How Large a Training Set is Needed to Develop a Classifier for Microarray Data?

Kevin K. Dobbin, Yingdong Zhao, and Richard M. Simon

Abstract

Purpose: A common goal of gene expression microarray studies is the development of a classifier that can be used to divide patients into groups with different prognoses, or with different expected responses to a therapy. These types of classifiers are developed on a training set, which is the set of samples used to train a classifier. The question of how many samples are needed in the training set to produce a good classifier from high-dimensional microarray data is challenging.

Experimental Design: We present a model-based approach to determining the sample size required to adequately train a classifier.

Results: It is shown that sample size can be determined from three quantities: standardized fold change, class prevalence, and number of genes or features on the arrays. Numerous examples and important experimental design issues are discussed. The method is adapted to address ex post facto determination of whether the size of a training set used to develop a classifier was adequate. An interactive web site for performing the sample size calculations is provided.

Conclusion: We showed that sample size calculations for classifier development from high-dimensional microarray data are feasible, discussed numerous important considerations, and presented examples.

This review provides guidance on how to determine the number of arrays needed for microarray studies in which the objective is to construct a classifier based on gene expression. These types of studies are referred to as classifier development studies or class prediction studies (1). A classifier is a gene expression–based rule that can be applied to a future sample to provide a prediction of the class to which the sample belongs. Here, "class" is used as a general term, which, depending on context, may refer to patient response to a treatment, patient vital status 2 years after surgery, histologic type, etc. If constructing a classifier is the primary goal of the study, then this goal should be used to guide study design and sample size determination. Typical classifiers in clinical contexts are either prognostic, predictive, or both. For example, a prognostic classifier might classify a group of patients currently considered to have a homogeneous prognosis into subgroups with distinct prognoses—resulting in a refinement of the current prognostic system (e.g., American Joint Committee on Cancer/International Union Against Cancer stage). A predictive classifier classifies patients based on their predicted response to a particular targeted therapy, such as an epidermal growth factor receptor inhibitor. An example of a classifier that is both prognostic and predictive is the Oncotype DX classifier (2), which identifies a subset of breast cancer patients who, under current standards of care, receive adjuvant chemotherapy but whose prognosis is in fact so good that the small probability of benefit from chemotherapy is negligible.

The purpose of this review is 2-fold. First, to present a less technical, and more clinically oriented, version of our earlier sample size results accessible to non–statisticians involved in the design of these types of studies; and second, to extend our previous results (3) with new material. Our 2007 report presented a model-based approach to sample size determination, and contrasted with the one previous article on the topic (4). The sample sizes explored in the 2007 report were restricted to the situation of equal representation from each class. This review presents analyses of the effect of prevalence imbalance on sample size. Using mathematical techniques to eliminate as many unknowns as possible, we present a novel listing of the minimal set of information needed to make a sample size determination, and give examples of how data from previous experiments can be used to estimate required quantities. This review also adapts the methodology to the problem of determining ex post facto whether a training set sample size actually used in an experiment was adequate. A new web-based program interface for calculating sample size has been developed to accompany this review.[1] Novel real and synthetic data examples are provided throughout.

Authors' Affiliation: Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, NIH, Rockville, Maryland

Received 2/21/07; revised 8/22/07; accepted 9/18/07.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org).

Requests for reprints: Kevin K. Dobbin, National Cancer Institute, 6130 Executive Boulevard, EPN Room 8124, Rockville, MD 20852. Phone: 301-451-6244; E-mail: dobbinke@mail.nih.gov.

doi:10.1158/1078-0432.CCR-07-0443

[1] Available at http://linus.nci.nih.gov/brb/samplesize/.

Materials and Methods

A two-step model for classifier development. A mathematical model for developing a classifier has two steps. The first step models the process of selecting genes that are differentially expressed (DE) between the classes, and the second step models the construction of the classifier out of these DE genes. There are a variety of methods for identifying DE genes, but each is trying to do the same thing, i.e., include "informative" genes and eliminate "noise" genes from the classifier. Our method models the first step as follows: genes are selected based on gene-specific P values from t tests. Although the resulting multigene predictors are more complex than single-gene diagnostics traditionally studied, the genes included in such classifiers represent a vast reduction in scale compared with the tens of thousands of genes represented on modern microarrays. Although such a reduction is not required, it is almost always done in practice.

How many genes to include in the classifier. How many genes is a good number to have in a predictor? Intuitively, the larger the differences between the classes relative to the biological heterogeneity within each class, the easier it is to find a good classifier, and the more likely it is that a short gene list will be adequate. For example, if one is comparing different types of cancer samples, such as acute myeloid leukemia and acute lymphocytic leukemia, one may expect to have many genes with large differences in expression between the classes. So a classifier with a handful of genes may do about as well as one with thousands of genes, and there are probably lots of different handfuls that could do. More specifically, if the standardized fold change[2] for DE genes is 2.0, then 6 DE genes provide enough separation between the classes to obtain an accuracy of 98%, whereas 50 DE genes provide enough separation to obtain >99% accuracy—a marginal improvement. An example of this type of large standardized fold change is seen in a study of acute myeloid leukemia and acute lymphocytic leukemia (1), in which we estimate that standardized fold changes of >2.0 are present, and it has been found that very short gene lists perform very well.

When the differences between the classes are not large compared with the biological heterogeneity within each class, then short gene lists will not provide adequate separation between the classes. An example of this type of small standardized fold change is seen in prognostic signatures for lung cancer (5), in which we estimate that standardized fold changes of ~0.80 are present. To simplify this example, suppose there really are 50 genes each with a standardized fold change of 0.80 (which is an unlikely oversimplification of the truth). In this case, 6 genes provide enough separation between the classes to obtain 83% accuracy, whereas 50 DE genes provide enough separation to obtain 99% accuracy—a considerable improvement. So a longer gene list will be critical to accurate classification. Beer et al. (5) found ~50 genes to be the optimal number.

As can be seen from these examples, the optimal number of genes to include in a predictor is largely determined by the size of the largest standardized fold change present. Our method exploits this fact by optimizing the gene selection step of the model depending on the user-specified standardized fold change—so that the model reflects a near-optimal number of genes.

Robust classification versus stable gene lists. Our approach is based on the expected performance of the classifier for independent cases. As long as the classification is accurate for independent data, the actual classifier itself need not be unique. That is, it is not relevant for our approach whether the selected genes are stable—in the sense that if a classifier were built on a different set of samples, the same genes would be selected. The lack of reproducibility of gene lists that have been noted in some recent publications (6–8) is not an indication that the classifiers are not robust, as was recently shown by Fan et al. (9).

Sample size and optimal classifier performance. It is critical that the sample size required for a study be determined by the specific study objectives. Generic rule-of-thumb approaches are not adequate. For example, it has been suggested that thousands of samples may be needed to identify a definitive list of DE genes (7). But the identification of such a list is not the objective of a classifier development study. Hence, this finding is not a reason to use very large sample sizes. Using thousands of samples may be wasteful when far fewer are actually needed. Moreover, thousands of samples may not even produce a good classifier. Sometimes, even in high dimensions, there is just no good classifier. Some examples are given in Table 1 (note that explicit calculations associated with each figure and table in this review are provided in the Supplement). For example, row 1 of the table describes a situation in which the best classifier has a 37% error rate. Table 1 gives some information about how the number of DE genes and standardized fold change together influence potential performance. For example, larger effect sizes have a greater positive effect on potential performance than larger numbers of informative genes, as can be seen by comparing 35 genes with a standardized fold change of 0.25 (optimal error rate, 23%) to 14 genes with a standardized fold change of 0.75 (optimal error rate, 8%). Our approach is focused on the objective of classifier development, and explicitly takes these facts into account by adjusting for the performance of optimal classifiers like the ones listed in Table 1.

Prediction accuracy depends on sample size. Sample size planning for classifier development is different from sample size planning for testing a null hypothesis in a clinical trial. In the latter case, the sample size is usually determined so that statistical power for rejecting the null hypothesis is at least 0.80 or 0.90 when the true treatment effects are of a specified size. Increasing sample size in a classifier development study

Table 1. Examples of error rates for optimal prediction rules for a two-class situation

Number of DE genes	Standardized fold change	Lowest possible error rate (%)
7	0.25	37
7	0.50	25
7	0.75	14
7	1.00	9
7	1.50	2
7	2.00	<1
1	0.25	45
2	0.25	43
3	0.25	41
4	0.25	40
5	0.25	39
6	0.25	38
14	0.25	32
21	0.25	29
28	0.25	25
35	0.25	23
14	0.75	8
21	0.75	4
28	0.75	2
35	0.75	1

NOTE: Optimal error rates are >0% due to overlap between the two populations. For example, for the population described by row 1, seven genes are DE, with fold change 0.25 times the SD (standardized fold change), and the resulting optimal predictor has a 37% error rate. Multivariate normal model with diagonal covariance matrix assumed. Error rates do not depend on the number of genes on arrays.

[2] The standardized fold change of a gene is the difference between classes in log gene expression divided by of the gene's standard deviation of expression within a class, as in Table 1.

is for improving predictor accuracy rather than to increase power for testing a null hypothesis.

When considering the classifier accuracy associated with a particular sample size, say of $n = 50$, one must take into account the fact that this accuracy will depend on *which* 50 samples are ultimately selected for the training set. Different training sets of 50 samples will produce different classifiers with somewhat different accuracies. But there will be some overall average accuracy associated with training sets of size $n = 50$, and we will denote this average by $PCC(n)$. So $PCC(n)$ is the average probability of correct classification for classifiers built on training sets of size n. Given n, $PCC(n)$ is a fixed quantity that can be used for sample size determination. For example, $PCC(n) > 0.90$ means that on average samples of size n will produce a classifier with a probability of correct classification >90%. But, as we have seen, there may be too much overlap in the population to produce a classifier with such a high probability of correct classification.

Define $PCC(\infty)$ as the probability of correct classification for an optimal classifier[3] developed on an infinitely large training set. $PCC(\infty)$ takes into account the overlap between the classes in the population. One can always find a sample size large enough so that $PCC(n)$ is close to $PCC(\infty)$. In particular, if one specifies a tolerance $\gamma > 0$, one can find n big enough that $|PCC(\infty) - PCC(n)| < \gamma$ for that n. This motivates the following sample size objective: given a tolerance γ, find a sample size n large enough so that $|PCC(\infty) - PCC(n)| < \gamma$. In other words, pick n so that the expected accuracy of the resulting classifier is within γ of the best possible classifier.

Results

The minimal set of information required to determine sample size. We have developed sample size methods that require only the following minimal set of information be stipulated:

- the largest standardized fold change, as measured by the difference in average expression between the classes divided by the within-class standard deviation of expression of that gene (on the log scale),.
- the number of genes or features on the microarrays, and.
- the proportion of cases and controls in the population.

Each one of these elements is crucial to sample size determination. The reason that the largest standardized fold change is required is that the larger this is, the easier the classification problem, and the smaller the required sample size. The number of genes or features on the microarray affects the degree of difficulty of identifying DE genes and determines the expected proportion of false-positive genes included in the classifier. The proportion of each class in the population also affects the expected proportion of false-positive genes, and hence, the overall sample size requirements.

We will discuss briefly the investigations that have made it possible for us to eliminate the need for other information, particularly

(a) what significance level will be used to select genes to include in the classifier,
(b) what method will be used to construct the classifier,
(c) the number of DE genes, and
(d) the correlation/coregulation structure among the genes.

With regard to (a), methods for the identification of DE genes can be based on parametric or nonparametric tests, and may control the false discovery rate (10) or other related quantities (11). But all these methods have some elements in common. For example, the more stringent the method for identification of DE genes, the more likely really informative genes will be missed. On the other hand, the less stringent the method, the more likely the gene list will be cluttered with false-positive genes that are not truly informative and that may negatively affect classifier performance. So the best method is a balanced approach stringent enough to weed out uninformative genes but not so stringent that it also weeds out informative genes. It turns out that the optimal level of stringency depends on the sample size. We developed (3) a model-based method for finding the optimally stringent significance level cutoff to use in gene selection, defined as the one producing the best $PCC(n)$. Larger sample sizes and larger effect sizes were associated with smaller, more stringent optimal P value cutoffs for gene selection. Our sample size method, which searches over various sample sizes to find one that is adequate, automatically selects this optimal significance level each time.

With regard to (b), there are a wide variety of methods for developing classifiers from gene expression data. On the one hand, different methods applied to the same data set may produce classifiers with different error rates. On the other hand, well-established methods can be expected to result in classifiers with similar error rates. So, if one takes one of these well-established methods, and modifies it in a way that negatively affects the expected performance, the result will be conservative (large) sample size estimates. Our method is a general linear method, like the compound covariate predictor (12), but modified so as to ensure conservative sample sizes. We have shown, on a number of real and synthetic microarray data sets, that it produces sample sizes that are conservative under a variety of methods including compound covariate predictor, support vector machine, and nearest neighbor methods (3).

With regard to (c), in most situations, it is difficult to estimate the number of DE genes that will be observed in an experiment. In order to avoid this, our approach takes the worst case scenario and uses the number that will result in the poorest tolerance. This is done by a search over all possible numbers of informative genes.

With regard to (d), the general formulas we have developed do take into account gene coregulation by allowing a general covariance structure among the genes. However, empirical investigations with both simulated and real data sets indicated that the assumption of a single informative gene without this correlation structure adjustment resulted in sample sizes that were adequate or conservative. Hence, it was clear that the worst case scenario assumption about the number of DE genes led to a procedure which was conservative enough, and that the additional adjustment for the gene correlation structure was not required. Importantly, we were only able to identify this fact by developing formulas in terms of covariance matrix eigenvalues which made it explicit that the covariance structure adjustment would influence the sample size requirements.

The standardized fold change. To use many of the figures and tables in this review or our web-based program interface, one needs an estimate of the standardized fold change. If there is no data available from an experiment that used similar samples and platforms, then a hypothesized standardized fold

[3] Specifically, this is the probability of correct classification for an optimal linear classifier under the assumptions of the multivariate normal homogeneous variance model.

change must be used. The hypothesized value can result from reasoning such as: if the largest standardized fold change size isn't at least as big as x, then there is probably too much overlap between the classes to produce a classifier with clinically adequate accuracy; here, one would set the standardized fold change equal to x and compute sample size using Fig. 1, the tables, or the online program. For example, one might require a fold change of at least 2 (= 1 on the log-base 2 scale) in the context of human cancer samples on the Affymetrix GeneChip U133A platform. We have observed typical median variance on this platform of ~ 0.71 in several human cancer data sets, leading to a standardized fold change of $1/0.71 \approx 1.4$.

If data are available from a previous similar experiment that contained samples from each class, then one can estimate the standardized fold change from that data. Examples are presented below. Typically, such a data set will contain numerous genes that seem to be differentially expressed and each will have a different standardized fold change size associated with it. The largest standardized fold change estimate should be used. However, this will require some care. Because the observed largest effect size is likely to overestimate the true largest effect size because of random measurement error, multiplying this largest value by a shrinkage factor is preferable. We recommend using 0.80 as the shrinkage factor, which is the right factor when the ratio of biological to experimental error variation is 4.0. An example is provided in the caption to Table 4. We have shown (3) that if standardized fold change sizes are uniformly small, then it is often impractical to develop a classifier that separates the classes well.

Sample size determinations for specific experimental designs. Samples selected for use in a training set are usually chosen following one of two sampling designs: (a) random or consecutive sampling, (b) selecting equal numbers of patient samples from each class. Different sampling plans will require different sample sizes, so we treat these two designs separately.

Under sampling design (a), the proportion in each class in the training sample is likely to be similar to the population. Advantages of design (a) are that one can estimate the classifier's overall accuracy, and positive and negative predictive values, from the observed data. With design (b), these quantities can only be estimated by a weighted analysis that uses external estimates of population prevalence. A disadvantage of design (a) is that one runs some risk of having only a very small number of samples from one class, resulting in poor estimates of the predictor's actual performance in the underrepresented class and possibly uncertainty about the selection of cutoff points for classification.

Design a: random or consecutive sampling. In many situations, the two classes will not be equally represented in the population under study. For example, a prognostic classifier may predict recurrence 5 years after surgery, but the recurrence rate may be only 10%; in this case, the good outcome class represents 90% of the population and the poor outcome class represents 10% of the population, so they are unequal. In developing a predictive classifier of the patients who respond to a new drug, the response rate is often <50%.

Under random or consecutive sampling, the quality of the gene list is affected by the prevalence in each class. The quality will decrease the more unequal the prevalence is. This decrease in quality will affect the expected accuracy of the predictor because the predictor is a function of this gene list.

We have extended the application of our sample size methodology to situations with unequal prevalence from the two classes to examine the effect of prevalence imbalance. Results are presented in Tables 2 and 3. As can be seen from the

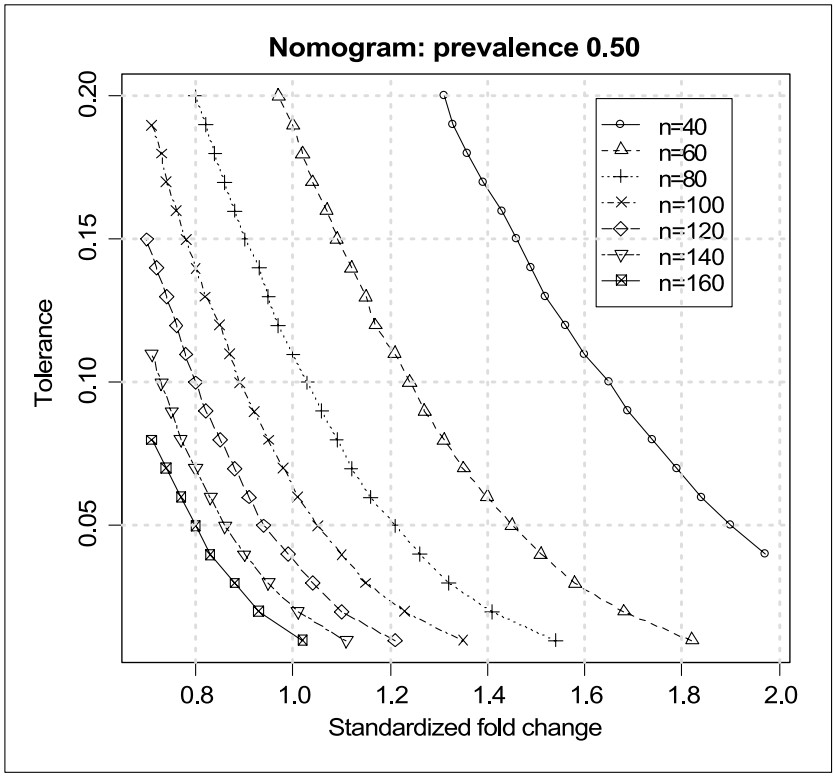

Fig. 1. Nomogram of the effect of standardized fold change and tolerance on sample size when the prevalence is 0.50. Curved lines, total sample size n, with $n/2$ per group. Arrays with 22,000 genes assumed.

Table 2. Sample size requirements for tolerance of 0.10

Standardized fold change	Prevalence in underrepresented class						
	15%	20%	25%	30%	35%	40%	50%
2.0	38	36	34	33	32	32	31
1.9	42	38	36	35	35	34	34
1.8	44	43	39	38	37	37	35
1.7	49	46	43	41	40	39	39
1.6	54	49	47	45	43	42	42
1.5	60	55	52	49	48	46	46
1.4	68	63	58	55	53	51	50
1.3	77	70	65	62	59	57	55
1.2	90	81	75	69	67	64	63
1.1	104	93	86	80	76	74	71
1.0	126	113	102	94	89	86	83

NOTE: Sample size for random (or consecutive) sampling. Standardized fold change is the difference in log-base 2 expression divided by the within-class standard deviation. The sample sizes assure that average accuracy, PCC(n), is within the specified tolerance (0.10) of the optimal accuracy, PCC(∞). Microarray with 22,000 genes assumed.

tables, the required sample size increases as the prevalence imbalance increases. Suppose, for example, that one wishes to develop a predictive classifier of complete response (CR) to a regimen with a CR rate expected to be ~20%. Suppose that one targets a standardized fold change of 1.4-fold. For a tolerance of 0.05 (Table 3), 83 samples, consisting of ~17 CR's and 66 non-CR's would be required. For a less stringent tolerance of 0.10 (Table 2), 63 samples, consisting of ~13 CRs and 50 non-CRs would be required.

Design b: selecting equal numbers of patient samples from each class. When the training set is constructed by selecting equal numbers from each class, our sample size formulas can be used with the prevalence set to 50%. For example, Fig. 1 can be used for planning classifier development studies. On the X-axis is the standardized fold change. On the Y-axis is the tolerance. The lines represent different size training sets, with $n/2$ from each class. Suppose that equal numbers of patients from each recurrence status ($\pm$)2 years after surgery will be selected to assess the potential of prognostic classification. Furthermore, suppose that a standardized fold change of 0.80, similar to the article by Beer et al. (5), is anticipated, then for a tolerance of 0.05, the required sample size would be ~160 samples (80 per class). Note that this sample size estimate is smaller than the estimate from Table 4 (n = 203), because in that table, the sampling plan was random sampling with unequal prevalence (75% versus 25%). Raising the tolerance to 0.10 in Fig. 1 results in a sample size requirement of ~120 samples (60 per class).

Small sample sizes and average accuracy. If our method is used for sample size determination, then the average accuracy of the resulting classifier will be within the tolerance of the best possible classifier. In other words, if one repeatedly took samples of the same size from the population and constructed a classifier, on average, these classifiers would have that accuracy. But this level of control might not always be adequate. If there is substantial variation in classifier performance, then there might be a high probability that the classifier will have accuracy well below this average. For example, we generated multivariate normal data sets with three informative genes (out of 1,000 total genes) with a standardized fold change size of 1.5; for each data set, we constructed a compound covariate predictor and calculated (mathematically) the accuracy of each predictor derived from eight independent replications. When the sample size was n = 24, the lowest and highest observed accuracies were 70% and 91%, respectively, producing a range of 21%. When the sample size was n = 100, the lowest and highest accuracies were 88% and 90%, respectively, with a range of 2%. Therefore, with a smaller sample size, control of the averages of these highly variable classification accuracies may not be adequate.

Example sample sizes estimated from real microarray data sets. Table 4 gives examples of sample size calculations using the methods described here for several real microarray data sets. In the leftmost two columns are descriptions of the data sets and classes. The sample size actually used in the studies appears in column 4, along with the actual prevalence (in parentheses). Our estimated sample size requirements for a tolerance of 0.10 and 0.05 appear in the right-most two columns. In these example data sets, sample size requirement estimates range from 20 to 80 for easier morphology or signature-based classification problems, such as Golub et al. (1), Pomeroy et al. (13), and Rosenwald et al. (14). For more difficult distinctions related to outcome or follow-up, such as Beer et al. (5) and van't Veer et al. (15), estimates range from 80 to 200. By comparing the rightmost two columns to the fourth column, one can determine whether the sample size actually used in the study was larger than the conservative sample size estimates produced by our method. If the actual size used was larger, then the study size was likely adequate. For example, the Golub et al. (1) study had a much larger overall sample size than needed. The 1999 article used 72 samples altogether, whereas even our conservative method indicates that only 23 to 28 samples are required. This may explain the strong results of the 1999 article. On the other hand, our method indicates that the Beer et al. (5) and van't

Table 3. Sample size requirements for tolerance of 0.05

Standardized fold change	Prevalence in underrepresented class						
	15%	20%	25%	30%	35%	40%	50%
2.0	50	46	43	41	39	38	38
1.9	56	49	46	44	42	41	39
1.8	60	53	51	48	46	44	43
1.7	66	58	54	52	50	49	47
1.6	74	66	60	57	55	53	51
1.5	84	73	67	62	60	59	58
1.4	95	83	75	70	67	65	63
1.3	110	95	86	80	76	73	71
1.2	130	111	99	92	87	83	82
1.1	156	133	118	108	101	96	94
1.0	190	163	142	129	119	113	107

NOTE: Sample size for random (or consecutive) sampling. Standardized fold change is the difference in log-base 2 expression divided by the within-class standard deviation. The sample sizes assure that average accuracy, PCC(n), is within the specified tolerance (0.05) of the optimal accuracy, PCC(∞). Microarray with 22,000 genes assumed.

Table 4. Application of sample size methods to some well-studied microarray data sets

Data set descriptions		Statistics from data set		Estimates using our method	
Data source (references)	**Classes**	**Maximum standardized fold change estimate**	**Sample size used in study (smaller prevalence)**	**Sample size for tolerance 0.10**	**Sample size for tolerance 0.05**
Beer et al. (5)	Alive vs. dead at 2 years	0.80	67 (25%)	133	203
van't Veer et al. (15)	Metastases vs. no metastases at last follow-up	1.02	97 (47%)	82	106
Pomeroy et al. (13)	Medulloblastoma vs. other	1.53	90 (33%)	41	54
Golub et al. (1)	AML vs. ALL	2.38	72 (35%)	23	28
Rosenwald et al. (14)	GCB vs. non-GCB	1.14	240 (48%)	61	79

NOTE: The second column is the two classes the predictor is trained to distinguish. The third column is the estimated maximum standardized fold change. Sample sizes (with prevalence) actually used in the previous studies appear in the fourth column. Sample sizes calculated using our method appear in the rightmost two columns. The number of genes on the arrays was 7,129, 24,482, 7,129, 7,129, and 7,399 (top to bottom). The maximum standardized fold change size was estimated from the original data using the formula $0.80 \cdot \max |t_g| \cdot \frac{1}{n_1} + \frac{1}{n_2}$. Here $\max|t_g|$ is the t test statistic with the largest absolute value. n_1 and n_2 are the number of samples from each of the two classes, and the 0.80 at the beginning of the equation is a shrinkage factor used to adjust for the fact that the largest observed t statistic is likely to produce an overestimate of the true largest standardized fold change due to experimental noise. This shrinkage factor results from assuming that the ratio of biological signal to measurement error is 4:1—which we have found to be a reasonable estimate (16)—and after applying Bayesian methods (17).
Abbreviations: AML, acute myeloid leukemia; ALL, acute lymphocytic leukemia; GCB, germinal-center B-cell-like large-B-cell lymphoma.

Veer et al. (15) studies might not have used adequate sample sizes. But this conclusion would be more tentative because our method is conservative, tending to produce estimates that may be larger than necessary, so that it is possible that the sample sizes actually used were adequate.

Ex post facto evaluation of sample size adequacy. A collection of samples used as a training set sometimes results in a classifier that performs poorly either in cross-validation or on independent data. A natural question is then whether the poor performance is due to overlap between the classes, or to the training sample size having been too small. In this ex post facto context, our mathematical modeling approach can be adapted to assess whether the null finding is due to the training set being too small. Table 5 shows examples of how, using our method, the absolute value of the largest t statistic from an experiment can be used to get a conservative estimate of the tolerance associated with the training set size. In applying this table, it is important to keep in mind that the estimated tolerances here may be significantly larger than the true tolerance because this method was developed explicitly to ensure conservative results. But the method can provide an idea of whether the tolerance was adequate. For example, suppose a predictor was developed on a training set and applied to a large independent validation set. The observed accuracy on the validation set was 70%. The largest observed t test statistic on the training set was $\sim$7.0 and 50 samples were used (25 per class). The associated tolerance estimate is 0.06. This suggests that it may be possible to find a predictor with an accuracy as high as 76%. Because our method tends to be conservative, 76% would probably be an optimistic estimate of the best that is achievable. Depending on the context, 76% may still not be deemed adequate. The poor performance would seem to be due to overlap between the classes rather than inadequate sample size.

Discussion

We have presented methods for sample size determination for class prediction microarray studies and discussed a number of issues related to classifier development and performance assessment. We have presented methods for situations in which one class is underrepresented in the population. Our approach rests on simplifying assumptions that should lead to conservative sample size estimates that are large enough, but may be too

Table 5. After-experiment, ex post facto assessment of sample size adequacy

Largest observed t statistic (in absolute value)	Estimated tolerance		
	Sample size (n = 20)	**Sample size (n = 50)**	**Sample size (n = 80)**
5.0	0.42	0.15	0.11
6.0	0.37	0.11	0.08
7.0	0.27	0.06	0.04
8.0	0.16	0.03	0.02
9.0	0.09	0.01	0.01
10.0	0.05	<0.01	<0.01

NOTE: t statistics are from the gene-specific pooled variance t tests. Assumes 22,000 genes on each array and a prevalence of 50%. Total sample size is n with $n/2$ per class.

large, particularly in very easy classification situations with highly diverse classes. We showed how the approach could also be applied to the problem of ex post facto determination of whether a training sample used was large enough. An accompanying web-based interface for sample size determination is being made available. Our approach controls the expected accuracy to be within a specified tolerance of the best possible accuracy for the population.

References

1. GolubTR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by expression monitoring. Science 1999;286:531–7.
2. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
3. Dobbin KK, Simon RM. Sample size planning for developing classifiers using high dimensional DNA microarray data. Biostatistics 2007;8:101–17.
4. Mukherjee S, Tamayo P, Rogers S, et al. Estimating data set size requirements for classifying DNA microarray data. J Comput Biol 2003;10:119–42.
5. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival in patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
6. Ein-Dor L, Kela I, Getz G, et al. Outcome signatures in breast cancer: is there a unique set? Bioinformatics 2005;21:171–8.
7. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA 2006;103:5923–8.
8. Michiels S, Kosciehny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92.
9. Fan C, Oh DS, Wessels L, et al. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 2006;355:560–9.
10. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a powerful approach to multiple testing. J R Stat Soc Ser B 1995;57:289–300.
11. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plan Inference 2003;124:378–98.
12. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002;9:505–11.
13. Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002;415:436–42.
14. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937–47.
15. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
16. Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005;6:27–38.
17. Carlin BP, LouisTA. Bayes and empirical Bayes methods for data analysis. New York: Chapman & Hall; 1996.