Published in final edited form as:
J Proteome Res. 2011 December 2; 10(12): 5562–5567. doi:10.1021/pr200507b.

Accounting for control mislabelling in case-control biomarker studies

Mattias Rantalainen[†] and **Chris C. Holmes**[*,‡]
Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom, and Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom, tel: +44(0)1865 285386, fax: +44(0)1865 285384

Abstract

In biomarker discovery studies uncertainty associated with case and control labels are often overlooked. By omitting to take into account label uncertainty, model parameters and the predictive risk can become biased, sometimes severely. The most common situation is when the control set contains an unknown number of undiagnosed, or future, cases. This has a marked impact in situations where the model needs to be well calibrated, e.g. when the prediction performance of a biomarker panel is evaluated. Failing to account for class label uncertainty may lead to underestimation of classification performance and bias in parameter estimates. This can further impact on meta-analysis for combining evidence from multiple studies. Using a simulation study we outline how conventional statistical models can be modified to address class label uncertainty leading to well-calibrated prediction performance estimates and reduced bias in meta-analysis. We focus on the problem of mislabelled control subjects in case-control studies, i.e. when some of the control subjects are undiagnosed cases although the procedures we report are generic. The uncertainty in control status is a particular situation common in biomarker discovery studies in the context of genomic and molecular epidemiology, where control subjects are commonly sampled from the general population with an established expected disease incidence rate.

Introduction

Case-control study designs[1] provide a powerful approach to elicit predictive biomarkers. In such studies, biomarkers[2] are measured retrospectively within affected individuals (‘cases’) and the patterns of measurements are compared to those obtained from a set of unaffected individuals (‘controls’). It is well known that in order to maximise statistical power the control set should be chosen to be as similar in key variables, such as age and gender, to that of the case set.[1] However, a widespread but often overlooked problem arise when there is uncertainty present in the control labels, so that some of the subjects labelled as controls are in fact cases.[3] When there is a risk of mislabelling in a case-control study, fitting statistical models without taking account of the uncertainty in control status will result in downward bias in estimates of biomarker effect sizes and the resulting model will underestimate the

[*]To whom correspondence should be addressed: cholmes@stats.ox.ac.uk.
[†]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom
[‡]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom, tel: +44(0)1865 285386, fax: +44(0)1865 285384

Europe PMC Funders Author Manuscripts

true predictive risk for an 'at risk' individual. Both of these features are undesirable, the latter undermining confidence in the true effectiveness of the biomarker panel to discriminate those at risk.

A common cause for case-control mislabelling is when the sample of control subjects contain undiagnosed cases.[4] Ideally such mislabelling should not occur in case-control studies, however, in reality mislabelling can occur for several reasons including, low sensitivity of a diagnostic test, uncertainty in determining the trait defining disease, or if the control set is based on a population sample with an intrinsic expected (undiagnosed) disease incidence rate. Low sensitivity may be the result of a sub-optimal diagnostic test or when the 'gold-standard' test is too invasive to be utilized on control subjects, for example if a biopsy is required for 'gold-standard' diagnosis. In such circumstances control subjects might instead be diagnosed by an alternative less invasive test, with lower sensitivity. Uncertainty in diagnosis may also arise due to subjective scoring of patients based on phenotypical evidence or when case and control assignment are made by dichotomizing a continuous trait.[5] When a population based sample is used as the control set it is expected that a proportion of mislabelled control subjects are present corresponding to the population based incidence level for the disease, this is a type of study design common in genome-wide association studies (GWAS),[6] but also in biomarker studies based on samples from biobanks.[7]

In studies where a 'gold-standard' diagnostic test is available and used, there will be no uncertainty in the case-control labels except if there is a risk of future (prospective) mislabelling. Prospective mislabelling should ideally be accounted for when developing a model with aim to predict subjects at risk of future disease. However, in the case of prospective cohort-based studies, where subjects are followed over time and a 'gold-standard' diagnostic test is available, alternative analyses strategies to the case-control study design can be used with benefit, such as time-to-event analysis.[8]

In this note we discuss how to formally account for uncertainty in the status of controls. The outlined methodology is applicable to case-control studies in general, including studies utilising 'omics' technologies such as proteomics,[9] metabonomics[10] and transcriptomics[11] for biomarker discovery. Our recommendation reduces bias in estimates and improves accurate assessment in the overall effectiveness and utility of the biomarker panel. We demonstrate how such adjustments can be easily incorporated within standard statistical model fitting algorithms. The general problem of class mislabelling within classification models has a considerable literature.[4,12] However, we are the first to concentrate specifically on case-control study designs in the context of biomarker discovery and the issues relating to contamination of cases within the control group. To our knowledge we are also the first to assess the impact of label uncertainty on prediction performance as described by the commonly used Receiver Operating Characteristic (ROC) curve, and the impact on meta-analysis estimates when the level of label uncertainty varies between studies.

Throughout this note we show results from logistic regression although the procedures we adopt are generic to other standard classification models. A software implementation for R[13] of the logistic regression model accounting for label uncertainty is available for free under GPL v3 license (https://sourceforge.net/projects/labelu/files/).

Methods

In this section vectors are indicated by lower-case bold font, matrices by upper-case bold font and scalars in italic. We suppose that we have measured a set of k biomarkers on n individuals and we use x_{ij} to denote the jth biomarker measurement on the ith individual, $j =$

$1,\ldots,k$, $i = 1,\ldots,n$ and we use $x_i = \{x_{i1},\ldots,x_{ik}\}$ to denote the panel of biomarker measurements on individual i. Each individual is assigned a class label $y_i = 1$ or $y_i = 0$ depending on their case or control status respectively. The problem can be formulized as follows, let z_i denote the *true*, possibly future, case status of individual i. Then $Pr(z_i = 1|y_i = 1) = 1$ whereas $Pr(z_i = 0|y_i = 0) < 1$, in words, the probability of being a true case given you are labeled a case is 1 but the probability of being a true control ('never case') given you are labeled a control is less than 1, equivalently $Pr(z_i = 1|y_i = 0) > 0$. Clearly in biomarker discovery we wish to estimate $Pr(z_i = 1|x_i)$, the true risk of disease given a biomarker measurement, but if instead we use a model of $Pr(y_i = 1|x_i)$ then the predictive risk will be biased downwards. Intuitively this can be seen by imagining a situation where you take some of your cases and artificially re-label them as controls and then re-fit your predictive model. This will clearly underestimate risk. In a sense this is the situation we are dealing with when the ascertainment of controls is uncertain.

Let P represent the re-labeling probability, for example the background prevalence of the disease within the population. When P is known efficient model fitting procedures using the Expectation-Maximization (EM) algorithm exist,[14] and can be used within a variety of standard statistical models such as logistic regression,[15] naive Bayes,[16] or Linear Discriminant Analysis.[17] The EM algorithm works by iteratively re-fitting the model to re-weighted samples, where at each iteration each control individual is given a weight in proportion to the evidence that they are in fact a case; defined as their predicted case-control status times their prior case-control status, $\propto Pr(z_i = 1/y_i = 0, x_i) \times P_i$. The probability status of each sample is then used to re-weight the model fit. This process repeats until the weights and estimates of effects converge after which we have an individual probability $\widehat{P}_i$, the posterior probability that the ith individual is a case given the information in the data and the prior probability P. Note that when we have no uncertainty in the control label we have $P_i = 0$. In this paper we show results from the commonly used logistic regression model with label uncertainty fitted by EM. The procedure for fitting the logistic regression model with label uncertainty using EM is outlined in Algorithm 1. The labels are treated as the unknown variables, where $\mathbf{y}$ is the vector of class labels of dimension $[n \times 1]$ and $\mathbf{X}$ is the matrix of descriptor variables of dimension $[n \times k]$. Here we initialize the coefficients $\beta = 0$, but β could also be initialized using a preliminary estimate using e.g. conventional logistic regression. We recommend choosing P based on available population incidence data, for example if 5% of the population is expected to have the disease P would be set to 0.05.

Bias in predicted risk, $Pr(y_i = 1|x_i)$, due to mislabelling will not directly impact the ROC curve as it depends on the rank of the predictions. However, the estimated optimal classification boundary will be miscalibrated. Such miscalibration of the decision boundary will have an impact on future predictions and in evaluation of prediction performance using an external test data set or crossvalidation, resulting in a negative bias in the ROC curve. In both the case of cross-validation and evaluation of an external test data set we also expect mislabelled observations in the test set, which will directly impact upon the classification performance evaluation. To account for mislabelling in test set observations when evaluating the prediction performance, the test set class labels can be updated by repeated draws from the Bernoulli distribution using the posterior predictive risk, $Pr\left(y_i=1|x_i,\widehat{\beta}\right) \times P_i$, and the ROC curved averaged over these samples. By updating the test set class label we allow label uncertainty to be taken into account when evaluating the prediction performance, thus reducing (downward) bias in the ROC curve that would otherwise be present.

Results

Bias in predicted risk

The first scenario we consider is bias in predicted risk. Figure 1 show the effect on fitting a logistic regression model under four modelling scenarios: (A) using the true (unobtainable) labels to provide a reference benchmark; (B) if we was to simply ignore the mislabelling; (C) if we account for mislabelling using a non-Bayesian approach; (D) if we use a Bayesian approach incorporating a prior probability distribution on model coefficients. Data were generated from two normal densities with a shift in mean and with the same variance (~ $N(\mu = 0, \sigma^2 = 1)$ (controls) and ~ $N(\mu = 2.56, \sigma^2 = 1)$ (cases)) for the true cases and controls, and where each control subject has a 10% probability of coming from the case conditional distribution. In each one of 10000 simulation rounds 100 samples from the control conditional distribution and 100 samples from the case conditional distribution was drawn. The class separation corresponds to a 10% misclassification rate under the true model and with no mislabelling. Figure 1 shows the true risk (x-axis) versus the predicted risk (y-axis) with 95% confidence intervals obtained by repeated resampling, Figure 1A illustrate the 'oracle' situation with no mislabelling. If we simply ignore the mislabelling, we can observe a clear downward bias in Figure 1B. For example, we can see that an individual who is at 80% risk of disease has an expected predicted risk of around 60% but could be as low as 50%.

It is underappreciated that standard statistical model fitting can address this problem and accommodate control mislabelling so long as the prior probability of a control becoming a case is known with reasonable precision. In many situations of biomarker discovery studies, the proportion of controls that may be (future) cases is expected to be well known; for example if controls are sampled from a normal human population (population-wide disease risk being known), or in the case of matched controls (e.g. gender, age, body mass index) the conditional disease risk is also likely to be well established. Figure 1C shows the same plot as for the data in Figure 1B but when using the EM correction. We can clearly see that the bias has been removed.

While the EM algorithm removes bias it can give rise to an increased variance of $Pr(z_i = 1|y_i = 1)$, which is partly due to instability in the EM algorithm under some conditions, resulting in overly large (or small) estimates. By placing a suitable prior distribution on the logistic regression coefficients, increased stability of the EM results can be achieved. In many biomarker studies the range of realistic effect sizes is usually known, allowing us to specify a suitable prior distribution on the model coefficients. For example, if the measurement x is standardised to unit variance, it would be very rare for a biomarker to exhibit an effect greater than 2 on the log-odds scale, i.e. a 7-fold change in risk for a unit change in x. Here we place a Gaussian prior with $\mu = 0$ and $\sigma^2 = 2$ on the logistic regression coefficients ($\sigma^2 = 10^5$ for the constant term). Figure 1D show how the bias is eliminated, while the overly wide confidence intervals observed for the standard EM algorithm (Figure 1C) has also been reduced. In corresponding simulations with N=20 and N=50 observations in each class (see Supplementary Online Material) we also observed consistently tighter confidence intervals for the Bayesian model compared to the EM model.

Model calibration and prediction performance estimates

The second scenario we consider is the evaluation of model prediction performance. Receiver operating characteristic (ROC) curves are a common method to visualize and characterize the performance of a classifier via the trade-off between sensitivity (true positive rate) and specificity (true negative rate). In biomarker discovery studies, classification performance of a biomarker panel in the discovery phase is often central to

decision making regarding further investment in research and development of the biomarker panel. In particular when the objective is to discover biomarker panels with clinical applications, for example in clinical diagnostics, the classification performance of the model needs to have sufficiently high sensitivity and specificity to be viable. Class label uncertainty in data may, however, lead to underestimation of the classification performance, and may therefore subsequently lead to a premature termination of research and development efforts.

By allowing for label uncertainty also when evaluating the prediction performance we can reduce (downward) bias in the ROC curve. Figure 2A shows that the estimated ROC curve from the model incorporating label uncertainty and where the label uncertainty is also taken into account in test data set when estimating the ROC curve. It is clear that ROC curve is almost perfectly overlaid with the 'oracle' result (no label uncertainty and a conventional logistic regression model), indicating that by accounting for label uncertainty in the model and when estimating the ROC curve we can reduce downward bias in the ROC curve. Figure 2B shows the estimated ROC curve when ignoring control mislabelling, clearly indicating a substantially reduced estimate of prediction performance compare to the 'oracle' situation. This data set was simulated to provide 0.9 sensitivity and specificity in the case of no mislabelling, and we note that the estimated sensitivity at 0.9 specificity, when ignoring mislabelling, is here estimated as low as ~ 0.7. Corresponding results from simulations with N=20 and N=50 observations in each class can be found in the Supplementary Online Material.

Meta-analysis

The third scenario we consider for which label uncertainty may have adverse effects is meta-analysis. Meta-analysis is increasingly applied in the area of genomic epidemiology and genetics to aggregate information across multiple studies. The typical example, which we consider, is meta-analysis of parameter estimates across two or more studies. Bias in parameter estimates due to case-control mislabelling will bias meta-analysis estimates, particularly if the levels of label uncertainty varies between the included studies. An example of when this can occur is when results from a highly curated study with no mislabelling is combined with a second study where samples are sourced from the general population where a known proportion of mislabelled control individuals is expected, i.e. undiagnosed disease cases. In this instance the latter model's regression coefficients will be biased towards zero, consequently meta-analysis estimates across the two studies will also tend to a smaller estimate, leading one to falsely conclude that the biomarker has less of an effect than it does. In Figure 3 results are shown from two simulated data sets (10000 simulation rounds) of the same size, simulated from the same distribution as in previous sections, with the only difference that data set $D1$ has 10% mislabelled controls while $D2$ does not have any mislabelling. In the top panel (A) the parameter estimate is displayed for the scenario where we know the true class labels (i.e. the unobtainable 'oracle' model), using a conventional logistic regression model. In the middle panel (B) we observe that the effect size estimate for $D1$ data set is negatively biased due to mislabelling, as is the meta-analysis estimate. The lower panel (C) displays the results from a model allowing for label uncertainty resulting in eliminating of downward bias.

Conclusions

To summarize we have shown that accounting for, possibly prospective, control mislabelling in case-control studies can remove substantial downward bias in estimates of effect sizes and hence predictive risk. We have demonstrated how accounting for label uncertainty in prediction performance, particularly ROC curves, can reduce downward bias otherwise observed under mislabelling. Accurate prediction performance estimates are of uttermost

Europe PMC Funders Author Manuscripts

importance in many applications, including biomarker discovery studies using 'omics' platforms. Underestimation of prediction performance may lead to premature termination of further validation and development of biomarker panels not passing predefined performance thresholds required, for example in the case of clinical biomarker panels. We also note that accurately taking into account label uncertainty is important in meta-analysis across studies with potentially different levels of mislabelling.

Control mislabelling is a problem arising mainly under two conditions, when the control subjects are sampled from the general population (e.g. biobanks and cohort studies) and when there is uncertainty in diagnosis due to e.g. low sensitivity in available diagnostic tests.[18] Currently the capacity to profile large sets of samples using omics technologies is increasing through technological developments while large sample collections are also becoming increasingly available through biobank efforts.[7,19] These general trends make us believe that we will likely have an increase in population based biomarker studies over time, where uncertainty in case-control labels should ideally be accounted for in order to maximize the outcome of such studies.

Another interesting aspect not covered here but which can impact the mislabelling rate is when the predictive interval of the model is greater than the age-profile of the sample. In this situation the control set may contain future prospective cases. The ideal scenario would be to follow the controls over time, in a prospective manner, or to revalidate the individual control status at a series of intervals. In the instance of cohort based studies as well as biobank based studies, where sample collections may be used in several different sub-studies, relabelling which requires diagnostic tests may not be possible to carry out arbitrarily frequently for practical reasons. In such instances it would be desirable to have a revalidation frequency proportional to the the expected probability that a control was in fact a case, so that in the studies of more common disease, the relabelling frequency is higher than for rare disease, and for controls with higher predicted risk are re-validated more frequently than those of predicted low risk.

A related problem to that of label uncertainty is addressed in the field of semi-supervised learning, where data include labelled as well as unlabelled observations. Semi-supervised learning has previously been applied for classification problems with DNA microarray data[20] as well as for learning biological networks.[21] Detection of mislabelled observations can also be handled in a separate step prior to the statistical analysis.[22]

As part of this study we also investigated the impact on statistical power to discover predictive biomarkers. Interestingly, detecting which biomarkers have no predictive effect is not impacted by mislabelling, due to a bias-variance trade off. But as we have shown, once a biomarker has been flagged as having a non-zero effect, accurately estimating how predictive it is can be strongly influenced by mislabelling.

We conclude that the benefits of accurately accounting for label uncertainty is substantial in biomarker discovery studies, and include reduced bias in model coefficients, in predicted risk, in prediction performance evaluation and in meta-analysis. As such we believe these methods are important to researchers working in biomarker discovery and deserve to become common best practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

M.R. is supported by a MRC biomedical informatics fellowship (Medical Research Council, fellowship G0802460).

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

References

[1]. Schulz KF, Grimes DA. Case-control studies: research in reverse. Lancet. 2002; 359:431–434. [PubMed: 11844534]

[2]. Ptolemy AS, Rifai N. What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. Scand J Clin Lab Invest Suppl. 2010; 242:6–14. [PubMed: 20515269]

[3]. (a) Szatmari P, Jones MB. Effects of misclassification on estimates of relative risk in family history studies. Genet Epidemiol. 1999; 16:368–381. [PubMed: 10207718] (b) Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol. 1977; 105:488–495. [PubMed: 871121]

[4]. Yasui Y, Pepe M, Hsu L, Adam B, Feng Z. Partially supervised learning using an EM-boosting algorithm. Biometrics. 2004; 60:199–206. [PubMed: 15032790]

[5]. (a) Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. AJNR Am J Neuroradiol. 2011; 32:437–440. [PubMed: 21330400] (b) Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006; 25:127–141. [PubMed: 16217841]

[6]. Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]

[7]. Hewitt R, Hainaut P. Biobanking in a fast moving world: an international perspective. J Natl Cancer Inst Monogr. 2011; 2011:50–51. [PubMed: 21672898]

[8]. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003; 89:232–238. [PubMed: 12865907]

[9]. Blackstock WP, Weir MP. Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol. 1999; 17:121–127. [PubMed: 10189717]

[10]. Nicholson JK, Lindon JC, Holmes E. ‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999; 29:1181–1189. [PubMed: 10598751]

[11]. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]

[12]. (a) Copas J. Binary regression models for contaminated data. Journal of the Royal Statistical Society. Series B. 1988; 50(2):225–265.(b) Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. American Journal of Epidemiology. 1997; 146:195–203. [PubMed: 9230782]

[13]. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2010. ISBN 3-900051-07-0

[14]. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B. 1977; 39(1):1–38.

[15]. Casella, G.; Berger, RL. Statistical inference. Duxbury Press; 2002. p. 660

[16]. Hand D, Yu K. Idiot’s Bayes: Not So Stupid after All? International Statistical Review / Revue Internationale de Statistique. 2001; 69:385–398.

[17]. Duda, RO.; Hart, PE.; Stork, DG. Pattern Classification. 2nd Edition. Wiley-Interscience; 2001. 2nd ed.

[18]. Joseph S, Robbins K, Zhang W, Rekaya R. Effects of misdiagnosis in input data on the identification of differential expression genes in incipient Alzheimer patients. In Silico Biol. 2008; 8:545–554. [PubMed: 19374137]

Europe PMC Funders Author Manuscripts

[19]. Riegman PHJ, Morente MM, Betsou F, de Blasio P, Geary P. on Biobanking for Biomedical Research, M. A. I. W. G. Biobanking for better healthcare. Mol Oncol. 2008; 2:213–222. [PubMed: 19383342]

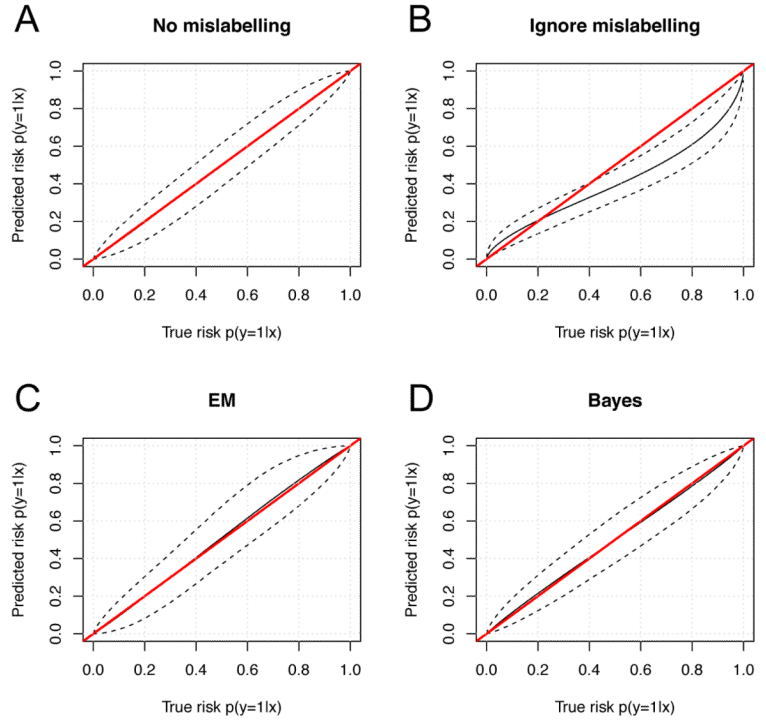
[20]. Harris C, Ghaffari N. Biomarker discovery across annotated and unannotated microarray datasets using semi-supervised learning. BMC Genomics. 2008; 9(Suppl 2):S7. [PubMed: 18831798]

[21]. (a) Hwang T, Sicotte H, Tian Z, Wu B, Kocher J-P, Wigle DA, Kumar V, Kuang R. Robust and efficient identification of biomarkers by classifying features on graphs. Bioinformatics. 2008; 24:2023–2029. [PubMed: 18653521] (b) Kashima H, Yamanishi Y, Kato T, Sugiyama M, Tsuda K. Simultaneous inference of biological networks of multiple species from genome-wide data and evolutionary information: a semi-supervised approach. Bioinformatics. 2009; 25:2962–2968. [PubMed: 19689962]

[22]. Malossini A, Blanzieri E, Ng RT. Detecting potential labeling errors in microarrays by data perturbation. Bioinformatics. 2006; 22:2114–2121. [PubMed: 16820424]

Europe PMC Funders Author Manuscripts

Synopsis



Uncertainty associated with case and control labels are often overlooked in biomarker discovery studies. By omitting to take into account label uncertainty, model parameters and the predictive risk can become biased. We focus on the problem of mislabelled control subjects in case-control studies and outline how conventional statistical models can be modified to address class label uncertainty, leading to well-calibrated prediction performance estimates and reduced bias in meta-analysis.

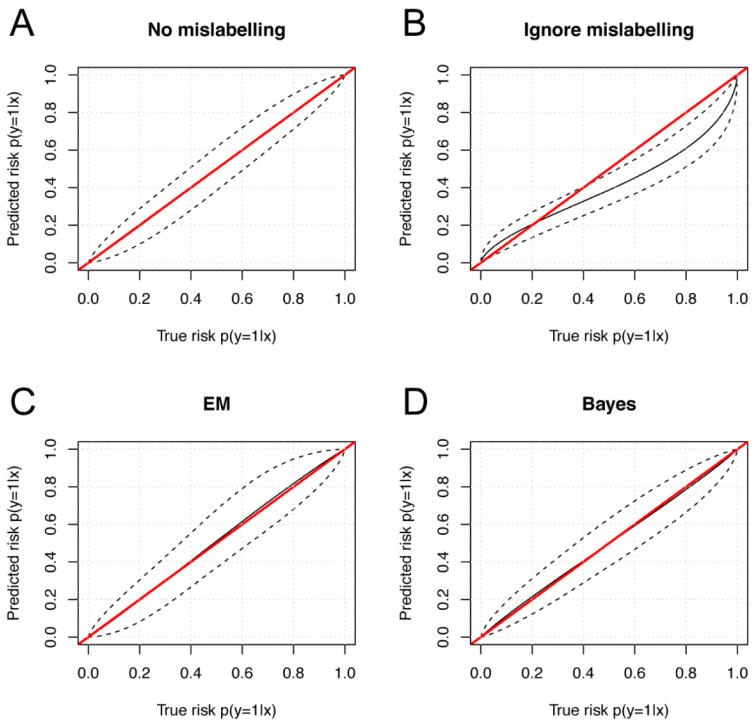


Figure 1.

Bias in predicted risk under 10% mislabelling. Each panel show the predicted risk (median (black line) and 95% confidence interval (dashed line) from resampling (10000 samples)) plotted against the true risk. A) Conventional model and data with no mislabelling. B) Conventional model and data with 10% mislabelling of controls. C) EM model with 10% mislabelling of controls D) Bayesian model with 10% mislabelling of controls.

Europe PMC Funders Author Manuscripts

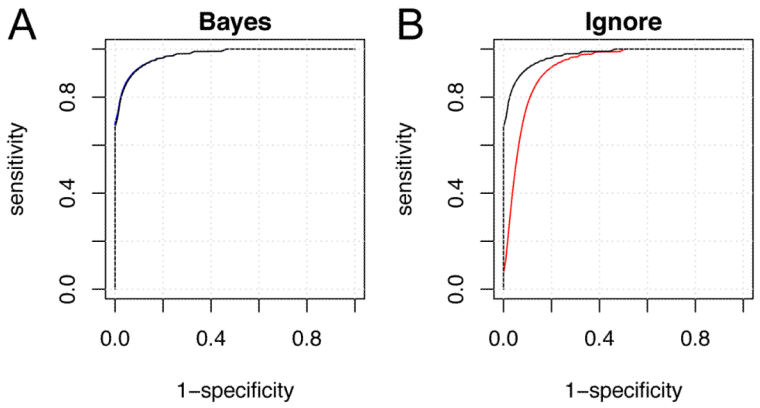


Figure 2.
The impact of class label uncertainty on prediction performance evaluation. The ROC curve represent the median across 10000 simulation rounds. A) ROC curve for a model incorporating label uncertainty (blue) and ROC curve estimated from a data set with no mislabelling (black). B) ROC curve for a conventional model (red) and ROC curve estimated from a data set with no mislabelling (black).

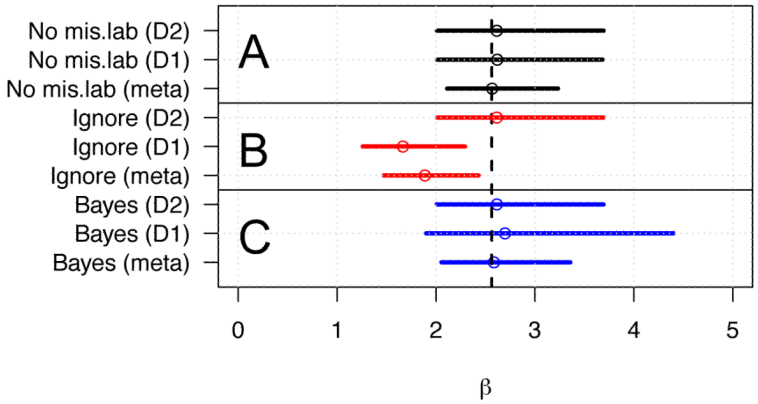


Figure 3.

The impact of class label uncertainty on meta-analysis. Parameter estimates for data set 1 (D1) with 10% mislabelling, data set 2 (D2) having no mislabelling, and a meta-analysis of D1 and D2. In this example the Bayesian model (C) has a relatively flat prior distribution ($N(\mu = 0, \sigma^2 = 10^5)$) on the coefficients to allow for comparison with the other two models. A) Estimates from a conventional logistic regression model for meta-analysis of D1 and D2 with no mislabelling in either data set. B) Parameter estimates and meta-analysis estimates for a conventional logistic regression model with 10% mislabelled data in D1. C) Parameter estimates and meta-analysis estimates for the Bayesian logistic regression model accounting for control mislabelling and with 10% mislabelled data in D1. The circles indicate the median parameter estimate from 10000 simulations and the error bars show the 95% confidence interval.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Algorithm 1

Algorithm 1 Logistic regression with label uncertainty using EM

Augment $\mathbf{X}$ for each $\mathbf{x}_i$ with label uncertainty (i.e. $P_i \neq 0$)
Let $\pi_i = 1 - P_i$ for $y_i = 0$ (controls) and $\pi_i = 1$ for $y_i = 1$ (cases)
Initialize coefficients β (e.g. as zero vector)
$\mathbf{X}^* = [\mathbf{X}; \mathbf{I}_k]$ {Add pseudo data accounting for β prior}
while $\varepsilon > \varepsilon_{lim}$ **do**
  $\mathbf{p} = \frac{e^{\mathbf{X}\beta^j}}{1+e^{\mathbf{X}\beta^j}}$
  $\mathbf{w} = \mathbf{p}(1-\mathbf{p})$
  $\mathbf{w}_{EM} = \mathbf{y}\frac{\mathbf{p}*\pi}{\mathbf{p}*\pi+(1-\mathbf{p})*(1-\pi)} + (1-\mathbf{y})\frac{(1-\mathbf{p})*\pi}{(1-\mathbf{p})*\pi+\mathbf{p}*(1-\pi)}$
  $\mathbf{z} = \mathbf{X}\beta^j + \mathbf{w}^{-1} * (\mathbf{y} - \mathbf{p})$ {Adjusted response}
  $\mathbf{z}^* = [\mathbf{z}; \beta_0]$ {Add pseudo data accounting for β prior}
  $\mathbf{w}^* = [\mathbf{w}, \sigma_\beta^{-2}]$ {Add pseudo data accounting for β prior}
  $\mathbf{w}^*_{EM} = [\mathbf{w}_{EM}, \mathbf{I}_k]$ {Add pseudo data accounting for β prior}
  Let $\mathbf{W}$ be the diagonal matrix with $\sqrt{\mathbf{w}^*} \times \sqrt{\mathbf{w}^*_{EM}}$ as diagonal
  $\mathbf{X}_{EM} = \mathbf{W}\mathbf{X}^*$ {Update $\mathbf{X}^*$ by weights $\mathbf{W}$}
  $\hat{\beta^{j+1}} = (\mathbf{X}'_{EM}\mathbf{X}_{EM})^{-1}\mathbf{X}'_{EM}\mathbf{W}\mathbf{z}^*$
  $\varepsilon = |\hat{\beta^{j+1}} - \hat{\beta}^j|^2 / max(var(\mathbf{X}))$
end while