was less than the number of cases ($n = 261$). In the highest risk category, sampling with replacement was used.

**STANDARD DATA ANALYSES**

We fit standard unconditional logistic regression models to derive linear combinations of risk factor predictors (denoted by $x$) and the biomarker ($y$)\footnote{15}. Unconditional logistic regression is appropriate for the settings considered in this study, in which data were frequency matched on explicit variables that could be included in the logistic model\footnote{16}, chapter 6. ROC curves were estimated empirically from the case control data for $x$ alone, for $y$ alone, and for the linear combination. These analyses were carried out separately for the cases and unmatched controls data set and for the cases and matched controls data set. We

---

**Fig. 1.** ROC curves calculated with data from cases and unmatched controls and from a study in which controls were selected to match cases in regard to clinical risk factor predictors (CP).

Data for $n = 2000$ cases and controls were simulated from bivariate binormal models corresponding to the scenario in Table 1 with $\text{ROC}_x (0.2) = \text{ROC}_y (0.2) = 0.64$, and correlation = 0.5. TPR, true-positive rate; FPR, false-positive rate.
calculated the empirical AUC statistic for each ROC curve in addition to the ROC(0.2) statistic, which is the true-positive rate corresponding to the threshold that allows a false-positive rate of 0.2.

These analysis procedures were carried out for the bivariate binormal and critical illness data sets. In addition, we constructed CIs and performed statistical tests using bootstrap resampling methods for the smaller bivariate binormal data set that included 100 cases and controls using the Stata software package (17).

Bootstrapping involved resampling separately from case and control subgroups and refitting the logistic regression models to derive the combination of x and y in each resampled data set. We calculated P values with the Wald test using the bootstrap SE estimate. We note recent concerns about the validity of P values when testing for performance improvement (18,19) with changes in AUC or other metrics. By refitting the models we greatly improved performance of the P values but further improvements are still warranted (20).

ANALYZING COHORT AND MATCHED CASE CONTROL DATA TOGETHER

Our methods for combining cohort data on the matching factors with the matched case control data to arrive at bias-corrected estimates of ROC curves are described in the Supplemental Data that accompanies the online version of this article at http://www.clinchem.org/content/vol58/issue8. Briefly, because ROC curves for y are derived from the case and control distributions of y, the key task is to estimate the distribution of y in the population of controls, which is distorted in the matched set by selecting controls matched to cases on x. This estimation is accomplished by using the distribution of x for controls in the cohort to reweight the controls in the matched set. The ROC curve for the combination of (x,y) requires the additional step of adjusting the combination derived from the matched set for associations between x and D in the population that are nullified by matching the controls to the cases on x.

Results

BIVARIATE BINORMAL SCENARIOS

For data simulated from a bivariate binormal model, performances of the marker and the clinical risk factor variables are displayed in Fig. 1. The true performances of the marker in the population are shown in the upper panels. These are also the values that are estimated from an unmatched study. We see that use of the marker on its own provides good prediction performance (upper left panel). For example, the AUC is 0.80 and when the threshold is chosen so that 20% of controls are allowed to be biomarker positive, 64% of cases are detected [i.e., ROC(0.2) = 0.64]. In the matched study, however, performance appears to be much less (lower left panel), with AUC = 0.66 and only 40% of cases detected, ROC(0.2) = 0.40. See the first column of Table 1 for a range of other scenarios. Biomarker performance is severely underestimated in the matched study when standard simple estimation methods are used. We see from the first 2 columns of Table 1 that the bias is most severe when the marker is highly correlated with the matching factor (ρ = 0.9) and less so when the correlation is small (ρ = 0.1). Note that, by definition, if the marker is uncorrelated with the matching factor, matching has no effect on the marker distribution of selected controls and no bias results from matching.

The performance of the marker in the population may be partly attributed to its association with risk factors that are themselves associated with the outcome. How much does the marker add to the information already contained in the clinical data? This question concerning the incremental value of the marker is best addressed by comparing the ROC curve for the combination of the marker and clinical risk factor variables with that for the clinical variables alone (21). See the upper right panel of Fig. 1 for an example in which, by adding the marker, we increase the proportion of cases detected by only 6%, from 64% with the clinical predictors alone to 70% when the marker is added to them. The incremental value of the marker is small in the scenario of Fig. 1. In the lower right panel of Fig. 1 we see the ROC curves expected if controls are selected to exactly match the cases in regard to the clinical variables. The clinical variables on their own are completely uninformative in the matched study, a simple consequence of the design in which controls are chosen with the same clinical risk factor values as the cases. We see that the difference between the ROC curve for the marker combined with risk factors and the ROC curve for the clinical risk factors alone is much larger in the matched study than in the population. For example, when we again set thresholds to allow 20% of controls to be positive, the case detection rate with the combination was 44% and when we compared this rate with that for the clinical predictors alone (20%), the improvement seemed to be 24%. This is much larger than the actual increment in performance in the population, namely 6%. The apparent change in the AUC with addition of the marker was 0.183 in the matched study compared with the true value of 0.033. Therefore the apparent increment in performance due to the marker is severely biased by matching, this time toward overestimation of performance. Table 1 shows numerical results for a variety of other scenarios in the Combination vs Risk Factors column. In all the scenarios we studied, increments calculated from matched studies were many times larger than the true increments, rang-
ing from 2 to more than 10 times larger than those observed in unmatched studies. The bias was more severe when the marker was highly correlated with risk factors (Table 1, column 6).

We used very large sample sizes in the simulated studies reported in Table 1 and Fig. 1 to quantify the expected magnitudes of bias. Real studies with moderate samples have the additional issue of variability that may or may not overwhelm the biases. Fig. 2 shows results for a simulated study from the same scenario as Fig. 1 but with only 100 cases and 100 random or matched controls included.

We found that for the biomarker alone (Fig. 2, left panels) with random unmatched controls we estimated AUC = 0.76 [95% CI = (0.69,0.83)] and ROC(0.2) = 0.53 [95% CI = (0.43,0.74)], whereas with matched controls AUC = 0.64 [95% CI = (0.56,0.71)] and ROC(0.2) = 0.35 [95% CI = (0.23,0.49)]. The downward bias in the matched data estimates resulted in CIs that do not contain the true values [AUC = 0.80, ROC(0.2) = 0.64]. This result contrasts with the unmatched study estimates that do contain the true values. Moreover, the CIs for the matched study barely overlap those for the unmatched study.

### Table 1. Performance calculated from simulated bivariate binormal data with unmatched designs that choose controls randomly or with matched designs that choose controls to match cases on risk factors.

<table>
<thead>
<tr>
<th>Controls</th>
<th>Marker</th>
<th>% Bias</th>
<th>Risk factors</th>
<th>Combination</th>
<th>Combination vs risk factors</th>
<th>% Bias</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC(0.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.64</td>
<td>—</td>
<td>0.64</td>
<td>0.78</td>
<td>0.14</td>
<td>—</td>
<td>0.1</td>
</tr>
<tr>
<td>Matched</td>
<td>0.59</td>
<td>−7.8</td>
<td>0.20</td>
<td>0.59</td>
<td>0.39</td>
<td>179.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.64</td>
<td>—</td>
<td>0.64</td>
<td>0.70</td>
<td>0.06</td>
<td>—</td>
<td>0.5</td>
</tr>
<tr>
<td>Matched</td>
<td>0.40</td>
<td>−37.5</td>
<td>0.20</td>
<td>0.44</td>
<td>0.24</td>
<td>300.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.63</td>
<td>—</td>
<td>0.63</td>
<td>0.64</td>
<td>0.007</td>
<td>—</td>
<td>0.9</td>
</tr>
<tr>
<td>Matched</td>
<td>0.24</td>
<td>−61.9</td>
<td>0.20</td>
<td>0.29</td>
<td>0.090</td>
<td>&gt;1000.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.31</td>
<td>—</td>
<td>0.31</td>
<td>0.36</td>
<td>0.05</td>
<td>—</td>
<td>0.1</td>
</tr>
<tr>
<td>Matched</td>
<td>0.30</td>
<td>−3.2</td>
<td>0.20</td>
<td>0.30</td>
<td>0.10</td>
<td>100.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.31</td>
<td>—</td>
<td>0.31</td>
<td>0.33</td>
<td>0.02</td>
<td>—</td>
<td>0.5</td>
</tr>
<tr>
<td>Matched</td>
<td>0.26</td>
<td>−16.1</td>
<td>0.20</td>
<td>0.26</td>
<td>0.06</td>
<td>200.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.31</td>
<td>—</td>
<td>0.31</td>
<td>0.31</td>
<td>0.00</td>
<td>—</td>
<td>0.9</td>
</tr>
<tr>
<td>Matched</td>
<td>0.21</td>
<td>−32.2</td>
<td>0.20</td>
<td>0.22</td>
<td>0.02</td>
<td>&gt;1000.0</td>
<td>0.9</td>
</tr>
<tr>
<td>AUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.801</td>
<td>—</td>
<td>0.800</td>
<td>0.873</td>
<td>0.073</td>
<td>—</td>
<td>0.1</td>
</tr>
<tr>
<td>Matched</td>
<td>0.775</td>
<td>−3.2</td>
<td>0.500</td>
<td>0.776</td>
<td>0.276</td>
<td>278.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.800</td>
<td>—</td>
<td>0.802</td>
<td>0.836</td>
<td>0.033</td>
<td>—</td>
<td>0.5</td>
</tr>
<tr>
<td>Matched</td>
<td>0.660</td>
<td>−17.5</td>
<td>0.500</td>
<td>0.683</td>
<td>0.183</td>
<td>454.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.799</td>
<td>—</td>
<td>0.799</td>
<td>0.805</td>
<td>0.006</td>
<td>—</td>
<td>0.9</td>
</tr>
<tr>
<td>Matched</td>
<td>0.535</td>
<td>−33.0</td>
<td>0.500</td>
<td>0.578</td>
<td>0.078</td>
<td>1200.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.599</td>
<td>−5.95</td>
<td>0.595</td>
<td>0.630</td>
<td>0.035</td>
<td>—</td>
<td>0.1</td>
</tr>
<tr>
<td>Matched</td>
<td>0.589</td>
<td>−1.7</td>
<td>0.500</td>
<td>0.589</td>
<td>0.089</td>
<td>154.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.598</td>
<td>—</td>
<td>0.590</td>
<td>0.609</td>
<td>0.019</td>
<td>—</td>
<td>0.5</td>
</tr>
<tr>
<td>Matched</td>
<td>0.550</td>
<td>−8.0</td>
<td>0.500</td>
<td>0.559</td>
<td>0.059</td>
<td>210.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Unmatched</td>
<td>0.599</td>
<td>—</td>
<td>0.599</td>
<td>0.601</td>
<td>0.002</td>
<td>—</td>
<td>0.9</td>
</tr>
<tr>
<td>Matched</td>
<td>0.509</td>
<td>−15.0</td>
<td>0.500</td>
<td>0.523</td>
<td>0.023</td>
<td>1050.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>

* Shown are proportions of cases detected when the biomarker threshold allows 20% of controls to be positive, i.e., ROC (0.20) are displayed. AUC values are shown in the lower panel.
* 100 × [(matched/unmatched)/unmatched].
* Percentage bias in the performance increment (combination risk factors).
* Correlation between x and y in cases and in controls.
* Unmatched random controls studies give true performance.
Considering the increment in performance gained with the biomarker (Fig. 2, right panels), with unmatched data we found that the change in the AUC was 0.019 and nonsignificant \((P = 0.12)\), whereas with matched data the estimated change in the AUC was much larger and statistically significant, AUC = 0.17 \((P = 0.003)\). Similarly, we found that the change in the proportion of cases detected, ROC(0.2), was only 6\% \((P = 0.36)\) with unmatched data but much inflated, 19\% \((P = 0.08)\) with matched data. Substantially different conclusions are drawn about the magnitude and statistical significance of the marker in the matched and unmatched studies.

**Predictors of Critical Illness**

In the data set concerning predictors of critical illness, the performance of the biomarker alone is not biased by matching. The explanation is that the biomarker was uncorrelated with the matching factors among controls in this particular setting, so matching did not distort the distribution of the marker in study controls. However, the ROC curves for the risk factors and for the combination of marker with risk factors were biased by matching. In particular, the true proportion of cases detected at threshold allowing 20\% of controls positive was 67\% without the marker and 75\% with the marker, a gain of 8\% in the detection rate. In the matched analysis, however, the apparent gain was much larger, 19\% (Table 2, Fig. 3). Similarly, the improvement in the AUC was 0.05 with use of the marker in the population, but the matched study indicates an improvement of 0.14, almost 3 times the true value. Note that the clinical predictors are somewhat predictive in this matched study, in contrast to results for the bivariate binormal setting. The reason is that we used a categorized score for selecting the matched controls here. Because we did not match controls exactly to cases on the values of the clinical risk factors, there was

---

**Fig. 2.** ROC curves calculated with data from cases and unmatched controls and from a study in which controls were selected to match cases in regard to clinical risk factor predictors (CP).

Data for \(n = 100\) cases and controls were simulated from bivariate binormal models corresponding to the scenario in Fig. 1. TPR, true-positive rate; FPR, false-positive rate.
We implemented our proposed methodology for calculating bias-corrected ROC curves from matched studies when cohort data on the matching variable is available. Results are summarized in Table 2. We found that the curves were indistinguishable from those calculated with random unmatched controls (also see online Supplemental Fig. 1).

### COMPARISONS BETWEEN MARKERS

Turning now to possible effects of matching on comparisons between markers we illustrate in the left panels of Fig. 4 an example where, in unmatched data (and in the population), the ROC curve for marker A is higher than for marker B, but the reverse holds when controls are chosen to match cases on risk factors. In a second example (right panels of Fig. 4) the combination of marker A and risk factor variables has a higher ROC curve than that of the combination of marker B and risk factor variables, but again the reverse holds when controls are chosen to match cases.

### Discussion

We have demonstrated here that simple analyses of matched data can give grossly biased impressions of a biomarker’s capacity for prediction or classification. When the biomarker was positively associated with risk factors and with outcome, the apparent performance

<table>
<thead>
<tr>
<th></th>
<th>TPR&lt;sup&gt;b&lt;/sup&gt; at FPR = 0.2</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random controls</td>
<td>Matched controls</td>
</tr>
<tr>
<td>CP</td>
<td>0.667</td>
<td>0.342</td>
</tr>
<tr>
<td>Marker</td>
<td>0.491</td>
<td>0.498</td>
</tr>
<tr>
<td>CP + marker</td>
<td>0.748</td>
<td>0.536</td>
</tr>
<tr>
<td>Combination vs CP</td>
<td>0.081</td>
<td>0.194</td>
</tr>
</tbody>
</table>

* Shown are performance measures calculated with random unmatched controls and with controls that matched to cases on quintile of baseline risk. Also shown are values calculated with the use of data on matching variables in the cohort from which cases and controls were drawn in addition to the matched case–control dataset.

<sup>b</sup> TPR, true-positive result; FPR, false-positive result; CP, clinical predictors.
of the biomarker on its own was generally deflated in a matched study compared with its true performance, and the biomarker’s capacity to improve on existing clinical predictors appeared to be inflated in matched studies. Biases were largest when the biomarker was most strongly correlated with the matching factors. Although the fact that matching leads to biased results has been mentioned previously (9), it is not widely appreciated among researchers and the magnitude of bias has not been investigated.

We focused on 2 commonly used measures of prediction performance: the case detection rate when the marker positivity threshold is fixed by setting a criterion on the false-positive rate and the AUC. Qualitatively similar results were found for other measures of prediction performance (22), including the net reclassification index (23), the integrated discrimination improvement statistic (23), and the improvement in net benefit (24).

We observed (Fig. 4) that these biases have consequences not only for evaluating a single biomarker, but also, potentially, for comparing biomarkers and for selecting biomarker combinations. The fact that matching can distort the rank order of markers for their prediction capacity raises concerns, especially for discovery studies. In particular, discovery studies typically rank biomarkers according to their apparent performance, with the goal of selecting the best-performing markers to go forward for validation. It is possible that we are not selecting markers with the best prediction capacity or the best potential for improving on existing risk factors when we use matched designs and standard analyses. This possibility should be investigated further in real data sets.

We proposed modified analysis methods that correct for biases caused by matching. Additional data on the distribution of the matching factors in the popula-
tation are required for these analyses. Such data should be available when good principles of study design are employed in constructing the case study control. For a good case control design, a cohort representing the clinical setting of interest is identified and cases and controls are selected from that cohort (10, 25). For matched case control designs, one selects controls from that cohort having the same matching variable values as the corresponding cases. Therefore the supplementary data that are needed for the modified analysis, namely matching variable data on all subjects in the cohort, should be readily available if the matched study is conducted appropriately. However, many studies are done by selecting controls from a different population than the cases, and consequently the distribution of the risk factors in that population is not relevant or useful for correcting the bias. We and others have noted that studies that employ controls from a different population or clinical setting from which the cases are drawn are subject to many biases (10). Our results here indicate that matching on known risk factors is not a means to eliminate bias in gauging the magnitude of a biomarker’s performance or the magnitude of its potential for improving performance over those risk factors.

Our results are directed at estimation of the magnitude of a biomarker’s prediction performance. We have shown that matching can lead to bias in the magnitude of estimated prediction performance when standard analysis methods are used. However, matching is not problematic when the focus is on the narrower question of assessing associations. In particular, matching is not a problem if one simply wants to determine if any association exists between the marker and outcome that is not due to risk factors. Rather, matching is a problem if the intent is to examine performance within subpopulations defined by the risk factors. In fact, matching is known to be an efficient design strategy for evaluating subpopulation-specific ROC curves (9). Furthermore, standard methods for estimating risk factor–adjusted odds ratios for a biomarker applied to matched data and, again, matching is an efficient design for this purpose (26). However, our concern is not with simply estimating associations. We are interested in estimating prediction performance metrics that depend on the distributions of the risk factors and biomarkers in the population in addition to their associations with outcome. We have shown here that matching invalidates the usual analysis methods for estimating prediction performance measures.

Although matching invalidates estimates derived with standard methods, we have developed bias-correcting methods to estimate prediction performance measures if cohort data on matching factors are available. Therefore, when cohort data on matching factors are available, we have the choice to do an unmatched design or a matched design. An advantage of the unmatched design is that standard analysis methods can be applied. However, this approach may not be as efficient in its use of data as a strategy that matches controls to cases and derives estimates with the non-standard bias-correcting methods. Preliminary results from simulation studies suggest in fact that the latter strategy may be optimal (21). Further theoretical work and simulation modeling will be needed before this approach is recommended.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: M.S. Pepe, NIH/NIGMS GM054438.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References