The MSPress Journal
Vol 4 No 1 2017

Essay

Improper Statistical Analysis:
A Cause of Poor Translation of New Biomarkers into Clinical Practice

Justin Barnes
Saint Louis University School of Medicine

Introduction

A biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention."[1] The use of biomarkers in detecting or diagnosing disease and monitoring a body's response to therapy is a fundamental aspect of modern medical practice. For example, diabetes can be diagnosed with serum glucose levels, cardiovascular events are indicated by serum cardiac troponin, and pregnancy can be detected early using human choriogonadotropin. With an unprecedented amount of data—notably omics (genomic, proteomic, lipidomic, metabolomic, etc.) data—and the potential to provide more accurate, less expensive, less invasive, and earlier diagnoses, it is no wonder that the biomedical community has dedicated many efforts to discover and validate new markers.

However, very few of the newly published "biomarkers" are utilized in clinical settings. For example, in 2011, there were 7,720 publications of new biomarkers but only 407 new biomarker patents.[2] Furthermore, despite thousands of publications regarding novel cancer biomarkers, very few markers have received FDA approval or clearance.[3] And before the approval of those markers within the past few years, there had been no new biomarkers approved for clinical oncology use for over 25 years.[4] While one setback is the amount of time required to obtain FDA approval, the primary reason for the lacking utilization of newly published biomarkers is their failure to perform properly in the clinical setting. For example, a multi-marker diagnostic panel for ovarian carcinoma had a published sensitivity of 95% and specificity of 99.4%.[5] However, when the Early Detection Research Network tested the markers, they found that in contrast to the excellent performance in the authors' data, the "true" sensitivity of the 6 markers was only 52%—approximately as "good" as a coin toss.[4] Of many issues leading to the authors' optimistic results, perhaps the greatest factor was an improper statistical analysis due to a biased external validation procedure.

While the reasons for the failure of the discovered biomarkers include problems in sample collection (from biased selection of samples to improper care or storage of specimens) and sample analysis (such as using incorrect or subpar methods to prepare the specimens, machine settings to obtain data, or steps to process the data), one of the main and perhaps most misleading issues is improper statistical analysis.[5] Several specific problems occur in biomarker discovery and predictive modeling—multiple hypothesis testing, data

overfitting, and inappropriate validation.[6] In this report, illustrations of these problems and biomarker-appropriate statistical methods to overcome such problems are presented.

Before proceeding, it is helpful to consider the two major approaches to biomarker discovery.[7] The "classical" approach involves testing each potential marker individually to test its association with the selected outcome/response (diagnosis, prognosis, etc.). One common approach is using the t-test to identify differences between groups. The importance of the new biomarker is directly related to the significance of the corresponding test. And the final "product" of the classical approach is one marker, the measured values of which are predictive of the response of interest. While the classical method is typically robust, it is subject to problems arising from multiple hypothesis testing and ignores multifactorial relationships among markers, which may be extremely important in biological systems.[7] The second approach, which I will call the multivariable approach, accounts for and includes multifactorial relationships. The multivariable approach typically involves a statistical model (logistic regression, partial least squares, etc.) or machine learning algorithm (support vector machine, neural network, etc.) that has been specially adapted to find and utilize only the most important markers. The importance of the set of markers is often determined by the performance of the model (measured by sensitivity, specificity, an ROC curve, etc.) rather than by significance levels. Furthermore, the product of the multivariable approach is a panel of markers, whose values are combined in some way to arrive at a score (such as the probability of disease for a diagnostic marker, a set of probabilities for a set of disease stages for a monitoring biomarker, or a predictive indicator of longevity for a prognostic marker). While this approach often suffers from a variety of challenges, such as overfitting and improper validation, there are continuous advancements in computational technology and statistical methodology to improve the models and algorithms.

Multiple Hypothesis Testing

As the name suggests, multiple hypothesis testing is performing a large number of statistical tests to look for some kind of statistically significant relationship. For instance, subjecting each variable/biomarker to a Student's t-test is one method of multiple hypothesis testing. However, when one finds a seemingly significant result (i.e. $p < 0.05$ or another selected threshold) using this method, it is very likely that the result is a false-positive. In fact, with a p-value cutoff of 0.05, we expect that 5% of completely useless markers will have significant results!

One of the most widespread approaches—perhaps a current standard of biomarker research—is the inclusion of a validation study. This approach involves collecting 2 samples; biomarkers are discovered in the first set of data, and if they also exhibit significant relationships in the validation set, they are deemed important. This method is extremely effective at reducing false positives, as the chances of useless marker being found significant in both samples of data is very small. Assuming a p-value cutoff of 0.05, it is expected that only 0.25% of markers would be found significant. However, this can still pose a problem when the number of markers examined is very large, as is the case in many biomarker studies. Thus, some researchers use a much lower p-value threshold (such as 0.001) to account for the problem of multiple hypothesis testing. However, while the lower threshold does reduce the number of false positives, the method is not stringent enough to properly control the number of false positives identified.[8]

One of the more popular methods to properly account for multiple hypothesis testing is using a false-discovery-rate (FDR) adjustment, which controls the number of false-positives discovered. Mathematically, it is an adjustment of the p-value to make it a larger "q-value." Like the p-value, the q-value is also an indication of significance. However, the q-value cutoff used is the expected proportion of false positive markers among all markers with a significant q-value. Note the key difference between the implications of the p-value and FDR adjustment: while we expect that 5% of all markers

tested will be incorrectly labeled as significant with a p-value less than 0.05, only 5% of markers with a q-value less than 0.05 will have been incorrectly labeled.

Suppose we simulate a study examining 10,000 gene expression markers in 200 patients (100 controls and 100 with cancer), where only 5 of the markers have a real relationship with cancer. Figure 1 shows the number of false positives from applying Student t-tests to the data. Using a p-value cutoff of 0.05, 507 markers are falsely identified as significant. Using a lower p-value threshold, 0.001, provides better results—only 8 false positive markers. However, using the FDR adjusted p-values with a q-value cutoff of 0.05, no false positives are identified. Also, note that when using the validation set of data to find markers that continue to be significant, the number of false-positives decreases dramatically—only 28 with a p-value cutoff of 0.05 (still too many!) and none with a p-value cutoff of 0.001.

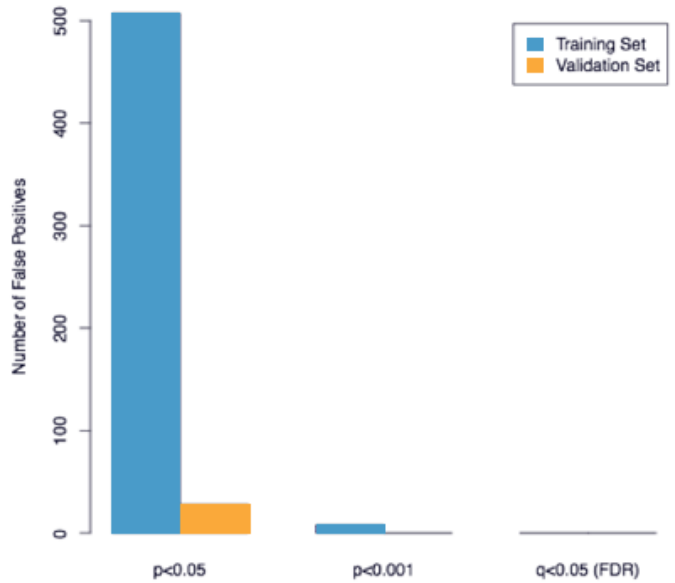

Figure 1: Number of False Positives from Student t-tests of 10,00 Markers Using Various Significance Criteria

Figure 2 shows the proportion of true positives (# true positives / # of markers with significant results) identified by the various methods. In every case, the 5 "true" markers were identified, but note the persisting problem of false positives. However, with the FDR approach (both in the first study and the validation study) and the 0.001 p-value cutoff approach (only with the validation study) identified only correct markers (all 5, in this case). Hence, it is apparent that if efforts are to be focused on true markers, it is imperative to appropriately account for the problem of multiple hypothesis testing. Simply performing a validation study (while continuing to have no p-value cutoff adjustment from the classic 0.05) or lowering the p-value threshold likely won't be sufficient. Identifying studies with potentially misleading conclusions based on multiple hypothesis testing should be easy to identify, since authors should list their criteria for claiming a marker is statistically significant. In general, be wary of any result that is based on p-values and be suspicious of results with an abnormally large number of significant markers.

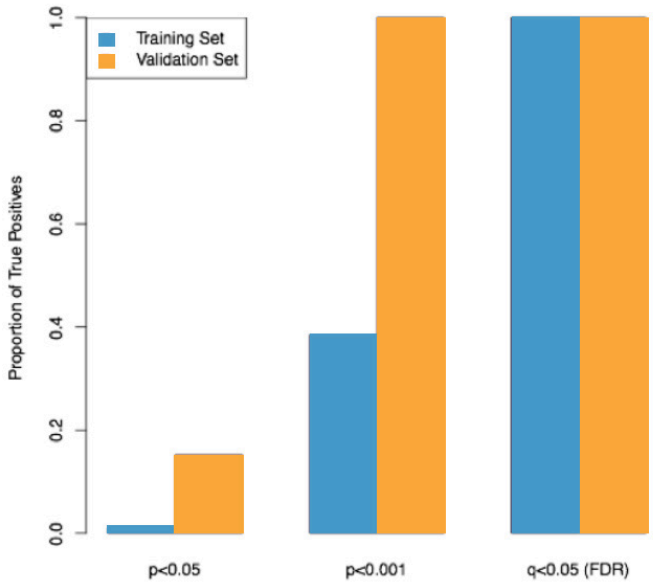

Figure 2: Proportion of True Positives from Student t-tests of 10,000 Markers Using Various Significance Criteria

Data Overfitting

Data overfitting occurs when a model or classifier has incorporated information that is irrelevant to the outcome of interest and is characteristic only to a particular sample of data. Due to the large number of markers, some markers may have a seemingly important relationship with the outcome variable, and an overfit model is one that includes this information. Typically, if there are fewer than 10-

20 samples per variable, the model will be overfit.[9] While the model may perform exceptionally well on the sample of data it was developed on, when applied to other sets of data (which will likely not contain the same falsely important relationships), it will perform worse. Perhaps the biggest issue in biomarker research regarding overfitting is the large discrepancy between reported performance and true performance, if the reported performance was estimated by applying the model to the same data used to train the model.

Differences in Apparent AUC and True AUC Estimates

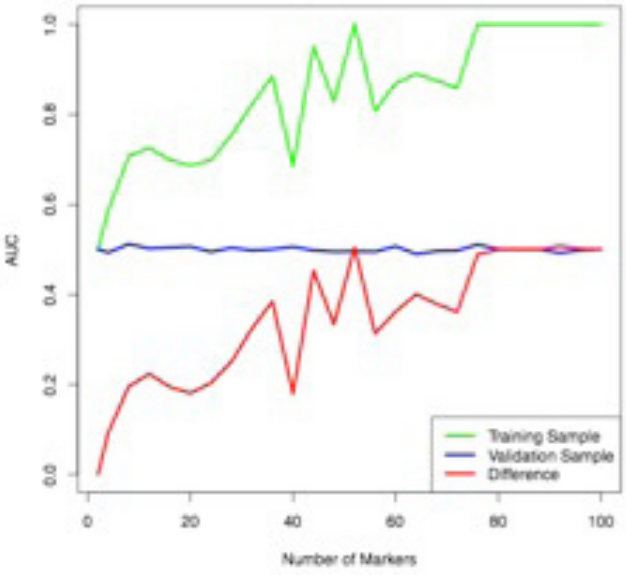

Figure 3 (a): Stepwise Logistic Regression of Data Without Important Markers

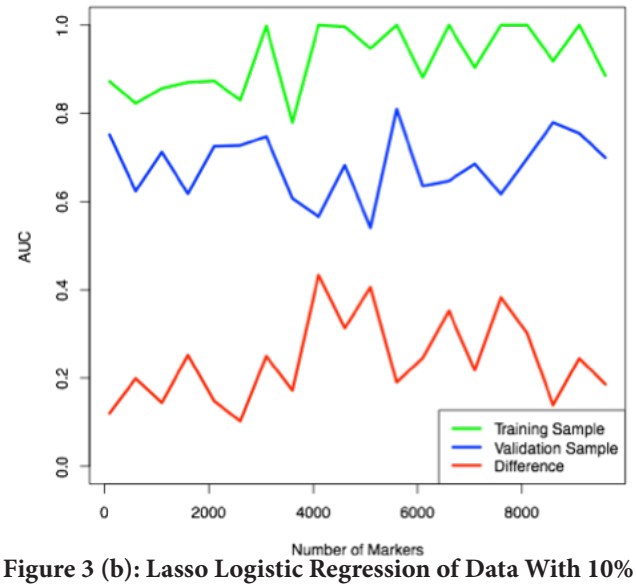

Figure 3 (b): Lasso Logistic Regression of Data With 10% Important Markers

Figure 3 shows the drastic differences in area under the curve estimate (AUC; a measure of model accuracy) between the apparent performance in the training set and the true performance in a validation study for increasing numbers of markers. In Figure 3 (a), a stepwise logistic regression model, which does not control for overfitting, suffers more from overfitting with increasing numbers of markers.

While the lasso logistic regression model demonstrated in Figure 3 (b) is designed to account for overfitting, notice that the apparent performance is always better than the true performance. Thus, it is critical to be wary of claims of high performance; even worthless models can have near-perfect performance in the training sample! The only way to get an accurate estimate of a model's performance is through proper validation. Therefore, also be wary of results that do not explicitly demonstrate results from a validation study.

Model Validation

Since the value of a multivariate panel of markers is largely related to its predictive performance, it is imperative to obtain accurate estimates of model performance. The optimal method is an independent and external validation, where another study is performed by other researchers to test the proposed panel of markers. The results of such an analysis should be indicative of the model's true performance. A common method when an external validation is not yet available is where researchers arbitrarily divide their initial sample into two—one becomes the sample on which to develop the model, and the other is the sample on which the model is tested to provide performance estimates (a "synthetic" validation study). However, this method is inefficient unless there are thousands of subjects. A proper implementation of internal validation, such as cross-validation, is much better (though still less ideal than an independent validation).[10]

Unfortunately, there are a variety of ways to invalidate the validation process, usually resulting in biased and optimistic estimates of model perfor-

mance. Independent (often called "external") validation is the testing of a finalized model on a new population to assess its predictive capacity. Any variation from this straightforward procedure will invalidate the results. For example, performance estimates from validation using an independent study are likely biased and optimistic if: 1) the independent set of data is accessed or utilized in any form other than testing the finalized model, 2) the model performance results from the independent set of data inform the selection of markers, statistical methodology, or model parameters, or 3) the independent study is not in fact independent (i.e. if any of the subjects are utilized in both studies). Cross-validation involves splitting the data into a training and validation samples, following the steps of proper independent validation, and then iteratively repeating the data splitting process so that each subject receives a predicted outcome based on a model that was trained on data that did not include that subject. Thus, every sample is included (at separate times/iterations) in both training and validation samples. Like external validation however, any variation from the procedure will invalidate the results. For example, performance estimates from cross-validation are likely biased and optimistic when every stage of the model training procedure (including method selection, marker selection, parameter optimization, and any step that directly or indirectly relies on information from the response variable) is not repeated in each iteration of the cross-validation algorithm.

Figure 4 demonstrates the effects of incorrect validation procedures on the AUC. The grey bar represents the AUC estimate from a correctly performed independent validation (with 10,000 samples to ensure accuracy), and the orange bar demonstrates the erroneous optimism included in the AUC estimate from incorrect validation (either independent validation or cross-validation). Note that inappropriate validation leads to biased and optimistic results. In case 1, the assumption of an "independent" validation was violated, as the subjects from both the training and validation sets were pooled initially to identify the top markers by using a t-test. Following that procedure, a lasso logistic regression model was trained using the training set and only the selected markers (which were, in part, chosen by the validation set!). Notice how poorly the lasso logistic regression model performed on a truly independent set of data. In case 2, the cross-validation procedure was performed incorrectly since the parameters for a sparse partial least squares discriminant analysis (sPLS-DA) model were optimized initially, and the optimal values (likely just optimal for the given set of data!) were fixed during the cross-validation procedure, rather than optimized at every iteration. In case 3, the top markers were selected initially by a t-test, and only those markers were included in the cross-validation procedure; the cross-validation procedure should have included the marker selection step in every iteration of the procedure. Thus, for the sake of accurate estimates of performance, it is crucial that validation is done correctly. Unfortunately, it is nearly impossible to detect errors in others' validation studies.

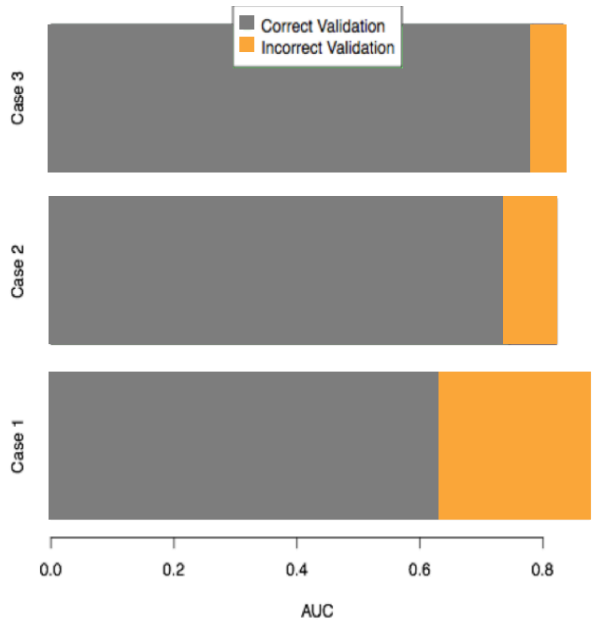

Figure 4: Difference Between AUC in Correct and Incorrect Validation Procedures

Case 1: Violation of Independent External Validation
Case 2: Violation of Repeated Parameter Optimization in Cross-Validation
Case 3: Violation of Repeated Marker Selection in Cross Validation

Conclusion

Biomarkers have the potential to greatly enhance care provided to patients—through providing better diagnostic and prognostic information, both of which can inform decision making and be correlated to important patient outcomes. Yet, despite dedicated efforts, very few new biomarkers are being introduced in clinical settings. While their initial performance may seem promising, real-life implementation is often lacking. Why such a disparity? In large part, the inappropriate use and/or understanding of statistical methods for biomarker research can result in biased and optimistic estimates of diagnostic/prognostic performance or overstatements on the importance of a certain biomarker. It is thus crucial for the rising generation of physicians and scientists to understand such issues, recognize and correct mistakes, and help bring about the changes necessary so effective biomarkers can be discovered and translated into clinical practice.

References

1. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical Pharmacology and Therapeutics. 2001; 69(3): 89-95.

2. Drucker E, Krapfenbauer K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA J. 2013; 4(1): 7.

3. Li D, Chan DW. Proteomic cancer biomarkers from discovery to approval: it's worth the effort. Expert review of proteomics. 2014; 11(2): 135-136.

4. Diamandis EP. Cancer biomarkers: can we turn recent failures into success? Journal of the National Cancer Institute. 2010; 102: 1-6.

5. Visintin I, Feng Z, Longton G, et al. Diagnostic markers for early detection of ovarian cancer. Clinical Cancer Research. 2008;14(4): 1065-72.

6. Pavlou MP, Diamandis EP, Blasutig IM. The long journey of cancer biomarkers from the bench to the clinic. Clinical chemistry.2013; 59(1): 147-157.

7. Zhang Z, Chan DW. Cancer proteomics: in pursuit of "true" biomarker discovery. Cancer Epidemiology Biomarkers & Prevention. 2005; 14(10): 2283-6.

8. Lindquist M. Statistical Methods in functional MRI: Multiple Comparisons. Talk presented at: John Hopkins Bloomberg School of Public Health; April 23, 2013; Baltimore, MD. http://www.stat.columbia.edu/~martin/Tools/Lec7-MultipleComparisons.pdf. Accessed November 25, 2016.

9. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996; 15(4): 361-87.

10. Harrell FE. Statistical Methods and Statistical Pitfalls in Biomarker Research. Talk presented at: Vanderbilt University Biomarker Research Summit; June 22, 2007; Nashville, TN. https://www.researchgate.net/profile/Frank_Harrell/publication/237423804_Statistical_Methods_and_Statistical_Pitfalls_in_Biomarker_Research/links/551a8c360cf2f51a6fea6077.pdf. Accessed December 11, 2015.