Journal Title
XX(X):1–10

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Sample size requirements for learning to classify with high-dimensional biomarker panels

Paul McKeigue

Abstract

A common problem in biomedical research is to calculate the sample size required to learn a classifier using a (possibly high-dimensional) panel of biomarkers. This paper describes a simple method, based on a gaussian approximation for calculating the predictive performance of the learned classifier given the size of the biomarker panel, the size of the training sample, and the optimal predictive performance (expressed as C-statistic C_{opt}) of the biomarker panel that could be obtained if a training sample of unlimited size were available. Under the assumption that the biomarker effect sizes have the same correlation structure as the biomarkers, the required sample size does not depend upon these correlations, but only upon C_{opt} and upon the sparsity of the distribution of effect sizes, defined by the proportion of biomarkers that have nonzero effects . To learn a classifier that extracts 80% of the predictive information, the required case sample size varies from about 0.1 cases per variable for a panel with $C_{\text{opt}} = 0.9$ and a sparse distribution of effect sizes (such that 1% of biomarkers have nonzero effect sizes) to nine cases per variable for a panel with $C_{\text{opt}} = 0.75$ and a diffuse distribution of effect sizes.

Keywords

Sample size, linear classifier, Bayesian, high-dimensional

Usher Institute of Population Health Sciences and Informatics, University of Edinburgh

Corresponding author:
Paul McKeigue, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Old Medical School, Teviot Place, Edinburgh EH8 9AG, UK
Email: paul.mckeigue@ed.ac.uk

Prepared using sagej.cls *[Version: 2016/06/24 v1.10]*

Introduction

A key objective in many areas of biomedical research is to learn to predict clinical outcomes such as disease risk or drug response using panels of biomarkers. Platforms that can assay thousands of proteins, gene transcripts, or metabolites are now available. A new era is heralded with the emergence of "precision medicine, an innovative approach to disease prevention and treatment that takes into account individual differences in people's genes, environments, and lifestyles." (1). Statisticians are often asked to estimate required sample sizes for such studies. Classical methods for calculating statistical power to reject a null hypothesis at a specified level of significance are not relevant to this problem. Methods described so far for estimating sample size for learning to classify from high-dimensional biomarker platforms require preliminary data on the covariance structure of the biomarkers (2).

This article describes a method for calculating the sample size required to learn to classify from a high-dimensional biomarker panel that, given some simplifying assumptions, does not require the availability of preliminary data. The researcher has only to specify the number P of biomarkers on the panel, a plausible value for the proportion ϕ of these biomarkers that have nonzero effects, and a plausible value for the optimal predictive performance of the biomarker panel that could be learned from a sample of unlimited size.

Methods and results

For simplicity the terms case and control are used though the argument applies to any dichotomous outcome. A commonly-used design for biomarker studies is a nested case-control design, in which a cohort of individuals with consent for storage of tissue samples is established at baseline, and the members of this cohort are then followed up to ascertain cases of an adverse outcome such as incident disease, recurrence of disease, or non-response to drug therapy. These cases are then compared with controls from the same cohort, using biomarkers measured on the stored tissue samples.

A measure widely used to evaluate the performance of a diagnostic classifier is the area under the receiver operating characteristic curve or C-statistic, which can be defined as the probability of correctly classifying a case-control pair. In a Bayesian framework, or more generally according to the likelihood principle that is common to both Bayesian and frequentist approaches, this classification should be based on the ratio of the likelihoods of the two possible assignments of case-control status, given the data. Denoting the two alternative assignments of case and control status of such a pair as hypotheses $(\mathcal{H}_1, \mathcal{H}_2)$, C is the probability that when $\mathcal{H}_1$ is true the log Bayes factor favouring $\mathcal{H}_1$ over $\mathcal{H}_2$ is greater than zero.

Among a series of results attributed to Turing (3) are that:

1. the sampling distribution of the log Bayes factor based on a sum of many independent observations is asymptotically gaussian with (when natural logarithms are used) variance twice its expectation

2. the expectation under hypothesis $\mathcal{H}_2$ of the log Bayes factor favouring $\mathcal{H}_1$ over $\mathcal{H}_2$ is asymptotically minus 1 times its expectation under $\mathcal{H}_1$

It follows that if the log Bayes factor favouring $\mathcal{H}_1$ over $\mathcal{H}_2$ is computed from many biomarkers, its asymptotic sampling distribution (under $\mathcal{H}_1$) has mean 2Λ and standard deviation $2\sqrt{\Lambda}$. where Λ is the expected log Bayes factor favouring case over control status in a case, or control over case status in a control. Thus $C = 1 - \Phi\left(-\sqrt{\Lambda}\right)$ where Φ is the standard gaussian cumulative distribution function. The C-statistic can thus be interpreted as a mapping of the expected log Bayes factor Λ, which as a Kullback-Leibler divergence can take non-negative real values, to the interval from 0.5 to 1. As Λ can be interpreted as the mean information for discriminating in favour of true (case or control) status, this gives an information theoretic interpretation to the C-statistic.

Gaussian distribution of effect sizes

The argument below adapts an approach described previously for inferring the presence of an individual in a case sample (4). The approach described in that paper is equivalent to evaluating the ability to predict the case-control status of an individual who was in the training sample. This article however is concerned with the ability to predict the case-control status of a new individual. For simplicity, the argument below is developed for a study design with equal numbers of cases and controls, although it can be extended to any dichotomous outcome. We assume that the likelihoods can be approximated by the following simple model, in which the correlations between the biomarker effect sizes are the same as the class-conditional correlations $\mathbf{\Sigma}$ between the P biomarkers:-

- The population means $\boldsymbol{\beta}$ of the case-control differences are distributed as $\mathcal{N}\left(0, \frac{1}{J}\mathbf{\Sigma}\right)$, where J is the precision (inverse variance) of the univariate distributions.
- The biomarker vector of a control individual is distributed as $\mathcal{N}\left(\boldsymbol{\eta} - \frac{1}{2}\boldsymbol{\beta}, \mathbf{\Sigma}\right)$.
- The biomarker vector of a case individual is distributed as $\mathcal{N}\left(\boldsymbol{\eta} + \frac{1}{2}\boldsymbol{\beta}, \mathbf{\Sigma}\right)$

Without loss of generality we specify that the population means $\boldsymbol{\eta}$ (over equal proportions of cases and controls) of the P biomarkers are distributed as $\mathcal{N}\left(0, \frac{1}{K}\right)$, where K is the precision of the univariate distributions. Write:-

- $\boldsymbol{x_{0i}}, \boldsymbol{x_{1j}}$ for the biomarker values in the ith control and jth case
- $\boldsymbol{\epsilon_{0i}}, \boldsymbol{\epsilon_{1j}}$ for the residual deviations in the ith control and jth case
- $\bar{\boldsymbol{x}}_{\mathbf{0}}, \bar{\boldsymbol{x}}_{\mathbf{1}}$ for the sample means of the biomarkers in N controls and N cases respectively.
- $\boldsymbol{y_0}, \boldsymbol{y_1}$ for the biomarker vectors of a new case and a new control, standardized by subtracting the combined mean of the case and control samples
- $\boldsymbol{\epsilon_0}, \boldsymbol{\epsilon_1}$ for the residual deviations in this new case and new control.

These vectors are then defined as sums of random variables:-

$$
\begin{aligned}
\boldsymbol{x_{0i}} &= \boldsymbol{\eta} - \frac{1}{2}\boldsymbol{\beta} + \boldsymbol{\epsilon_{0i}} \\
\boldsymbol{x_{1j}} &= \boldsymbol{\eta} + \frac{1}{2}\boldsymbol{\beta} + \boldsymbol{\epsilon_{1j}} \\
\bar{\boldsymbol{x}}_{\mathbf{0}} &= \boldsymbol{\eta} - \frac{1}{2}\boldsymbol{\beta} + \sum_i \frac{\boldsymbol{\epsilon_{0i}}}{N} \\
\bar{\boldsymbol{x}}_{\mathbf{1}} &= \boldsymbol{\eta} + \frac{1}{2}\boldsymbol{\beta} + \sum_j \frac{\boldsymbol{\epsilon_{1j}}}{N} \\
\bar{\boldsymbol{x}} := \frac{1}{2}\left(\bar{\boldsymbol{x}}_{\mathbf{0}} + \bar{\boldsymbol{x}}_{\mathbf{1}}\right) &= \boldsymbol{\eta} + \frac{1}{2}\left(\sum_i \frac{\boldsymbol{\epsilon_{0i}}}{N} + \sum_j \frac{\boldsymbol{\epsilon_{1j}}}{N}\right) \\
\boldsymbol{b} := \bar{\boldsymbol{x}}_{\mathbf{1}} - \bar{\boldsymbol{x}}_{\mathbf{0}} &= \boldsymbol{\beta} + \sum_j \frac{\boldsymbol{\epsilon_{1j}}}{N} - \sum_i \frac{\boldsymbol{\epsilon_{0i}}}{N} \\
\boldsymbol{y_0} := \boldsymbol{x_0} - \bar{\boldsymbol{x}} &= -\frac{1}{2}\boldsymbol{\beta} + \boldsymbol{\epsilon_0} - \frac{1}{2}\left(\sum_i \frac{\boldsymbol{\epsilon_{0i}}}{N} + \sum_j \frac{\boldsymbol{\epsilon_{1j}}}{N}\right) \\
\boldsymbol{y_1} := \boldsymbol{x_1} - \bar{\boldsymbol{x}} &= \frac{1}{2}\boldsymbol{\beta} + \boldsymbol{\epsilon_1} - \frac{1}{2}\left(\sum_i \frac{\boldsymbol{\epsilon_{0i}}}{N} + \sum_j \frac{\boldsymbol{\epsilon_{1j}}}{N}\right)
\end{aligned}
$$

The vector $(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{y_0}, \boldsymbol{y_1})$ is distributed with zero mean and cross-covariance (omitting the upper diagonal)

$$
\begin{array}{c}
\begin{array}{ccccc} \boldsymbol{\eta} & \boldsymbol{\beta} & \boldsymbol{b} & \boldsymbol{y_0} & \boldsymbol{y_1} \end{array} \\
\begin{pmatrix}
\frac{1}{K}\boldsymbol{\Sigma} & & & & \\
0 & \frac{1}{J}\boldsymbol{\Sigma} & & & \\
0 & \frac{1}{J}\boldsymbol{\Sigma} & \left(\frac{1}{J} + \frac{2}{N}\right)\boldsymbol{\Sigma} & & \\
0 & -\frac{1}{2J}\boldsymbol{\Sigma} & -\frac{1}{2J}\boldsymbol{\Sigma} & \left(1 + \frac{1}{4J} + \frac{1}{N}\right)\boldsymbol{\Sigma} & \\
0 & \frac{1}{2J}\boldsymbol{\Sigma} & \frac{1}{2J}\boldsymbol{\Sigma} & \left(-\frac{1}{4J} + \frac{1}{N}\right)\boldsymbol{\Sigma} & \left(1 + \frac{1}{4J} + \frac{1}{N}\right)\boldsymbol{\Sigma}
\end{pmatrix}
\begin{array}{c} \boldsymbol{\eta} \\ \boldsymbol{\beta} \\ \boldsymbol{b} \\ \boldsymbol{y_0} \\ \boldsymbol{y_1} \end{array}
\end{array}
\tag{1}
$$

From standard results for the conditional distribution of a multivariate gaussian it follows that the distributions of $\boldsymbol{y}_1$, $\boldsymbol{y}_0$ given $\boldsymbol{b}$ have means $\boldsymbol{\mu}_1 := \frac{\frac{1}{2J}}{\frac{1}{J}+\frac{2}{N}}\boldsymbol{b}$, $\boldsymbol{\mu}_o := -\frac{\frac{1}{2J}}{\frac{1}{J}+\frac{2}{N}}\boldsymbol{b}$ and variance $\sigma^2\boldsymbol{\Sigma}$ where $\sigma^2 := 1 + \frac{1}{4J} + \frac{1}{N} - \frac{\frac{1}{4J^2}}{\frac{1}{J}+\frac{2}{N}}$

Using angled brackets $\langle\rangle_{\mathcal{H}}$ to denote taking the expectation under hypothesis $\mathcal{H}$, and $D^2\left(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ to denote the squared Mahalanobis distance of observation $\boldsymbol{x}$ from a

distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the expectation given $\boldsymbol{b}$ and $\mathcal{H}_{\text{case}}$ of the log Bayes factor favouring case over control status is

$$\begin{aligned}
&\langle \log p\left(\boldsymbol{y} \mid \boldsymbol{b}, \mathcal{H}_{\text{case}}\right)\rangle_{\mathcal{H}_\text{case}} - \langle \log p\left(\boldsymbol{y} \mid \boldsymbol{b}, \mathcal{H}_{\text{control}}\right)\rangle_{\mathcal{H}_\text{case}} \\
&= -\frac{1}{2}\log\det\sigma^2\boldsymbol{\Sigma} - \frac{1}{2}\langle D^2\left(\boldsymbol{y}_1, \boldsymbol{\mu}_1, \sigma^2\boldsymbol{\Sigma}\right)\rangle + \frac{1}{2}\log\det\sigma^2\boldsymbol{\Sigma} + \frac{1}{2}\langle D^2\left(\boldsymbol{y}_1, \boldsymbol{\mu}_0, \sigma^2\boldsymbol{\Sigma}\right)\rangle \\
&= -\frac{1}{2}D^2\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \mathbf{0}, \sigma^2\boldsymbol{\Sigma}\right)
\end{aligned} \tag{2}$$

as the expectation of the squared Mahalanobis distance of a P-dimensional observation from its own gaussian distribution is P. Taking the expectation over $\boldsymbol{b}$ of the Mahalanobis distance of $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ from a distribution with mean $\mathbf{0}$ and covariance $\sigma^2\boldsymbol{\Sigma}$, the expected log Bayes factor is

$$\frac{1}{2}P\frac{\frac{1}{J^2}}{\left(1 + \frac{1}{4J}\left(1 - \frac{N}{N+2J}\right) + \frac{1}{N}\right)\left(\frac{1}{J} + \frac{2}{N}\right)} \tag{3}$$

Thus the predictive performance of the learned classifier, given J, P and N does not depend on the correlation structure of the biomarkers, as long as the correlation matrix is of full rank. For infinite sample size N the maximum expected log Bayes factor Λ_{opt} is $\frac{1}{2}P/J$ and the C-statistic C_{opt} of an optimally-trained classifier is thus $1 - \Phi\left(-\sqrt{\frac{1}{2}P/J}\right)$.

Rearrangement of equation (3) shows that for large J, the ratio of case sample size N to number of biomarkers P is approximately

$$\frac{F}{(1-F)\,\Lambda_{\text{opt}}} \tag{4}$$

where F is the ratio of the expected log Bayes factor obtained with the learned classifier to the expected log Bayes factor Λ_{opt} obtained with the optimally-trained classifier. For illustration we consider two hypothetical biomarker panels: one that in a clinical setting would be considered a good predictor, with $C_{\text{opt}} = 0.9$ (equivalent to $\Lambda_{\text{opt}} = 1.64$ natural log units) and one that would be considered a moderate predictor with $C_{\text{opt}} = 0.75$ (equivalent to $\Lambda_{\text{opt}} = 0.45$ natural log units). If the criterion for adequate sample size is specified as that the learned model should extract 80% of the predictive information that would be obtained from the optimal model ($F = 0.8$). the required ratio of cases to biomarkers is 2.4 for the good predictor and 8.9 for the moderate predictor.

Sparse prior distribution of effect sizes

The assumption of gaussian priors on the biomarker effect sizes may be plausible for a panel of candidate biomarkers that have been selected as likely to be relevant to the outcome under study. In many situations, it is not possible to preselect the most

biologically plausible biomarkers, and a generic measurement platform is used, such as a genome-wide microarray to measure gene expression ($\sim 20,000$ transcripts), or mass spectrometry to measure metabolite profiles (~ 1000 metabolites). The biomarkers included on such platforms are not selected for relevance, and it is expected that only a small proportion of them will be predictive of the outcome under study. A more realistic prior to use with such biomarker panels is a spike-and-slab mixture, in which a proportion $(1-\phi)$ of biomarkers have zero effect size (the spike) and the remaining proportion ϕ have a gaussian distribution of effect sizes with mean zero and inverse variance J (the slab).

For a single biomarker in a new case the observed value y_1 is distributed conditional on the case-control difference b with mean $f\mu_1$, and variance (from the law of total variance)

$$f\left(1+\frac{1}{4J}+\frac{1}{N}-\frac{\frac{1}{4J^2}}{\frac{1}{J}+\frac{2}{N}}\right)+(1-f)\left(1+\frac{1}{N}\right)+f\,(1-f)\,\mu_1^2$$

where $\mu_1 := \frac{\frac{1}{2J}}{\frac{1}{J}+\frac{2}{N}}b$. For a single biomarker in a new control μ_1 is replaced by $\mu_0 := -\mu_1$. f is the posterior probability of the marker being in the slab mixture component given the observed effect size b, calculated as

$$\frac{\frac{\phi}{\sigma_{\text{slab}}}\exp\left(-\frac{1}{2}\frac{b^2}{\sigma^2_{\text{slab}}}\right)}{\frac{\phi}{\sigma_{\text{slab}}}\exp\left(-\frac{1}{2}\frac{b^2}{\sigma^2_{\text{slab}}}\right)+\frac{(1-\phi)}{\sigma_{\text{spike}}}\exp\left(-\frac{1}{2}\frac{b^2}{\sigma^2_{\text{spike}}}\right)} \tag{5}$$

where $\sigma^2_{\text{slab}} = \frac{2}{N}+\frac{1}{J}$, $\sigma^2_{\text{spike}} = \frac{2}{N}$ are the variances of the observed effect size b conditional on the biomarker being in the slab and the spike mixture components respectively.

If the class-conditional distributions of y are approximated by gaussians, the expected log Bayes factor contributed by a single biomarker conditional on the observed effect size b can be evaluated as the squared difference between the class-conditional means divided by the class-conditional variance, by the same argument as that used to derive equation (2). The expectation of this conditional expectation can be evaluated by averaging over the distribution of b under the spike and slab prior on effect sizes. Importance sampling can be used to compute this expectation efficiently. If the distributions of vectors $\boldsymbol{y}$ and $\boldsymbol{b}$ are further approximated by multivariate gaussians with the same correlation matrix, the expected log Bayes factor contributed by P biomarkers can be evaluated by multiplying this average by P. A justification for these gaussian approximations is that only the biomarkers that are in the gaussian mixture component contribute nonzero values to the expected log-likelihood.

Figure 1 compares the sample size required to learn a classifier when only 1% of the biomarkers have nonzero effect sizes ($\phi = 0.01$) with that required when the effect sizes have a gaussian distribution ($\phi = 1$), for a given size of biomarker panel ($P = 10000$) and two illustrative values of C_{opt}: 0.9 and 0.75 as before. This shows that it is far easier to learn the classifier when the effect sizes have a sparse distribution than when they have a gaussian distribution: the required sample size falls from 24400 (2.44 cases per variable) to 980 for $C_{\text{opt}} = 0.9$ and from 87500 to 3400 for $C_{\text{opt}} = 0.75$. It would of

course be easier still if the subset of biomarkers that have nonzero effects could be specified in advance, on the basis of biological plausibility: the required case sample size would then be 250 for $C_{\text{opt}} = 0.9$ and 880 for $C_{\text{opt}} = 0.75$.

Discussion

A widely quoted rule of thumb for logistic regression modelling and prediction of a binary outcome is that the case sample size should be at least ten times the number of variables (5; 6; 7). For linear discriminant analysis, a less stringent requirement that the number of observations should be about three times the number of variables has been stated (8). The results derived above show that such general rules do not take account of how the required case sample size depends on the performance of the optimal predictor that could be obtained from the panel and on the sparseness of the distribution of biomarker effect sizes. Thus to learn a predictor that extracts 80% of the information in the biomarker panel, the required ratio of cases to variables varies from 0.1 for a platform with optimal C-statistic of 0.9 and a sparse distribution of biomarker effect sizes such that 1% of biomarkers have nonzero effects, to about 9 for a platform with optimal C-statistic of 0.75 in which the biomarker effect sizes have a gaussian distribution. A key conclusion of this paper is that required sample sizes are critically dependent upon assumptions about the sparsity of the biomarker effects. For a researcher planning a new study, published results of previous studies that have used the same biomarker panel to predict other outcomes may help to set a plausible range of values for the number of biomarkers that have nonzero effects.

Several authors have examined the problem of how to calculate sample size requirements for learning a classifier based on linear discriminant analysis. Lachenbruch (1968) derived an expression for the misclassification rate, corrected for overfitting (9). As with the approach described here, this expression did not depend upon the covariance structure of the predictor variables. However the correction for overfitting could only be used where the sample size N was large compared with the number of variables P, so this method is not applicable to studies with high-dimensional biomarker panels.

Dobbin and Simon (2007) described a method for estimating the sample sizes required to develop classifiers using high-dimensional DNA microarray data (10). Their approach requires either that the biomarkers can be assumed to be independent, or that data are available from which to estimate the covariance structure of the biomarkers (or at least the eigenvalues of the correlation matrix). Dobbin and Song (2013) suggested using an errors-in-variables method to estimate the learning curve of the predictor, based on treating the risk predictor as a noisy "observation" of the true (optimally trained) risk predictor (2). This method however requires a pilot dataset.

The approach described here does not require the covariance structure of the biomarkers to be known in advance, but relies instead on the simplifying assumption that the effect sizes have the same correlation structure as the biomarkers. This assumption is plausible if the correlations between biomarkers and the correlations between effect sizes are generated by the same latent variables. In practice, biomarkers that are highly correlated usually have similar associations with clinical outcome. For instance, in a recent study

of whether biomarkers can be used to predict progression of chronic kidney disease, the baseline filtration rate was identified as a variable underlying correlations between many of the biomarkers most predictive of disease progression rate (11). Although this approach depends upon the likelihood being adequately approximated by a gaussian, the accuracy of this approximation should increase with the size of the set of biomarkers that have nonzero effects.

The calculations of required sample size given here do not make any specific assumptions about how the effects of the biomarkers will be modelled. For constructing classifiers from gene expression arrays, which typically measure a few tens of thousands of variables, methods based on penalized linear discriminant analysis are widely used (12), corresponding to the statistical model assumed for the sample size calculations in this paper. For more low-dimensional problems, penalized logistic regression is often preferred to linear discriminant analysis as it does not require the assumption of gaussian class-conditional distributions for the predictors, and the predictor has the property of maximum entropy given correct calibration on the training data.

For a researcher who is trying to develop a biomarker-based predictor, the initial objective may be to evaluate which of various high-dimensional biomarker platforms contain enough predictive information to be useful, before expending resources on a large case-control study to learn a classifier. For this, it is necessary only to collect enough data from which to estimate the parameters of the spike-and-slab distribution of effect sizes considered in this article. For prediction from genome-wide genotype data, these parameters can be estimated from published summary case-control data (13).

Online resources

An R script to generate a learning curve for classifier performance as a function of the ratio of cases to biomarkers for specified values ofC_{opt} and ϕ is available at https://pm2.phs.ed.ac.uk

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to this article.

Funding

The author received financial support from MRC/Arthritis Research UK for the MATURA (MR/K015346/1) during the preparation of this article.

References

[1] Jaffe S. Planning for US Precision Medicine Initiative underway. *Lancet (London, England)* 2015; 385: 2448–2449. DOI:10.1016/S0140-6736(15)61124-2.

[2] Dobbin KK and Song X. Sample size requirements for training high-dimensional risk predictors. *Biostatistics (Oxford, England)* 2013; 14: 639–652. DOI:10.1093/biostatistics/kxt022.

[3] Good IJ. A list of properties of Bayes-Turing Factors. *NSA Technical Journal* 1965; 10: 1–6. URL {https://www.nsa.gov/public{_}info/{_}files/tech{_}journals/list{_}of{_}properties.pdf}. Declassified.

[4] Clayton D. On inferring presence of an individual in a mixture: a Bayesian approach. *Biostatistics* 2010; 11(4): 661–673. DOI:10.1093/biostatistics/kxq035.

[5] Peduzzi P, Concato J, Kemper E et al. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; 49: 1373–1379.

[6] Courvoisier DS, Combescure C, Agoritsas T et al. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology* 2011; 64: 993–1000. DOI:10.1016/j.jclinepi.2010.11.012.

[7] Austin PC and Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res* 2017; 26: 796–808. DOI:10.1177/0962280214558972.

[8] Lachenbruch PA and Goldstein M. Discriminant analysis. *Biometrics* 1979; : 69–85.

[9] Lachenbruch PA. On Expected Probabilities of Misclassification in Discriminant Analysis, Necessary Sample Size, and a Relation with the Multiple Correlation Coefficient. *Biometrics* 1968; 24(4): 823–834. DOI:10.2307/2528873. URL http://www.jstor.org/stable/2528873.

[10] Dobbin KK and Simon RM. Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics (Oxford, England)* 2007; 8: 101–117. DOI:10.1093/biostatistics/kxj036.

[11] Looker HC, Colombo M, Hess S et al. Biomarkers of rapid chronic kidney disease progression in type 2 diabetes. *Kidney International* 2015; DOI:10.1038/ki.2015.199.

[12] Tibshirani R, Hastie T, Narasimhan B et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2002; 99: 6567–6572. DOI:10.1073/pnas.082099299.

[13] Vilhjálmsson BJ, Yang J, Finucane HK et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet* 2015; 97(4): 576–592. DOI:10.1016/j.ajhg.2015.09.001.

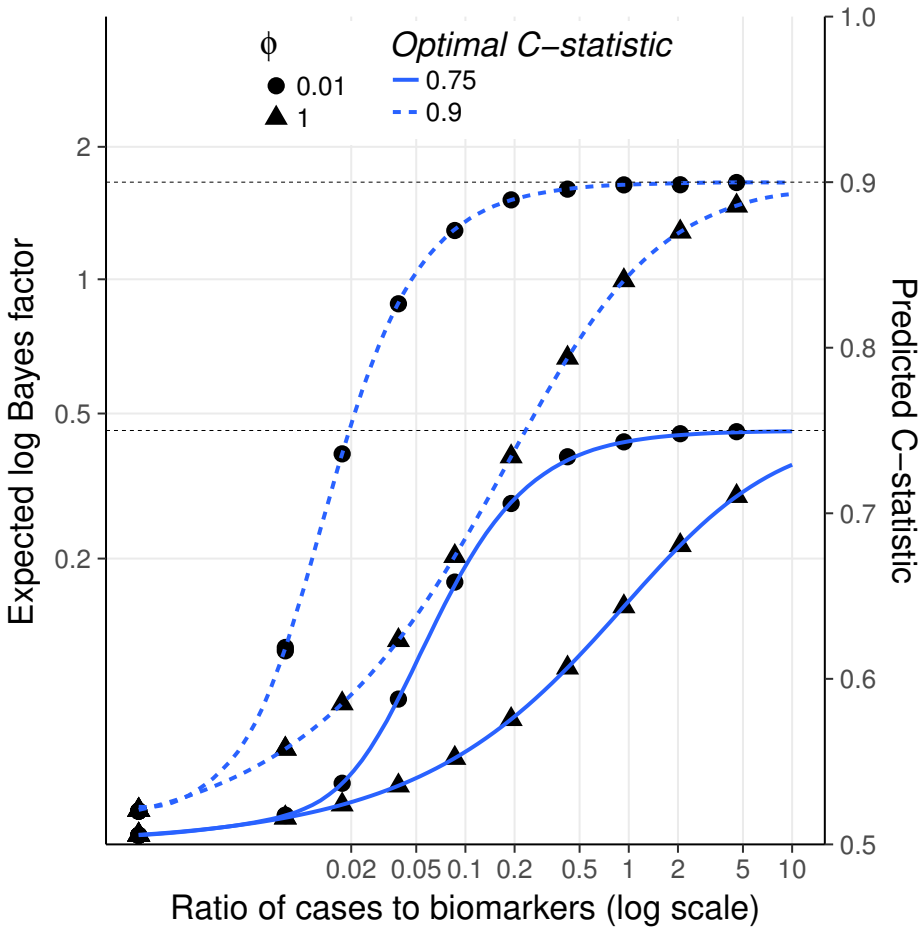

Figure 1. Learning curves for platform with 10000 biomarkers, comparing a gaussian distribution of effect sizes ($\phi = 1$) with a sparse spike-and-slab distribution ($\phi = 0.01$)