An empirical assessment of validation practices for molecular classifiers

Peter J. Castaldi, Issa J. Dahabreh and John P.A. Ioannidis
Submitted: 1st September 2010; Received (in revised form): 2nd November 2010

Abstract
Proposed molecular classifiers may be overfit to idiosyncrasies of noisy genomic and proteomic data. Cross-validation methods are often used to obtain estimates of classification accuracy, but both simulations and case studies suggest that, when inappropriate methods are used, bias may ensue. Bias can be bypassed and generalizability can be tested by external (independent) validation. We evaluated 35 studies that have reported on external validation of a molecular classifier. We extracted information on study design and methodological features, and compared the performance of molecular classifiers in internal cross-validation versus external validation for 28 studies where both had been performed. We demonstrate that the majority of studies pursued cross-validation practices that are likely to overestimate classifier performance. Most studies were markedly underpowered to detect a 20% decrease in sensitivity or specificity between internal cross-validation and external validation [median power was 36% (IQR, 21–61%) and 29% (IQR, 15–65%), respectively]. The median reported classification performance for sensitivity and specificity was 94% and 98%, respectively, in cross-validation and 88% and 81% for independent validation. The relative diagnostic odds ratio was 3.26 (95% CI 2.04–5.21) for cross-validation versus independent validation. Finally, we reviewed all studies (n = 758) which cited those in our study sample, and identified only one instance of additional subsequent independent validation of these classifiers. In conclusion, these results document that many cross-validation practices employed in the literature are potentially biased and genuine progress in this field will require adoption of routine external validation of molecular classifiers, preferably in much larger studies than in current practice.

Keywords: predictive medicine; genes; gene expression; proteomics

INTRODUCTION
The advent of high-throughput molecular technologies has led some to hypothesize that comprehensive assessment of the genome, the transcriptome and the proteome may lead to the discovery of new molecular classifiers capable of classifying patients more accurately than existing traditional prognostic factors and biomarkers [1, 2]. This has led to a corresponding bounty of novel molecular classifiers for which high levels of predictive performance have been claimed, some of which have even received approval for clinical use [3]. However, the development of classifiers from high-dimensional data is a complex process involving multiple analytic decisions [4], each of which is subject to a number of potential methodological errors which may result in spuriously high levels of reported performance [5, 6]. Amidst the current proliferation of molecular classifiers, it is difficult to know whether novel classifiers with impressive performance represent true discoveries or spurious
patterns of ‘noise discovery’ resulting from improper analysis of high-dimensional data [7–9].

Proper internal and external validation is essential for assessing the performance and generalizability of a classifier [10, 11]. However, the validation process itself can be susceptible to inappropriate application, and multiple authors have pointed to common misapplication of validation practices in the literature [10, 12–16]. Simple (re-substitution) analysis of a training set is well known to give biased (inflated) estimates of classification accuracy and overfitting can occasionally be severe [17]. To avoid overfitting, several internal validation methods have been widely used in the literature, including leave-one out and $k$-fold cross-validation, as well as the related family of bootstrap-based methods. However, several authors have demonstrated that inappropriate application of cross-validation leads to inflated estimates of classification accuracy [10, 14–20]. Various specific sources of bias have been described, with an accumulating literature using diverse terminology on ‘population selection bias’, ‘incomplete cross-validation’, ‘optimization bias’, ‘reporting bias’ and ‘selection of parameters bias’, among others [15–19]. There is increasing appreciation that proper and transparent design and reporting of these studies is important [8, 10, 21]. Moreover, it is also increasingly appreciated that it is essential to perform external validation of any classifier in external, independent data from those where it was developed [10, 20]. External validation may be performed by the team that developed the classifier or independent teams. Many of the biases that affect internal validation are addressed with stringent external validation, and the generalizability of the classifier becomes more robustly documented when the exercise is repeated successfully across multiple research teams and patient populations.

In this article, we assess the state of contemporary validation practices for novel molecular classifiers, with special focus on the use of external validation and its comparison against internal cross-validation. We have conducted an empirical assessment of a sample of studies which claim to develop and independently validate a novel molecular classifier in the same publication. We assessed the internal and external validation methods employed, and we obtained data relating to the classification accuracy of these classifiers as estimated from internal cross-validation, external validation from the same publication and any subsequent external validation efforts by the same, and different, research teams in other papers.

**METHODS**

**Literature search and eligibility criteria**

To obtain a sample of published studies reporting on the external validation of molecular classifiers, we performed a search of the MEDLINE database (through PubMed) on 23 June 2010 using the following terms: (molecular OR gene OR expression OR profile OR proteomic* OR metabolomic* OR microarray) AND (independent validation[tw] OR external validation[tw]). A single reviewer (PC) screened all abstracts and potentially eligible studies were retrieved in full text. Two reviewers (PC and IJD) reviewed all potentially eligible studies independently and discrepancies were resolved by consensus including a third reviewer (JPAI).

We set the following inclusion criteria: the article used molecular information to make binary classifications of human subjects into categories of interest; the molecular classifiers were not already well-established for clinical use for the condition of interest; external validation was performed in the paper; and classification results for internal and external validation were available in the form of sensitivity and specificity, or we could calculate these metrics based on other presented information. We did not consider animal studies, reviews, editorials, letters or other studies not reporting primary research findings or studies not published in English.

From each eligible study, we extracted the following information: first author and year of publication, categories into which subjects were being classified, the number of subjects in training and validation samples, whether these samples were independent, whether there were any gross differences between training and validation samples, whether the samples originated from a single or multiple centers, the technology used to identify molecules of interest (i.e. genotyping arrays, gene expression arrays, proteomic technologies), methods used for cross-validation, whether cross-validation results were averaged across all validations or represented the best performing classifier identified at the cross-validation phase, whether feature/variable selection was done so as to preserve the integrity of cross-validation test sets, whether multiple
models were tested in the external validation data and the reported sensitivity and specificity of the classifier in cross-validation and independent validation.

We use here the term cross-validation to describe all variants thereof (leave-one out, k-fold, etc.) as well as bootstrap-based methods that can serve a similar purpose of estimating classification accuracy while avoiding over-fitting without using additional external samples. Conversely, we use the term external validation to describe procedures that determine predictive performance of a model in data that were not used in any aspect of the development of the model.

**Power calculations for validation samples**

To evaluate whether the independent validation phase in each study was sufficiently designed to detect a meaningful difference in sensitivity or specificity compared to cross-validation estimates (whenever cross-validation had been performed), we estimated the power of the independent validation phase in each study to detect a 20% decrease in sensitivity or specificity compared to cross-validation estimates. To test this, we estimated the power of the independent validation (whenever cross-validation had been performed), and then calculated the relative DOR (rDOR), i.e. the ratio of the two DORs for each study [24–26]. DORcv was set to be higher than one for all studies and rDOR values higher than one indicate that the classifier of interest performed worse in the external validation phase. The standard error of each ln(DOR) was calculated as

\[ \sqrt{\frac{1}{TP} + \frac{1}{FP} + \frac{1}{FN} + \frac{1}{TN}} \]

with a continuity correction of \( k = 0.5 \) in studies where one of the denominators was zero. We used the sum of the variances of \( \ln(DOR_{cv}) \) and \( \ln(DOR_{iv}) \) to estimate the variance of the \( \ln(rDOR) \). Then, we obtained the 95% CI for the \( \ln(DOR) \) as

\[ \ln(rDOR) \pm 1.96 \times \sqrt{\text{var}(\ln(DOR_{cv})) + \text{var}(\ln(DOR_{iv}))} \]

and reported the back-transformed (exponentiated) values. When the two DORs are calculated in samples from different individuals their covariance is zero, i.e. they are independent [25]. For reasons similar to those discussed in the power calculations section, the var\( (\ln(DOR_{cv})) \) estimates we obtained are lower than the ‘true’ variance values that account for the correlation pattern in cross-validation [27]. Therefore, to explore the effect of underestimating the var\( (\ln(DOR_{cv})) \), we performed a sensitivity analysis where we doubled the standard error of the ln(DORcv) and the summary results were similar (data not shown).

To summarize the results across all studies, we used the rDORs and their variance to perform meta-analysis using both fixed and random effect models [28]. Between-study heterogeneity was assessed using Cochran’s \( \chi^2 \)-based Q-statistic and the \( I^2 \) index [29, 30]. Heterogeneity was considered statistically significant at \( P_Q < 0.10 \). \( I^2 \) ranges between 0 and 100% and expresses the percentage of between study heterogeneity that is not explainable by chance. We also estimated the 95% CI of \( I^2 \) estimates [29, 31].

In order for cross-validation to provide minimally biased estimates of prediction error, all aspects of the model-building process need to be cross-validated to
maintain the integrity of the process, and the reported result should be the average of all cross-validations whenever multiple cross-validations have been performed. Feature selection bias occurs when feature or variable selection takes place outside the cross-validation loop, and optimization bias occurs when the best, rather than the average, cross-validation result is reported. Cross-validation performance may be inflated when either feature selection bias \[16, 19\] or optimization bias \[10, 14, 19, 20\] is present. We categorized studies based on whether either of these biases appeared to be present, and we performed subgroup analyses comparing the summary rDOR in such studies versus the summary rDOR in studies that do not suffer such biases.

We also explored random effects meta-regression of the rDOR over the following variables: year of publication, sample size of cross-validation sample (log-transformed), sample size of external validation sample (log-transformed), whether samples in the training phase originated from multiple centers, whether samples in the independent validation phase came from multiple centers and whether clear differences between training and validation samples were present \[32, 33\]. These analyses are entirely exploratory and should be interpreted with caution given the limited number of studies.

Statistical analyses were conducted using Stata version SE/11.1 (Stata Corp., College Station, TX, USA) and R version 2.11 (R Foundation for Statistical Computing, Vienna, Austria) \[34, 35\].

### RESULTS

#### Eligible studies analyzed

The PubMed search yielded 285 citations. After abstract level review, 92 articles were considered potentially eligible and were examined in full-text. Of these articles, 57 were excluded for the following reasons: 29 did not report classification results as sensitivity and specificity (e.g. they reported percentage overall accuracy or only presented ROC curves without selecting a specific operating point on them), 15 reported on molecular classifiers that were not binary, 6 did not pertain to molecular classifiers, 5 did not report primary data, 1 had no independent validation sample, and 1 did not pertain to humans. This left 35 articles that met the inclusion criteria and were included in the analyses \[37–71\].

#### Study characteristics and validation practices

Table 1 describes the characteristics of the included studies. The vast majority studies used gene expression or proteomic data, as would be expected from our search strategy. Most of the published classifiers pertained to cancer or cancer-related outcomes. The majority of studies were single center studies, and the median total sample size was 119 (IQR 62–192), with a median training/validation ratio of 1.3. The ratio of classification groups (i.e. ‘cases’ defined as the group with a diagnosis or outcome of interest where model sensitivity was determined, and ‘controls’ defined as the comparison group where model specificity was determined) was roughly 1:1. The median reported classification performance for sensitivity and specificity was 94 and 98%, respectively, in cross-validation and 88 and 81% in independent validation, already suggesting a decrease in performance from cross-validation to independent validation.

Table 1 also presents data relating to other critical issues relating to study design and data analysis issues relating to classifier development. All of the included studies claimed independence of the training and external validation samples with no discernible overlap between the two. For the validation of a classifier, it is essential that the classifier tested in independent validation be identical to that developed in training.

### Subsequent validation efforts

In order to determine if subsequent validation efforts had been undertaken for any of the classifiers included in our study, we searched the ISI Web of Science database to identify all the research papers that cited the 35 included studies \[36\]. Timespan was ‘All Years’ and retrieved citations were limited to those classified as ‘Articles’ by ISI Web of Science. These papers were then reviewed at the abstract level in order to identify whether they reported on subsequent external validations of any of the classifiers we assessed. Articles that appeared to be validation efforts were examined in full text to identify whether the validation attempt used the exact classifier from the original publication, whether there was any overlap in authorship between the two papers (i.e. same or different research teams), and to extract the subsequent validation sample size and quantify the performance of the classifier in subsequent external validation.

#### Study characteristics and validation practices

Table 1 describes the characteristics of the included studies. The vast majority studies used gene expression or proteomic data, as would be expected from our search strategy. Most of the published classifiers pertaining to cancer or cancer-related outcomes. The majority of studies were single center studies, and the median total sample size was 119 (IQR 62–192), with a median training/validation ratio of 1.3. The ratio of classification groups (i.e. ‘cases’ defined as the group with a diagnosis or outcome of interest where model sensitivity was determined, and ‘controls’ defined as the comparison group where model specificity was determined) was roughly 1:1. The median reported classification performance for sensitivity and specificity was 94 and 98%, respectively, in cross-validation and 88 and 81% in independent validation, already suggesting a decrease in performance from cross-validation to independent validation.

Table 1 also presents data relating to other critical issues relating to study design and data analysis issues relating to classifier development. All of the included studies claimed independence of the training and external validation samples with no discernible overlap between the two. For the validation of a classifier, it is essential that the classifier tested in independent validation be identical to that developed in training.
(e.g. regression coefficients should be fixed and there should be no re-estimation of parameters or addition of features used for classification). In all but one study, it appeared clear that this principle had been maintained. It is common to develop many different classifiers on the training data; however, since there is often wide uncertainty around the performance estimates for classifiers, selecting the best classifier among many is likely to introduce a form of optimization bias. External validation can guard against this bias, but if many classifiers are moved forward to independent validation then independent validation may be subject to this bias as well. Ten (29%) of the included studies tested more than 1 classifier in the independent validation sample. Of these studies, six reported the results for all classifiers tested in independent validation.

Since the performance of classifiers may be dependent on the population to which the classifier is applied, training and validation samples should be similar, unless the authors intend to extend the applicability of a previously established classifier. Nearly 40% of the included studies had a clear difference between training and validation samples. Notable examples of such differences included discrepancies in the distribution of cancer stages between training and validation samples [45], differences in baseline characteristics (i.e. 68% males in healthy control group in training samples, 31% males in control group in validation) [51] or differences in sample specimen quality (in [71] 2% of gene expression arrays failed quality control in samples from one center, compared to 70% of samples from a second center; in classifier development, the training sample

### Table 1: Characteristics of the 35 evaluated studies

<table>
<thead>
<tr>
<th>Characteristic, all included studies</th>
<th>Included studies (n = 35)</th>
<th>Characteristic, subset of studies using cross-validation</th>
<th>Included studies (n = 28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sample size</td>
<td>119 (62–192)</td>
<td>Cross-validation (CV) performed</td>
<td>13 (46)</td>
</tr>
<tr>
<td>Training samples</td>
<td>69 (29–103)</td>
<td>leave-one-out CV, n (percentage of CV subset)</td>
<td>13 (46)</td>
</tr>
<tr>
<td>‘Cases’</td>
<td>36 (15–55)</td>
<td>4- to 10-fold CV, n (percentage of CV subset)</td>
<td>12 (7)</td>
</tr>
<tr>
<td>‘Controls’</td>
<td>24 (16–50)</td>
<td>Bootstrap, n (percentage of CV subset)</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Validation samples</td>
<td>54 (33–81)</td>
<td>Feature selection preserves integrity of CV test sets</td>
<td></td>
</tr>
<tr>
<td>‘Cases’</td>
<td>23 (13–38)</td>
<td>Yes, n (percentage of CV subset)</td>
<td>7 (25)</td>
</tr>
<tr>
<td>‘Controls’</td>
<td>25 (11–40)</td>
<td>No, n (percentage of CV subset)</td>
<td>18 (64)</td>
</tr>
<tr>
<td>Ratio of training to validation</td>
<td>1.3 (0.8–2.1)</td>
<td>Unclear, n (percentage of CV subset)</td>
<td>3 (11)</td>
</tr>
<tr>
<td>Training samples from multiple centers, n (%)</td>
<td>9 (26)</td>
<td>Reported performance from CV</td>
<td></td>
</tr>
<tr>
<td>Validation samples from multiple centers, n (%)</td>
<td>13 (37)</td>
<td>Result is the average of all CV results, n</td>
<td>3 (11)</td>
</tr>
<tr>
<td>Training and validation samples independent, n (%)</td>
<td>35 (100)</td>
<td>Result is the best of many CV results, n (percentage of CV subset)</td>
<td>22 (79)</td>
</tr>
<tr>
<td>Clear difference between training and validation samples, n (%)</td>
<td>13 (37)</td>
<td>Unclear, n (percentage of CV subset)</td>
<td>3 (11)</td>
</tr>
<tr>
<td>Predicted outcome</td>
<td></td>
<td>Sensitivity in training sample</td>
<td>94 (84–100)</td>
</tr>
<tr>
<td>Cancer related, n (%)</td>
<td>27 (77)</td>
<td>Specificity in training sample</td>
<td>98 (85–100)</td>
</tr>
<tr>
<td>Non-cancer related, n (%)</td>
<td>8 (23)</td>
<td>Sensitivity in independent validation</td>
<td>88 (78–92)</td>
</tr>
<tr>
<td>Molecular feature used for classification, n (%)</td>
<td></td>
<td>Specificity in independent validation</td>
<td>81 (68–95)</td>
</tr>
<tr>
<td>DNA variation</td>
<td>2 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA (gene expression)</td>
<td>19 (54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>12 (34)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classifier tested in validation is identical to that developed in training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes, n (%)</td>
<td>34 (97)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclear, n (%)</td>
<td>1 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;1 model tested in validation data, n (%)</td>
<td>10 (29)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Values are median (IQR) unless otherwise specified.
was comprised entirely of samples from the first center and roughly half of the validation sample was comprised of samples from the second center).

Feature selection bias and optimization bias were a potential concern in the majority of the included studies, with feature selection bias being likely in 64% and optimization bias in 79% of the 28 studies that performed internal cross-validation. Fifteen (54%) studies had evidence of both biases, 10 (36%) studies suffered from at least one of the two, two studies could not be classified for both types of bias and only one study was clearly exempt from either bias.

**Power estimates for independent validation**

In general, independent validation sample sizes were inadequate to detect even a large (20%) decrease in the sensitivity or specificity compared with the training set performance at the liberal false positive rate of 5%. Across studies, the median power of the independent validation phase was 36%, [IQR, 21–61%] for sensitivity and 29%, [IQR, 15–65%] for specificity at a one-sided alpha = 0.05 for a decrease of 20% between phases. Three studies had power of 80% or higher in the independent validation cohort for sensitivity, and five studies had power of 80% or higher for specificity.

**Comparison of classification accuracy in external versus cross-validation**

As shown in Figure 1, the large majority of studies reported worse classification performance in external validation versus internal cross-validation, and the meta-analysis estimate of the rDOR was significantly different from one (summary random effects rDOR 3.26, 95% CI 2.04–5.21). There was no strong evidence for substantial heterogeneity of rDOR.
between these 28 studies ($I^2 = 0\%$, 95% CI 0–42%) and meta-analysis under a fixed-effects model produced identical results.

Figures 2 and 3 show the rDORs forest plots for these 28 studies stratified by the likely presence of feature selection and optimization bias, respectively. In both of these analyses, the subgroups characterized by a high potential for bias show statistically significantly worse performance in validation than cross-validation (summary rDOR = 4.50, 95% CI 2.54–7.95 for selection bias and rDOR = 3.20, 95% CI 1.99–5.15 for optimization bias), whereas the subgroups without clear evidence of these biases do not show significantly worse performance (rDOR 1.75, 95% CI 0.66–4.60 and rDOR 5.62, 95% CI 0.40–78.80, respectively). However, the point estimate for the subgroup where optimization bias was unlikely is more extreme, and it should be acknowledged that the subgroup estimates have wide and overlapping CIs due to the small number of studies that did not suffer from these potential biases and the large variances of the rDORs.

To explore other causes of heterogeneity in rDORs between studies, we performed meta-regressions with other variables, including trends over time, the ratio of training to validation samples, total sample size, multicenter versus single-center samples in training and validation and presence of a clear difference between training and validation samples, but none of these factors was significantly associated with magnitude of the rDOR (all P-values >0.1).

### Subsequent external validations

We identified all research articles ($n=758$) which cited any of the 35 studies included in our analyses.

![Figure 2: Random effects meta-analysis of relative diagnostic odd ratios (rDORs) for the 28 included studies, stratified by whether feature selection bias is likely to be present. The group with feature selection bias demonstrates significantly worse performance in independent validation than in cross-validation (rDOR 4.50, 95% CI 2.04–5.21), whereas this is not the case in the group that is unlikely to have this bias.](image-url)
and reviewed them at the abstract level in order to identify any subsequent efforts to validate these classifiers. Only three subsequent validation efforts were identified, two by the same research team that published the initial paper [72, 73], and one by an independent research team [74]. Only one study used the exact classifier (predicting local tumor response to pre-operative chemotherapy in breast cancer) described in the initial publication. In this article, the classifier was one of four tested classifiers, and the sample size was 100 (15 cases of residual breast cancer in breast tissue or lymph nodes, 85 with no residual disease). Using the same threshold determined in the primary publication, the performance of the classifier in subsequent validation for sensitivity and specificity was 60% and 74%, respectively, compared to 92% and 71% in the independent validation from the original publication, corresponding to a rDOR of 6.86 (95% CI, 0.60–78.71) for these two independent validations [42, 73]. Accounting for uncertainty in both studies, the power to detect a decrease of 20% in sensitivity or specificity was 16% and 37%, respectively. The second study, performed by the same research team as the original publication, examined the predictive ability of the same mitochondrial deletion mutation for prostate cancer, but in a different clinical scenario. The original classifier had been trained on biopsy samples from subjects with prostate cancer, though since multiple biopsies were available from the same patient, some samples contained only benign tissue. In the validation publication, the authors selected patients with initially benign prostate biopsies. From these initial biopsies, the authors developed a prediction tool to identify the ~20% of subjects who developed biopsy-proven prostate cancer.

Figure 3: Random effects meta-analysis of relative diagnostic odd ratios (rDORs) for the 28 included studies, stratified by whether optimization bias (i.e. reporting the best result from cross-validations versus the average result across all cross-validation) is likely to be present. The group with optimization bias demonstrates significantly worse performance in independent validation than in cross-validation (rDOR 3.20, 95% CI 1.99–5.15). The group that is unlikely to have bias had a more extreme but non-significant point estimate.
within 1 year. Citing the difference in patient populations and biopsy samples, the authors selected a new discriminatory cutoff in the validation paper [58, 72]. The third study, performed by a different research team, used gene expression data from the same three genes used in the original classifier, but the authors used different model-building methods (logistic regression in validation paper versus linear discriminant analysis in the original paper) to develop a classifier from the expression data of these genes [40, 74].

**DISCUSSION**

The development of molecular classifiers from high-dimensional data is subject to a number of pitfalls in study design, analytical methods and reporting of results. This review of internal and external validation practices in the recent literature provides an empirical view of the field, and highlights areas for future improvements. We identified examples of questionable analytic practices in the model-building phase, underpowered independent validation efforts and overly optimistic assessment of classification performance. Cross-validation practices yield optimistically biased estimates of performance compared to external validation. Even external validations may sometimes provide inflated accuracy results and there is a dearth of external replication efforts by totally independent groups.

**Sample size**

Many studies of molecular classifiers are fundamentally limited by sample size. In 11 of the 35 included studies, at least one of the classification groups in either training or validation included 10 or fewer subjects. Cross-validation error estimation methods may be inaccurate in small sample settings [75]. Even if validation efforts yield encouraging point estimates, the large uncertainty associated with these estimates would still preclude drawing strong conclusions about the classifier performance. Furthermore, small sample sizes are more prone to yield spuriously good classification accuracy due to chance. There is some empirical evidence that in the published literature of microarray cancer classifiers, early smaller studies tend to show more impressive classifier performance [76] than larger later studies [77]. Smaller studies may differ from larger studies due to true heterogeneity, chance or publication bias [78, 79]. Moreover, the majority of molecular classifier evaluations are single-center studies. This not only leads to restrictions in the number of participants enrolled, but it also limits the generalizability of the findings. Obtaining clinical samples can be challenging in certain research contexts, particularly for rare diseases. However, for most conditions, increased collaboration between groups or networks of investigators could readily address the issue of inadequate sample sizes. In setting such collaborations, uniformity or at least compatibility of sample collection and processing across collaborating centers is essential. Whenever large sample sizes cannot be attained, the use of appropriate cross-validation methods may be preferable to further reducing the sample size by sample-splitting.

**Internal validation**

The majority of studies in our sample used resampling methods (either bootstrap or k-fold cross-validation) to quantify classification performance in the training sample. There has been considerable debate regarding the appropriate application of cross-validation methods in the literature. In a critical review of cancer classifiers, Dupuy *et al.* [9] showed that most studies pursued inappropriate cross-validation practices. Both feature selection and classifier optimization bias can result in substantial inflation of classification performance estimates [5, 10, 14–16, 19]. Optimization has been termed ‘reporting bias’, ‘optimization of the data set bias’, ‘optimization of the method’s characteristics’ and ‘optimal selection of the classification method’ by previous investigators, depending on the exact mechanism that leads to the unwarranted optimized results [14, 15]. Among the studies we reviewed, 89% were potentially affected by at least one of these types of bias. The performance estimates from current cross-validation practices were optimistically biased compared to the results of external validation, and feature selection and optimization bias contribute to this over-optimism.

**External validation**

We also observed practices that may lead to biased estimates of performance even in external validation. Previous authors have pointed out the risks of occult sources of ‘data leak’ in which information from the external validation sample can leak into the development of a classifier and lead to overly optimistic external validation results [11]. Our study documents other ways in which the process of
external validation may be subverted, the most common of which is the development and validation of multiple classifiers with incomplete reporting of all validation results. In this scenario, external validation may again yield overly optimistic estimates of performance when only the best replication results are reported. Often, it was difficult to deduce from a particular publication exactly how many models were tested.

**Genuine differences versus bias**

In ~40% of studies that we examined, there was some identifiable difference between training and validation samples. Such differences may either reduce or increase the classifier performance in external validation versus cross-validation, depending on what choices are made in selecting samples for either stage. It is not possible to generalize and each study should be scrutinized on a case-by-case basis to decipher what direction bias may take. However, some biases are well-recognized in the literature, e.g. spectrum of disease bias [80–82] may inflate estimates of classification performance when extreme cases and/or hyper-normal controls are used. Early studies may use such designs and reach over-optimistic results that do not hold true when the classifier aims to discriminate between patient groups that are representative of those seen in typical clinical practice and healthy individuals [83]. Of course, it is desirable to extend the applicability of classifiers by testing them in novel target populations and different settings. However, such extensions are probably more appropriate for well-validated rather than novel classifiers that are proposed for the first time and have not been extensively validated.

**Limitations**

Our literature search was designed to yield a representative sample of publications on molecular classifiers with independent validation for binary outcomes. Thus, our findings did not include studies pertaining to continuous outcomes or time-to-event outcomes such as duration of survival or disease-free survival, and they are also not generalizable to studies that report a molecular classifier without making a claim for independent validation. Given our search strategy, some classification methods such as nuclear magnetic resonance or infrared spectroscopy were not well-represented in our sample. Because our study aims to compare cross-validation and external validation results from the same publication, it is unclear whether studies without external validation would have used cross-validation more rigorously. Some studies intended to use cross-validation for model selection; in these cases, one should not use the biased results from cross-validation to describe classification accuracy. In addition, while the number of included studies was sufficient to show a statistically significant difference in classification performance between cross-validation and independent validation, it was clearly insufficient for subgroup analyses. Thus, there may yet be unidentified predictors of heterogeneity in validation performance that our subgroup analyses were not adequately powered to detect.

Our power analyses and rDOR calculations may have been influenced by the fact that data pairs (predicted cross-validation class and the ‘true’ class) for each sample used in cross-validation are not independent across subjects; the dependency arises due to the repeated use of the ‘true’ classes in the cross-validation process [22, 27]. This would have tended to make our power calculations anti-conservative (i.e., power would have been even lower than we calculated). For the same reason, we may have underestimated the var(DORcv), leading to confidence interval coverage far from the intended coverage probability. Sensitivity analysis by artificially inflating the standard error of the DORcv estimate did not affect our conclusions. Simulation studies have demonstrated that when the ‘true’ DOR is large the confidence interval coverage of the DORcv estimate will be approximately correct. In our dataset DORcv estimates were extremely large (median 177.7, IQR: 26.31–680) providing some reassurance; however, because the DOR parameter value in each study is unknown, inferences drawn from resampling-based procedures should be interpreted with caution [27]. We note that such inferences are highly prevalent in the published literature.

We extracted data using strict definitions of practices likely to incur optimization or feature selection bias, but some aspects of our data extraction were limited by poor reporting and diversity of analytic techniques resulting in a certain opacity of methods [10, 14]. Proliferation of methods and increased availability of software for data analysis do not necessarily imply a problem, but when coupled with unclear reporting they limit the ability to reproduce research findings [84]. Attempts to reproduce the published results of analyses of
high-dimensional data sets, both in proteomic [6] and gene expression [4] research, demonstrate the necessity of further, independent replication efforts. However, independent replication efforts for our study sample were almost non-existent. Across the 35 studies included in our analyses, we only identified one further attempt at strict replication (i.e. replication of the exact classifier reported in the original publication).

Our study also identified some areas in which proper validation practices were consistently applied. When independent validation was claimed, there were no instances in which overt overlap between training and validation samples was noted. Similarly, nearly all studies provided specific reassurance that the molecular classifier subjected to external validation was the exact same classifier developed in the training sample, though in most instances the details provided were insufficient to allow for close examination of this claim. Moreover, this does not guarantee that the same exact classifier is also used when external validation is attempted in other studies by the same or independent groups of researchers.

### Concluding comments

As a synopsis, Table 2 catalogues some major biases that can affect the development of molecular classifiers at the stage of cross-validation or external validation. Some of these biases are uncommon, while others are quite prevalent. Opaque reporting may occasionally hinder an exact appreciation of these biases.

The advent of high-throughput genomic and proteomic data is likely to be transformative for biology and medicine. However, reaping the benefits of quantitative changes in the availability of data requires also qualitative changes in the way research is performed, reported and validated. Our empirical evaluation reinforces the importance of transparent reporting and rigorous independent external validation of classifiers developed from high-dimensional data.

### Key Points

- The development of molecular classifiers is subject to a number of pitfalls which can lead to overly optimistic reports of classification performance.
- Many studies employ cross-validation methods which result in overly optimistic performance estimates.
- Validation in an independent data set is crucial in establishing the reliability and generalizability of a molecular classifier, though this process can itself be susceptible to bias when inappropriate methods are used.
- There is a dearth of subsequent independent replication efforts for published molecular classifiers.
FUNDING
National Institutes of Health (K08HL102265 to P.J.C. and UL1 RR025752 to Tufts-Clinical Translational Science Institute); Research Scholarship from the ‘Maria P. Lemos’ Foundation (to I.J.D.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health.

References
34. Stata Statistical Software: Release 11. 2009. College Station, TX, StataCorp, LP.


