OPEN ACCESS Freely available online

PLoS one

Factors Influencing the Statistical Power of Complex Data Analysis Protocols for Molecular Signature Development from Microarray Data

Constantin F. Aliferis[1,2,3]*, Alexander Statnikov[2], Ioannis Tsamardinos[2,4], Jonathan S. Schildcrout[3], Bryan E. Shepherd[3], Frank E. Harrell Jr.[3]

1 Center of Health Informatics and Bioinformatics, New York University, New York, New York, United States of America, **2** Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, United States of America, **3** Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America, **4** Department of Computer Science, University of Crete, Iraklio, Greece

Abstract

Background: Critical to the development of molecular signatures from microarray and other high-throughput data is testing the statistical significance of the produced signature in order to ensure its statistical reproducibility. While current best practices emphasize sufficiently powered univariate tests of differential expression, little is known about the factors that affect the statistical power of complex multivariate analysis protocols for high-dimensional molecular signature development.

Methodology/Principal Findings: We show that choices of specific components of the analysis (i.e., error metric, classifier, error estimator and event balancing) have large and compounding effects on statistical power. The effects are demonstrated empirically by an analysis of 7 of the largest microarray cancer outcome prediction datasets and supplementary simulations, and by contrasting them to prior analyses of the same data.

Conclusions/Significance: The findings of the present study have two important practical implications: First, high-throughput studies by avoiding under-powered data analysis protocols, can achieve substantial economies in sample required to demonstrate statistical significance of predictive signal. Factors that affect power are identified and studied. Much less sample than previously thought may be sufficient for exploratory studies as long as these factors are taken into consideration when designing and executing the analysis. Second, previous highly-cited claims that microarray assays may not be able to predict disease outcomes better than chance are shown by our experiments to be due to under-powered data analysis combined with inappropriate statistical tests.

Citation: Aliferis CF, Statnikov A, Tsamardinos I, Schildcrout JS, Shepherd BE, et al. (2009) Factors Influencing the Statistical Power of Complex Data Analysis Protocols for Molecular Signature Development from Microarray Data. PLoS ONE 4(3): e4922. doi:10.1371/journal.pone.0004922

Editor: Vladimir B. Bajic, University of the Western Cape, South Africa

Received July 18, 2008; **Accepted** February 5, 2009; **Published** March 17, 2009

Funding: The work was in part supported by grant 2R56LM007948-04A1. The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: constantin.aliferis@nyumc.org

Introduction

Microarrays and other high-throughput assaying technologies have generated immense opportunities for discovery spanning the spectrum from basic research to clinical studies [1–3]. As the field moves from simpler analyses (e.g., differential expression of single genes and clustering) to more complex analyses such as developing multivariate molecular signatures in supervised fashion, the interpretation of microarray data involves multifaceted analysis protocols with many sophisticated and interacting analytic steps [4]. Developing molecular signatures in particular, is playing an increasingly important role in a variety of research design objectives both in basic and translational studies. Such objectives include, for example, detecting complex and coordinated patterns of transcriptional response to chemotherapeutic agents on cell lines and predicting subsequent patient treatment response on the basis of this information [5], discovery of new drug targets [6], discovery of biomarkers [7], subtyping diseases [8] and personalizing treatments [9].

The reproducibility of gene expression microarrays across laboratories for individual gene expression measurements and the ability to differentiate between disease subtypes are well established in recent studies [2,10,11]. Essential to developing molecular signatures is not only assay reproducibility however, but also *statistical* reproducibility. The latter can be directly assessed by tests of statistical significance of the produced signatures. These tests are usually permutation based and were introduced in bioinformatics by [12,13] based on foundational works of [14,15].

Although substantial efforts have been invested in studying the statistical power of differential gene expression [16,17], much less is known currently about the power of testing molecular signatures for statistically significant (hence reproducible, "real") signal. The present work shows that four specific components of data analysis (error metric, error estimator, classifier, and event balancing) have

significant and compounding effects on probability (i.e., statistical power) to detect true signal in molecular signatures. These findings can help researchers design data analysis protocols that require fewer samples; they also shed light on the appropriateness of microarrays as an assay platform for outcome prediction. The present report uses theoretical analysis, simulation experiments, and empirical analysis of 7 human gene expression datasets [8,9,18–22]. The datasets were chosen so that the comparison to a previously published highly-cited protocol [23] constitutes a case study that demonstrates the practical benefits of improved statistical power on the resource efficiency and validity of analysis.

Results

We start with a theoretical analysis that shows how the choice of four specific components of data analysis protocols for molecular signature development and their statistical testing affects the statistical power to detect predictive signal. We then present a simulation study that demonstrates that depending on choice of the above components even strong signals can fail to be detected with routine sample sizes and that the effects of each component on statistical power are large and compounded. We subsequently test the insights and hypotheses generated by the theoretical analysis and simulation studies with real gene expression data. Specifically, we analyze 7 datasets [8,9,18–22]. These datasets are important for two reasons: first, they have been used to derive both clinically relevant signatures and to investigate underlying biological processes [8,9,18,20–22,24]. Second, a highly-cited prior analysis of the same datasets [23] reached the conclusion that statistically significant signal cannot be detected in 5 out of 7 datasets and thus either microarrays are incapable of predicting clinical outcomes (and by extension, giving insight into the biology of disease progression) or that studies with a few hundred samples are insufficient and only sample sizes in the order of thousands will lead to statistically reproducible findings [25,26]. Our investigation into the factors that affect power allows us to test the hypothesis that these prior results were due in large part to an underpowered analysis protocol, showing thus the great importance of careful planning of data analytics with an eye toward sufficient power.

Theoretical analysis

Effects of error metric on power. Fundamental to the assessment of predictive signal of molecular signatures is the choice of error metric that is used to quantify predictivity. An unfortunate frequent practice in the field of bioinformatics to date is to use as classification performance metric the proportion of misclassifications. Discontinuous error metrics such as proportion of misclassifications, sensitivity, and specificity are "improper scoring rules" however, since they impose arbitrary thresholds on predictor models and do not capture the uncertainty in the predictions [27]. The proportion of misclassifications moreover, is known to yield estimators with low power to detect signals in data when compared to other metrics such as area under the receiver operating characteristic (ROC) curve (AUC) [27]. The ROC curve is the plot of sensitivity versus 1-specificity for a range of continuous or discrete classification threshold values. AUC is equivalent to a rank correlation between predicted outcome probability and the observed outcome, requiring no categorization. AUC ranges from 0 to 1, with an AUC equal to 0 indicating the worst possible classifier, 0.5 representing a random (i.e., uninformative) classifier, and 1 representing perfect classification. Testing whether predictions are unrelated to true outcomes using AUC is equivalent to the Wilcoxon test, while testing for proportion of misclassifications is equivalent to using the Mood median test which has been shown to have poor efficiency compared to the Wilcoxon test [28]. A broader, non-parametric justification why AUC is more discriminative than proportion of misclassifications is provided by [29]. Supporting Information File S1 provides an example where two signatures have the same proportions of misclassifications but different predictivity which is captured by the AUC metric. Although counter-examples do exist, they are relatively rarer [29]. Hence the AUC is more powerful than proportion of misclassifications.

Effects of classifier on power. Statistical power is increased whenever the tested effect size (predictivity in our context) is larger and the variance is smaller (assuming fixed sample size for simplicity). Hence using a classifier that produces the most predictive signature (everything else being equal) directly translates to improved statistical power for detecting predictive signal. Statistical machine learning theory proves that different classifiers have different inductive biases (i.e., preferences for classes of models), and that a classifier family has to be matched to the characteristics of the domain in order to achieve optimal predictivity (and correspondingly optimal power to detect signal) [30]. Indeed, recent empirical studies with gene expression data have shown that specific classifiers, such as Support Vector Machines (SVMs) produce models with stronger predictive ability (signal) and higher robustness across many high-throughput datasets compared to several widely-used alternatives [31–33]. Other authors also corroborate the need to choose classifiers carefully by recommending against some complex classifiers in order to avoid overfitting [4]. The above results have been neglected by some authors [23] who claim that "*in principle, there is no biological or mathematical reason why one particular classification method should be better than others*" and do not optimize the choice of classifier for the data at hand when conducting statistical testing of microarray gene expression signatures. We will show that this adversely affects the power of their analyses.

Effects of error estimator on power. Procedures that estimate the generalization error of a signature are called "error estimators". A commonly used estimator is the *holdout estimator*. The holdout estimator is based on splitting the data in two random non-overlapping parts, deriving a signature from the first one and assessing its error in the second one. The holdout estimator is asymptotically unbiased, that is, with infinite test sample it produces an estimate that is the true error in the population. In small samples holdout estimates often deviate from the large-sample value. This variability is reduced as sample size grows [34].

From standard power-size analysis considerations it follows that the lowest-variance unbiased estimator has highest power [35]. Unfortunately, the holdout estimator has larger variance compared to several other unbiased estimators used in molecular signature studies [34], and this naturally leads to reduced statistical power. We elaborate on the reasons for this behavior by comparing the holdout to the *repeated 10-fold cross-validation estimator* [36]. The latter estimator is a variant of the well-known 10-fold cross-validation estimator which is calculated by balanced splitting of the data into 10 non-overlapping sample sets used for testing (while each complementing set is used for training) and averaging the test errors. The repeated 10-fold cross-validation estimator is obtained by running regular 10-fold cross-validation for 100 (or other sufficient number of) times with different splits of data into training and testing sets each time and by reporting the average estimate over all runs.

To see why holdout has higher variance than repeated 10-fold cross-validation, consider that there are several major sources of variance of estimators in practical use. These are: *sampling variance*, *split variance*, *testing set size*, and *internal variance*. Sampling variance

refers to the uncertainty associated with drawing a random sample of fixed finite size from a population. Split variance refers to uncertainty associated with drawing a random split of training and testing sets from all possible splits of a given sample with fixed training-testing ratio. Testing set size variance refers to variability in error estimates due to finite testing dataset size. Finally, internal variance refers to uncertainty associated with classifier instability (i.e., different training datasets lead to different signatures) and increases as training set size decreases [30,36]. The repeated 10-fold cross-validation essentially eliminates split variance by using many splits and averaging over them, and furthermore it reduces internal variance by using more sample for training than holdout (under typical training-testing split ratios). Both estimators have the same sampling variance. Finally, while testing set size variance is larger in the repeated 10-fold estimator than the holdout, the combination of higher split and internal variance makes overall the holdout to have higher variance and to be less powerful than repeated 10-fold cross-validation in many practical situations. Unfortunately this is often neglected in practical analysis and therefore using the holdout estimator leads to reduced ability to establish statistical significance of signatures.

Effects of event balancing on power. When in the context of error estimation one enforces that both the training and testing sets have the same proportion of events and non-events as the original full data, we will call such error estimation "event balanced". An important and subtle shortcoming of some data analysis protocols is to not balance the training and testing data, seriously affecting variance, statistical power (and potentially biasing error estimates). For example, in [23] the models were trained on samples with 50% event rates. They were then tested on samples the event prevalence of which was far below 50% thus yielding estimates that were less efficient than the standard holdout estimator in which the data are split at random. The result of this is evident in Figure 2 from [23] in which as the sampling moves to larger training sets, this forces the testing sets in addition to being smaller, to implicitly have a very low event rate and thus large variance of error estimates. Notice that most classifiers, including the one used by [23], are designed to work under the assumption that the training and testing sets are identically distributed [30]. It is thus unrealistic to expect in general that a classifier that is trained using data from a distribution where events and non-events are equally likely will perform well, without adjustments [37], in a different distribution where this ratio is heavily distorted. This is especially so when using an error metric that is sensitive to event priors such as proportion of misclassifications. Supporting Information File S2 shows via an example that this shift in distributions can affect the performance of even an optimal classifier, i.e., one that has learned perfectly the distribution of the training data, to the point of appearing to be no better than flipping a coin.

Simulation experiments

A primary purpose of the simulation experiments is to demonstrate and study the relative importance of the above factors that are hypothesized on the previous theoretical grounds to influence power of complex data analysis of high-throughput data. The simulation uses an idealized analysis in which the data-generating process is known and the true moderate-strength signal is present even in small samples. Such analysis is typical in literature discussing statistical issues surrounding microarray data because knowing the generative model allows a precise characterization of the strengths and limitations of data analysis techniques [38]. Details of the simulation are provided in the Supporting Information File S3. A second goal is to examine the statistical power of a previously published data analysis protocol [23] ("Protocol I", that employs non-balanced holdout estimator with proportion of correct classifications and a nearest-centroid classifier), specifically when varying these four factors. Finally, a third goal is to test the relative power of the theoretically expected more powerful protocol ("Protocol II", that employs balanced repeated 10-fold cross-validation estimator with AUC as the error metric and SVMs as classifier). The best protocol in the simulations will then be validated in the next sub-section with real data.

The left part of Figure 1 demonstrates the inability of Protocol I [23] to detect signal which is detectable by Protocol II. The right part of this figure shows results of application of Protocol II and assessment of its statistical significance by permutation testing (details about statistical significance testing are provided in the Materials and Methods section). Overall Protocol I has remarkably small power ranging from less than 0.002 to 0.3 (depending on the criterion used for rejecting the null hypothesis, please see Supporting Information File S3). In contrast, Protocol II has power 0.93. By replacing proportion of misclassifications with AUC in Protocol I, its power increases to 0.6, and by additionally adding the use of SVMs, it further increases to 0.75. Conversely, if we start with Protocol II and replace AUC with proportion of misclassifications and SVMs with the classifier from [23], these changes reduce the power from 0.93 to 0.46. These empirical power estimates do not provide the exact power in real datasets since the true nature of the corresponding distributions is not known and varies among datasets. However the simulation strengthens our hypothesis that the choice of error metric, classifier, event balancing and error estimator have large impact on study results and sheds light on the limitations of the analyses described in prior work [23]. In the next sub-section we test the Protocol II in real data (where Protocol I was previously independently applied).

Analysis of real gene expression data

Figure 2 reports the AUC estimates produced with Protocol II for each one of the 7 real datasets along with p-values for testing the null hypothesis that the produced signatures are uninformative (i.e., with no signal). As can be seen in Figure 2, statistically significant signal (at the 0.05 level) can be detected in 6 out of 7 datasets compared to 2 out of 7 in the prior study that had used the less powerful Protocol I [23]. The p-values are calculated by a standard label-permutation procedure (see Materials and Methods section). The histograms in Figure 2 depict with blue the distribution of the repeated 10-fold cross-validation AUC estimates from Protocol II for datasets produced under the null hypothesis of "no predictive signal" and with red the repeated 10-fold cross-validation AUC estimates from Protocol II for the original datasets.

The above repeated 10-fold cross-validation AUC estimates in the datasets that had statistically significant signal ranged from 0.67 to 0.76, indicating that even signal with weak strength can be shown in real data to be statistically significant with moderate sample sizes. The U-statistic based confidence intervals for repeated 10-fold cross-validation AUC estimates are provided in the Supporting Information File S4 and they are consistent with the above conclusions. Notice also that under the null hypothesis of no predictive signal the distribution of the repeated 10-fold cross-validation AUC estimates is centered at 0.5, which corroborates the theoretical expectation that the error estimates are unbiased and that Protocol II does not overfit (more details in the Supporting Information File S5).

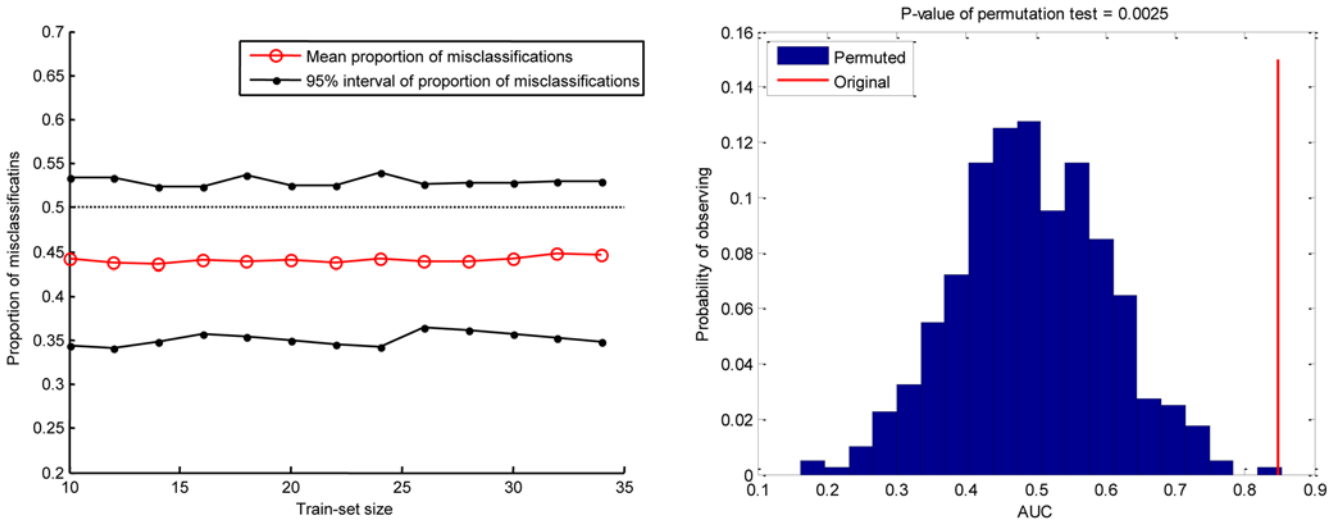


Figure 1. Comparison of Protocols I and II in simulated data. *Left:* Example where the Protocol I [23] applied to simulated data with true moderate-strength signal fails to detect statistical significance at all training set sizes. *Right*: a more powerful protocol (Protocol II, based on event balanced repeated 10-fold cross-validation with SVM classifiers and AUC metric) detects statistically significant predictive signal according to an outcome-value permutation test. Specifically, the p-value of the null hypothesis of no signal is 0.0025. The blue bars depict the distribution of repeated 10-fold cross-validation AUC estimates over 400 random datasets produced via outcome value permutation. The red line depicts the value of repeated 10-fold cross-validation AUC on the original data (i.e., without perturbing the outcome values).
doi:10.1371/journal.pone.0004922.g001

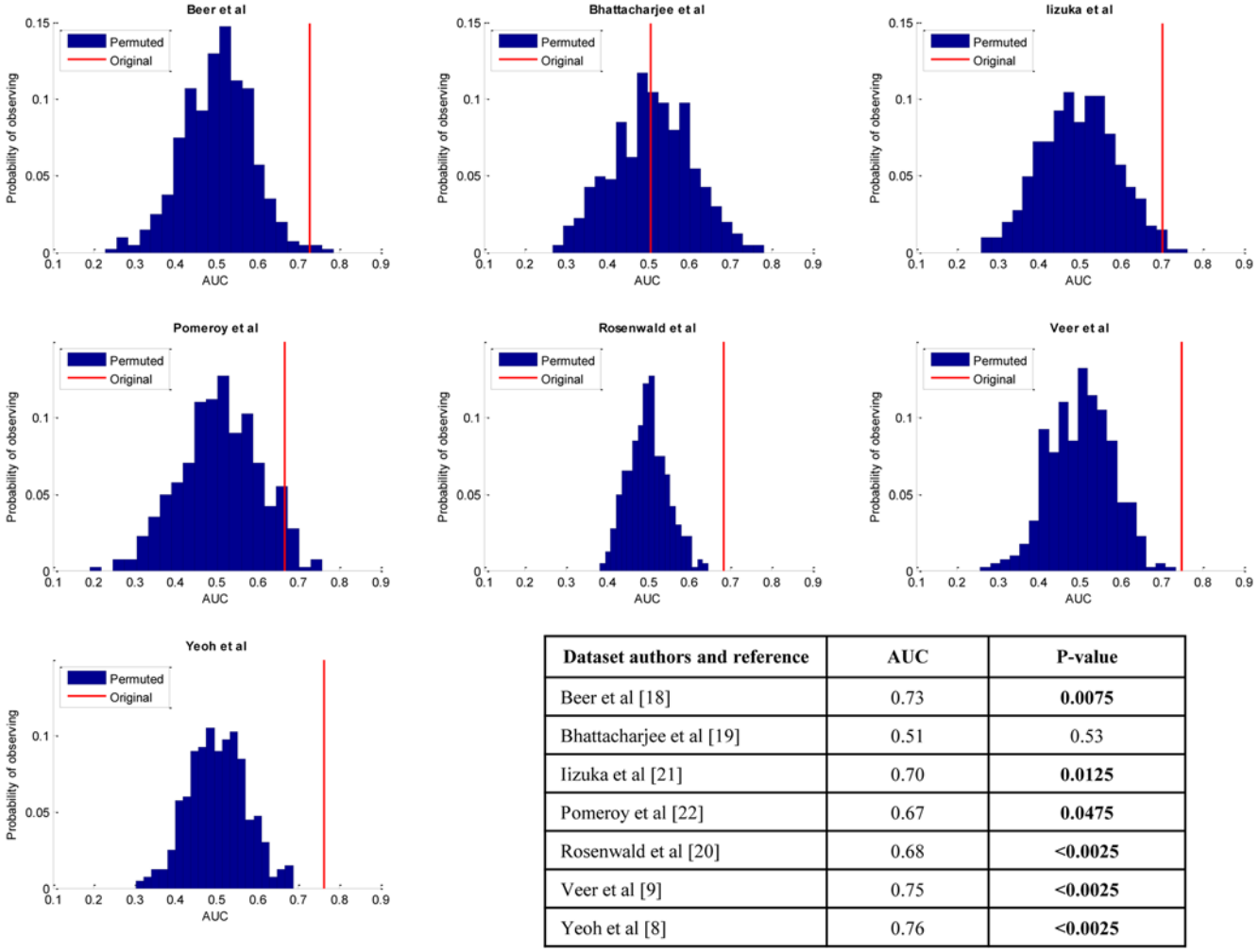
A note on the choice of null hypothesis for statistical significance testing. The combined simulated and real data results show very significant differences in the ability of Protocol I and II to detect real signal. This prompted us to investigate further the underlying differences between the two protocols. An unanticipated finding was that a major discrepancy between the two protocols exists in the precise null hypothesis tested: Ideally, one wishes to test the broad null hypothesis "there is no signal in the data". Rejecting this hypothesis entails that the observed signal in the sample will generalize in the population where the sample is drawn from. There exist several reasons why an observed signal may not be present in the population. First, the available sample may be non-representative of the population. Another reason is that a splitting procedure of the sample into training and testing parts may yield non-representative training or testing datasets (we will refer to this as "bad" split of the data). The previously published procedure for statistical significance testing of the signatures of Protocol I (see Materials and Methods section) eventually tests for only one source of non-reproducible signal: "bad" split of the sample data (and also conducts a sensitivity analysis on the training sample size). If this procedure instead was using sampling with replacement, it would amount to a simple Bootstrap estimator and thus would test for a non-representative sample as well. However because the Bootstrap introduces a bias in the error estimates that is difficult to correct, the above procedure samples without replacement and tests only for a restricted null hypothesis (i.e., "bad" data split). In contrast, the statistical significance procedure utilized by Protocol II (see Materials and Methods section) uses a repeated split cross-validation estimator effectively eliminating uncertainty introduced by non-representative splits. In addition by permuting labels, Protocol II effectively samples from a population where the gene expression patterns as well as the event rate are fixed, and there is no relationship between gene expression patterns and outcome (hence it is equivalent to the null hypothesis of no signal in the population). Under this label permuting any apparent relationship between gene expression patterns and outcomes is due to sampling variation. Thus Protocol II tests for a much more informative null hypothesis than the statistical test in Protocol I. Notice that the four factors affecting power we identified earlier affect both null hypotheses and have noteworthy effects on both protocols as shown in the simulation studies. The null hypothesis tested by the test of significance of Protocol I is too limited and redundant (i.e., as long as a repeated split cross-validation estimator is used) and should not be pursued in practice. However because of the broad implications previously drawn by applying Protocol I, it was necessary to test it in the present study in order to precisely identify the reasons why this protocol failed to establish signal in real microarray datasets.

Discussion

The present work shows that several important components of data analysis for molecular signature creation have significant and compounding effects on probability to detect true signal (i.e., statistical power). Four factors (choice of error metric, classifier, error estimator, and event balancing) were investigated by theoretical assessment, simulation study, and application to 7 human microarray datasets.

Our findings indicate that the choices made in the data analysis protocol corresponding to the four factors studied can improve power and by extension research efficiency. Increasing study sample size (as for example proposed by [26]) increases statistical power, but also dramatically increases study costs and delays study completion. In contrast, application of efficient statistical protocols has the potential to significantly improve the chances of detecting real signal with modest sample sizes. Conversely, even very large samples can be "wasted" when analyzed with under-powered (i.e., inefficient) data analysis procedures.

Our data also shows clearly that the highly-cited study [23] that concluded that "*Five of the seven largest published studies addressing cancer prognosis did not classify patients better than chance*" reached these conclusions because of two main reasons: first, the specific null hypothesis tested was inappropriate and second, because several



Dataset authors and reference	AUC	P-value
Beer et al [18]	0.73	**0.0075**
Bhattacharjee et al [19]	0.51	0.53
Iizuka et al [21]	0.70	**0.0125**
Pomeroy et al [22]	0.67	**0.0475**
Rosenwald et al [20]	0.68	**<0.0025**
Veer et al [9]	0.75	**<0.0025**
Yeoh et al [8]	0.76	**<0.0025**

Figure 2. Application of Protocol II to human microarray data. Each histogram is the distribution of the repeated 10-fold cross-validation AUC estimates for each dataset under the null hypothesis "there is no signal present in the data" (as computed by 400 random outcome value permutations). The red line in each graph is the observed value of AUC estimated by the repeated 10-fold cross-validation on the original data. AUC and p-values are shown for each dataset in the embedded table. Bold p-values indicate that the null hypothesis is rejected at the 0.05 level in these datasets.
doi:10.1371/journal.pone.0004922.g002

underpowered analysis components were employed. These two reasons were inextricably intertwined in the data analysis protocol employed. The ensuing controversy in the field of disease outcome prediction using microarrays seems thus to be an artifact of data analysis and not an intrinsic limitation of this assaying technology. The present findings therefore have direct positive implications for the feasibility of related research in new drug development, personalizing treatments and adapting clinical trials to patient genomic characteristics. Inappropriate data analysis methodology can create a climate of distrust about the underlying assay technology and findings that may lead to wasteful development processes. For example in our assessment, using a series of datasets for validation [39–42] likely wastes time and money with no substantial benefit. Validation using a single independent dataset from the same population of patients as used for construction of the signature is sufficient if the protocols used are unbiased and appropriately powered.

We note that it is possible that gene selection and better optimization/choice of classifiers could achieve predictivity and power improvements over the protocol used in the present paper [33]. For example, gene selection and error estimation using the more sophisticated but computationally more demanding nested cross-validation designs [43] was not pursued in order to keep the computational requirements of running extensive permutation tests under control.

We finally observe that the factors studied have been the subject of substantial prior research in biostatistics and bioinformatics. However their relationship to statistical power for molecular signature testing has not been systematically investigated previously. For example, recent work has proposed a much-needed and comprehensive set of guidelines for the analysis and reporting of microarray and other "omics" data [4]. However the choice of classifier is not addressed as of crucial importance, the choice of error metric and estimator is not linked to statistical power, and event balancing as a source of bias and low power is not addressed. These omissions demonstrate the subtle effects of these factors on statistical power and that these effects have gone largely unnoticed in the field so far.

In conclusion, factors that affect the statistical power of complex analysis protocols for molecular signature development from high-

throughput data constitute an important area for study. The present paper showed that choices of error metric, classifier, error estimator and event balancing have large and compounding effects on statistical power. They can further be combined with inappropriate null hypotheses to yield ineffective analysis protocols. An experimental comparison of data analysis protocols reveals that previous highly-cited claims that microarray assays may not be able to predict clinical outcomes better than chance are byproducts of data analysis limitations. Research designs of high-throughput studies will benefit by using the most powerful data analysis protocols available combined with appropriate statistical tests and doing so leads to substantial economies of required sample. New data analysis protocols should be tested for statistical efficiency before deploying for building molecular signatures. We recommend testing against existing protocols (such as one presented in this paper) in simulated or real data with known predictive signal using datasets in which the experimenter varies sample sizes [44,45].

Materials and Methods

Microarray Datasets

The characteristics of the human datasets analyzed [8,9,18–22] are summarized in Table 1.

Error / prediction performance metric

The area under the receiver operating characteristic (ROC) curve (AUC) is calculated by the formula provided in [46]. Proportion of misclassifications is calculated as the ratio: number of wrong classification divided by total number of classifications.

Classifiers & gene selection

Protocol I [23] involves selection of 50 genes with the highest correlation in training data with outcome variable according to Pearson's correlation coefficient. Then molecular signatures are developed based on these genes using a nearest-centroid prediction method [47].

Protocol II uses the LibSVM implementation of Support Vector Machines (SVMs) to build molecular signatures with a fixed misclassification penalty parameter $C = 100$, and a linear kernel [48]. Gene selection is not employed to avoid increased computational costs. We note that SVMs have built-in regularization however, which means that the learning algorithm penalizes large weights of predictors thus favoring simpler models by implicitly selecting genes, without using explicit gene selection procedures [49,50].

Statistical analysis

Statistical significance of the molecular signatures in Protocol I replicates the procedure of [23]. Namely, 500 training datasets of size n are obtained by sampling without replacement the original dataset (of size N) such that each training set has $n/2$ subjects with each outcome. For each training set, the testing set is defined as its complement (of size N-n). The molecular signatures are then fitted on the training sets and their classification performance is assessed on the corresponding testing sets. The above procedure is repeated for different training set sizes ranging from 10 to a maximum value which was chosen so that the testing set has at least one subject representing each outcome. Given a distribution of classification performances for each training set size, the corresponding 95% intervals are constructed. The original dataset is considered to contain predictive signal if the upper 95% interval limit is less than 0.5 proportion of misclassifications. Notice that the published description of this method [23] does not explicitly state whether the above condition for significance should hold in at least one training set size n, or all possible training set sizes, or the majority of them. Thus we examine all three possibilities in the present work.

For Protocol II, we use outcome value-permutation to test in each dataset the null hypothesis of no predictive signal [12,13]. This is also known as a randomization test or a Monte-Carlo permutation test. We construct the distribution corresponding to the null hypothesis by randomly permuting the values of the outcome variable (400 times) and then using SVMs (as described above) to compute the signature and repeated 10-fold cross-validation estimate of AUC for each permuted dataset. The repeated 10-fold cross-validation estimate from the original data is then compared to this distribution, and p-values correspond to the proportion of permuted estimators (under the null hypothesis) that are more extreme than the repeated 10-fold cross-validation estimate from the original (non-outcome value permuted) data.

Supporting Information

File S1 Comparison of proportion of misclassifications with area under ROC curve (AUC)
Found at: doi:10.1371/journal.pone.0004922.s001 (0.06 MB DOC)

File S2 Demonstration of pitfalls of non-balanced data
Found at: doi:10.1371/journal.pone.0004922.s002 (0.10 MB DOC)

File S3 Details of simulation experiments
Found at: doi:10.1371/journal.pone.0004922.s003 (0.08 MB DOC)

Table 1. Characteristics of gene expression microarray datasets analyzed in this study.

Dataset authors and reference	Sample size and number of events	Number of variables (genes)	Predicted event (outcome)
Beer et al [18]	**86** (24 events)	**7129**	Lung adenocarcinoma survival
Bhattacharjee et al [19]	**62** (31 events)	**12600**	Lung adenocarcinoma 4-year survival
Iizuka et al [21]	**60** (20 events)	**7070**	Hepatocellular carcinoma 1-year recurrence-free survival
Pomeroy et al [22]	**60** (21 events)	**7129**	Medulloblastoma survival
Rosenwald et al [20]	**240** (138 events)	**7399**	Non-Hodgkin lymphoma survival
Veer et al [9]	**97** (46 events)	**24188**	Breast cancer 5-year metastasis-free survival
Yeoh et al [8]	**233** (32 events)	**12240**	Acute lymphocytic leukemia relapse-free survival

doi:10.1371/journal.pone.0004922.t001

File S4 Confidence intervals for repeated 10-fold cross-validation AUC estimates
Found at: doi:10.1371/journal.pone.0004922.s004 (0.08 MB DOC)

File S5 Demonstration that Protocol II is not biased
Found at: doi:10.1371/journal.pone.0004922.s005 (0.09 MB DOC)

Author Contributions

Conceived and designed the experiments: CA AS IT JSS BS FH. Performed the experiments: AS JSS FH. Analyzed the data: CA AS IT JSS BS FH. Contributed reagents/materials/analysis tools: CA AS IT JSS BS FH. Wrote the paper: CA AS IT JSS BS FH.

References

1. Butte A (2002) The use and analysis of microarray data. Nat Rev Drug Discov 1: 951–960.
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531–537.
3. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 98: 15149–15154.
4. Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99: 147–157.
5. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, et al. (2006) Genomic signatures to guide the use of chemotherapeutics. Nat Med 12: 1294–1300.
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, et al. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4: 1293–1301.
7. Burczynski ME, Peterson RL, Twine NC, Zuberek KA, Brodeur BJ, et al. (2006) Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. J Mol Diagn 8: 51–61.
8. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1: 133–143.
9. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
10. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, et al. (2005) Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2: 351–356.
11. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24: 1151–1161.
12. Mukherjee S, Golland P, Panchenko D (2003) Permutation tests for classification. MIT AI Memo, 2003-019.
13. Radmacher MD, McShane LM, Simon R (2002) A paradigm for class prediction using gene expression profiles. J Comput Biol 9: 505–511.
14. Good PI (2000) Permutation tests: a practical guide to resampling methods for testing hypotheses. New York: Springer.
15. Lehmann EL, Stein C (1949) On the Theory of Some Non-Parametric Hypotheses. The Annals of Mathematical Statistics 20: 28–45.
16. Lee ML, Whitmore GA (2002) Power and sample size for DNA microarray studies. Stat Med 21: 3543–3570.
17. Baldi P, Hatfield GW (2002) DNA microarrays and gene expression. Cambridge, UK: Cambridge University Press.
18. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, et al. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8: 816–824.
19. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, et al. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 98: 13790–13795.
20. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346: 1937–1947.
21. Iizuka N, Oka M, Yamada-Okabe H, Nishida M, Maeda Y, et al. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet 361: 923–929.
22. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, et al. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415: 436–442.
23. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365: 488–492.
24. Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, et al. (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7: 278.
25. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103: 5923–5928.
26. Ioannidis JP (2005) Microarrays and molecular research: noise discovery? Lancet 365: 454–455.
27. Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361–387.
28. Freidlin B, Gastwirth JL (2000) Should the Median Test be Retired from General Use? The American Statistician 54: 161–164.
29. Ling CX, Huang J, Zhang H (2003) AUC: a statistically consistent and more discriminating measure than accuracy. Proceedings of the Eighteenth International Joint Conference of Artificial Intelligence (IJCAI).
30. Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: data mining, inference, and prediction. New York: Springer.
31. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21: 631–643.
32. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, et al. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16: 906–914.
33. Statnikov A, Wang L, Aliferis CF (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9: 319.
34. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI) 2: 1137–1145.
35. Casella G, Berger RL (2002) Statistical inference. Australia: Thomson Learning.
36. Braga-Neto UM, Dougherty ER (2004) Is cross-validation valid for small-sample microarray classification? Bioinformatics 20: 374–380.
37. Saerens M, Latinne P, Decaestecker C (2001) Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. Neural Computation 14: 21–41.
38. Jiang W, Simon R (2007) A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Stat Med 26: 5320–5334.
39. Habel LA, Shak S, Jacobs MK, Capra A, Alexander C, et al. (2006) A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. Breast Cancer Res 8: R25.
40. Paik S, Shak S, Tang G, Kim C, Baker J, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817–2826.
41. Paik S, Tang G, Shak S, Kim C, Baker J, et al. (2006) Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 24: 3726–3734.
42. Sparano JA, Paik S (2008) Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin Oncol 26: 721–728.
43. Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF (2005) GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. Int J Med Inform 74: 491–503.
44. Dobbin KK, Zhao Y, Simon RM (2008) How large a training set is needed to develop a classifier for microarray data? Clin Cancer Res 14: 108–114.
45. Dobbin KK, Simon RM (2007) Sample size planning for developing classifiers using high-dimensional DNA microarray data. Biostatistics 8: 101–117.
46. Hand DJ, Till RJ (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45: 171–186.
47. Simon R (2003) Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. Br J Cancer 89: 1599–1604.
48. Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. Journal of Machine Learning Research 6: 1918.
49. Vapnik VN (1998) Statistical learning theory. New York: Wiley.
50. Aliferis CF, Statnikov A, Tsamardinos I (2006) Challenges in the analysis of mass-throughput data: a technical commentary from the statistical machine learning perspective. Cancer Informatics 2: 133–162.