Impact of Bioinformatic Procedures in the Development and Translation of High-Throughput Molecular Classifiers in Oncology

Charles Ferté1,2,3, Andrew D. Trister1,4, Erich Huang1,5,6, Brian M. Bot1, Justin Guinney1, Frederic Commo1,2,3, Solveig Sieberts1, Fabrice André2,3, Benjamin Besse2,3, Jean-Charles Soria2,3, and Stephen H. Friend1

Abstract

The progressive introduction of high-throughput molecular techniques in the clinic allows for the extensive and systematic exploration of multiple biologic layers of tumors. Molecular profiles and classifiers generated from these assays represent the foundation of what the National Academy describes as the future of "precision medicine". However, the analysis of such complex data requires the implementation of sophisticated bioinformatic and statistical procedures. It is critical that oncology practitioners be aware of the advantages and limitations of the methods used to generate classifiers to usher them into the clinic. This article uses publicly available expression data from patients with non-small cell lung cancer to first illustrate the challenges of experimental design and preprocessing of data before clinical application and highlights the challenges of high-dimensional statistical analysis. It provides a roadmap for the translation of such classifiers to clinical practice and makes key recommendations for good practice. Clin Cancer Res; 19(16); 4315–25. ©2013 AACR.

Introduction

As high-throughput molecular technologies become ubiquitous and as antineoplastic agents are increasingly directed against specific molecular aberrations, modeling the relationship between genomic features and prognosis or therapeutic response provides the substrate for precision medicine (1). Over the past decade, very few biomarkers have reached the required level of evidence to be implemented in the clinic (2) and a dearth of genomic signatures generated from the aforementioned technologies have been approved for clinical use (3). Ironically, as the molecular data available in repositories rapidly expand, effective, validated translation of these data to bedside diagnostics or target discovery remains a vexing challenge. Apart from the typical statistical challenges facing biomarker studies (4), there are unique issues that accompany high-dimensional genomic platforms that present obstacles to generating performant genomic signatures. Unfortunately, many of these issues are obscure to the larger oncology community.

Herein, we highlight problems associated with developing molecular signatures at each phase of development: (i) data curation and preprocessing, (ii) statistical analysis, and (iii) the infrastructure required for effective translation in cancer research and clinical settings. To show each of these issues we focus on gene expression data, though the discussion is applicable to many types of high-dimensional data. Each section of this review includes pertinent figures of analysis conducted following recommendations for best practice (Table 1). For both educational and reproducibility purposes, we provide real data (available through Synapse, the collaborative compute space developed at Sage Bionetworks, under the Synapse ID ‘syn87682’: https://www.synapse.org/#/Synapse:syn87682) and companion R scripts (available on GitHub: https://github.com/SageBionetworks/Ferte-et-al-Review).

Experimental design and data preprocessing

Importance of experimental design. As in any scientific study, thoughtful experimental design increases the chance that the question being explored can be answered by the experimental data collected. A justified critique of many molecular signatures is that too less attention is paid toward typical statistical issues such as proper experimental design, sample size planning, patient selection, and clinical data curation (4). As with clinical trials, appropriate selection of a patient cohort, endpoint of interest, and sample size determination must be conducted a priori. Other common
Table 1. Practical issues and recommendations for the development and the translation of molecular classifiers in oncology

<table>
<thead>
<tr>
<th>Step of development and translation</th>
<th>Issue</th>
<th>Proposal for best-practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental design</td>
<td>Selection and curation of the datasets.</td>
<td>Cross-talk between oncologists, biostatisticians, and bioinformaticists is warranted to choose the most appropriate data. Appropriate selection of the samples according to the clinical and biologic variables. Heterogeneity of the clinicopathologic variables between datasets should be evaluated and possibly adjusted. Sample size assessment should be processed a priori.</td>
</tr>
<tr>
<td>Preprocessing step</td>
<td>Latent unwanted structure embedded in data has the potential to dramatically impact the analysis.</td>
<td>Raw data should be used for all subsequent analysis. All the preprocessing steps should be described explicitly (including the normalization, the rescaling, and the correction for adjustment variables used). Preprocessing code and preanalytic plots showing the structure of the data should be provided. Use of reference or housekeeper features. Use of universal or control samples as reference to be processed simultaneously to the patient sample.</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>Multiple comparisons</td>
<td>Report multiple correction adjustments for all statistics. Validation data should be kept entirely and wholly separated from the training data to ensure no potential for contamination. Apply methods that address the large ( p ) small ( n ) problem (e.g., ridge regression, lasso, principal component regression, partial least squares, etc.).</td>
</tr>
<tr>
<td></td>
<td>Resubstitution bias</td>
<td>Internal assessment of the performance of the model (cross-validation, bootstrapping). The analysis method and the code used to process it should be made publicly available.</td>
</tr>
<tr>
<td></td>
<td>Large ( p ) - small ( n )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Robustness of the model (multiple local optima).</td>
<td></td>
</tr>
<tr>
<td>Performance assessment</td>
<td>Generalization of the model in other dataset(s).</td>
<td>Stability of the performance must be validated in external dataset(s). ROC-AUC is relevant for binary endpoints; RMSE and ( R^2 ) are relevant for continuous endpoint; time-dependent AUC or concordance index are relevant for survival endpoints. The performance of the classifier must be compared with existing standard estimators.</td>
</tr>
<tr>
<td></td>
<td>Kaplan–Meier plots and heatmaps are not adequate to assess the performance of the model.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medical or biologic use.</td>
<td></td>
</tr>
<tr>
<td>Clinical development</td>
<td>Routine measurement of molecular classifiers is limited.</td>
<td>Expand the routine capture of pathologic specimens and data in the clinic. Incorporate biomarker validation in the design of clinical trials (cross-talk with biostatistician required). Ensure the samples of the patients enrolled in ongoing and future prospective trials are preserved for subsequent and unplanned analysis provided appropriate consents are given.</td>
</tr>
<tr>
<td></td>
<td>Current clinical trial designs do not incorporate biomarkers.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Limits on testing of archival pathologic specimens.</td>
<td></td>
</tr>
</tbody>
</table>

(Continued on the following page)
errors include the imbalance of clinicopathologic, treatment, and survival characteristics between training and validation cohorts. Particularly, the incompatibility of follow-up between data sets results in responses that may not be comparable.

With regards to sample size calculation, several web accessible tools are available to ensure adequate statistical power (5, 6). Figure 1A presents the results of the data curation process for a gene expression classifier designed to predict 3-year overall survival in patients with early-stage non–small cell lung cancer (NSCLC), which will be a motivating example throughout this review.

**Quality assessment of molecular data.** Preanalytic quality assessment of the molecular data is necessary not only when processing “raw” data (data collected directly from the assay platforms before normalization) but continuously throughout all steps of data analysis. Methods for assessing global structure in the data, such as principal component analysis (PCA) and clustering, are used to detect outliers, or confounding artifacts in the data that must be abated before data modeling may proceed (7–10). To this end, a number of publicly available tools such as arrayQualityMetrics (9), EDASeq (10), or FastQC (Babraham Institute) are widely used.

**Inherent biases in high-dimensional data.** Many high-dimensional omic technologies estimate the abundance of targeted elements by measuring the signal of labeled probes designed to hybridize to the specific targets (features; refs. 7, 11). These signal intensities are commonly represented by a matrix of $n$ by $p$ elements where $n$ is the number of samples and $p$ is the number of molecular features. The objective of any analysis using high-dimensional molecular data is to infer the relationships between biologic or clinical endpoints. Complicating these analyses is the presence of unwanted variables that arise from the specific technology platform, study design or uncontrolled biologic sample heterogeneity (7, 10, 12–16). In most cases, technical artifacts such as dye, probe, platform, technician, and run-time batch are the most common source of latent structure in the data. Known and unknown biologic variability not related to the design and endpoints of a study can also influence detecting signal above noise (14). For instance, histologic grade is often associated with higher necrosis in the tumor tissue and ultimately affects signal intensities but may have less influence on the predicted clinical endpoint (17). The objectives of these preprocessing methods are (i) to remove all latent structure and technical artifacts seen in the data, while (ii) preserving the influence of the biologic variables of interest.

**Background correction.** As the binding or hybridization events at the core of most array-based omic technologies are stochastic, there is a degree of nonspecific binding that alters the signal and must be accounted for (7, 9). Many vendors provide adequate software or hardware design to explore and reduce the influence of nonspecific binding. However, accepting these default corrections may also introduce additional biases.

While next generation sequencing (NGS) technologies obviate many biases present in array technologies, and therefore do not require background correction, they give rise to new unwanted biases such as base-call error and coverage biases and others that have yet to be fully elucidated (10, 11, 16).

**Normalization.** The objective of normalization or standardization is to make the data comparable across experiments by making the distributions the same. Many studies aiming to develop oncologic molecular predictors, including the original NSCLC studies discussed in detail in this review, use unsupervised normalization methods (18–21). Unsupervised normalization methods [e.g., linear scaling, cross-validated splines (22), running median lines (23), loess smoothers (7), and quantile normalization (7)] remove bias across samples perceived to be due to technical variation blind to the experimental design and biologic differences. The types of transformations applied in these methods vary widely and it is exceedingly important to understand that each impact the downstream model performance differently. Oncologists should note that, as with any normalization procedure, these techniques may obscure the biologic signal.

### Table 1. Practical issues and recommendations for the development and the translation of molecular classifiers in oncology (Cont’d)

<table>
<thead>
<tr>
<th>Step of development and translation</th>
<th>Issue</th>
<th>Proposal for best-practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation at bedside</td>
<td>Molecular data derived from current sampling standards are highly context specific (intratumoral heterogeneity, treatment effect, host effect, etc.)</td>
<td>Increase the number of tumor samples to achieve a better representation of the disease (Sequential biopsies, primary and metastatic sites).</td>
</tr>
<tr>
<td></td>
<td>Uncertainty of the results of the classifiers are rarely revealed to the oncologist.</td>
<td>Confidence intervals of the results should be provided to the oncologist for the decision to be made in the patient’s context.</td>
</tr>
<tr>
<td>Poor training in modern technologies and methods is a major limit in the translation of molecular classifiers at bedside.</td>
<td>Promote the cross-training of oncologists and cancer biologists in computational biology, systems biology, and biostatistics.</td>
<td></td>
</tr>
</tbody>
</table>

Published OnlineFirst June 18, 2013; DOI: 10.1158/1078-0432.CCR-12-3937
Figure 1. Overview of the preprocessing framework. Effects on the structure of the data are represented by principal component plots for 4 NSCLC gene expression datasets processed separately. A, a table to represent the number of raw data (CEL files) included in study as a result of the data curation process. As the classifier is for early-stage patients, an explicit decision was made to only include those who are in pathologic stage IA to IIIA, who did not receive induction or adjuvant chemotherapy and patients for whom overall survival data are available. In addition, only patients who underwent
Bioinformatic Procedures for Molecular Classifiers

Figure 2. Signatures developed using different methods have similar prediction performance (A) and present very less consistency with each other (B). A, receiver-operating characteristic curves of 6 widely used statistical methods (logistic regression, elastic net, bootstrapped elastic net, random forest, principal component regression, and partial least square regression) in predicting the probability of 3-year overall survival. The Director’s challenge and the Zhu and colleagues datasets are used as training and validation set, respectively (19). The ROC-AUC and their 95% confidence interval are computed for each method. Note that all curves overlap with one another. B, the number of features selected with each method are presented, as well as the number of genes that overlap from each method. Note the very small overlap of features across the different models, confirming that multiple and different solutions (local optimia) of a same problem may lead to similar prediction results. 95% CI, confidence interval.

while removing the latent structure of the data, making quality control challenging.

Several drawbacks render these methods difficult to translate into clinical practice. First, they require large datasets to perform adequately. Consequently, normalization cannot be applied on individual samples. Second, training and testing sets must be preprocessed together, necessitating simultaneous access to full training and testing datasets. As an example of the differences between five commonly used methods, we visualize the individual results on 4 publicly available early-stage NSCLC Affymetrix gene expression datasets (Fig. 1). PCA of the results clearly reveal that: (i) all methods transform the structure of the data (Fig. 1B), (ii) these transformations are different across normalization methods (Fig. 1B), (iii) intrastudy normalization [Fig. 1C; robust multi-array average (RMA) and supervised normalization method (SNM) callout] does not remove artifactual segregation between studies requiring further interstudy rescaling (Fig. 1D). These differences highlight the importance of beginning with raw data when developing molecular signatures to minimize hidden biases made by previous assumptions. Unfortunately, raw data are often not available for subsequent analysis (Supplementary Table S1; ref. 22). Searching for consistent patterns across multiple high-dimensional molecular datasets can also be done using meta-analysis techniques (24–26). Differences in individual study sample sizes and patient populations can often not be taken into account when study-level estimates are used. Pooling “raw” or patient-level data and fitting appropriately stratified models across studies, whereas complex, is the only way to sufficiently control for these biases (25, 26).

Supervised normalization. Supervised normalization incorporates information about confounding variables and variables of interest that can dramatically affect the performance of high-dimensional molecular models (12, 27, 28). Although experimental batch is the most frequently recognized source of latent structure, other environmental (29), genetic (30), and demographic (31) effects are inherent in each individual experiment. Several methods like surrogate variable analysis, SNM, and ComBat (13, 14, 32) are designed to solve the effects of these variables on data structure, but their employment is rarely described in detail in molecular signature articles (17, 33).

Specific preprocessing procedures for next generation sequencing. NGS is a rapidly evolving field and its data...
Figure 3. Comparison of receiver-operating characteristic curves, Kaplan–Meier survival prediction and heatmaps for 6 commonly used statistical methods (bootstrapped elastic net, elastic net, logistic regression, partial least square regression, principal component regression, and random forest) ordered by performance on ROC. Log-rank test is used to report P value of the differences of good outcome and poor outcome groups (as defined by median) for Kaplan–Meier predictors. The patients included in each group in Kaplan–Meier analysis are coded in the unsupervised clustering among the validation dataset.
are increasingly incorporated in the development of classifiers (34–36). Furthermore, these technologies have been used to elucidate many important biologic differences among pediatric tumors. Recently, the International Cancer Genome Consortium PedBrain Tumor Project showed the excellent use of NGS to elucidate the genetic complexities inherent in medulloblastoma (37). Drawbacks similar to those seen with gene expression microarrays also exist in NGS technologies and specific preprocessing methods are required (10, 11). Particularly, the short-length of the nucleotide sequences produced (reads are typically 50 to 150 nucleotides in length) necessitates assembly and annotation frameworks when reconstructing the genome for variant analysis purpose (such as with detecting SNPs, MNPs, and InDels; ref. 38). These technologies are also increasingly used to quantify gene expression and just as in microarray experiments, they require close consideration of latent variables when assessing a perceived signal (39). Specific preprocessing methods to estimate gene expression have been introduced: reads per kilobase per million, GC-content normalization, normalization to ‘housekeeping’ genes, and quantile normalization (10, 40–43).

**Issues with development of classifiers in the context of high-dimensional data**

**Impact of high-dimensional data on analysis design.** In the context of molecular profiling, “high-dimensional” data are generated such that the number of features (p) is much larger than the number of samples (n; p >> n). Any subsequent analysis suffers from the “curse-of-dimensionality,” that an association between a molecular feature and a clinical outcome of interest may occur by chance, a phenomenon known as ‘false discovery’ (44). The most commonly used methods to address false discovery are on the basis of the work by Benjamini and Hochberg and Holm and Bonferroni (45–47). In addition, there is the potential for ‘overfitting’ of a classifier in the training dataset, which reduces the resultant classifier’s performance on new data. Until the number of samples approaches the number of features (n ≈ p), strategies to minimize overfitting involve reducing the dimensions of the model space (44). These methods take advantage of the high correlation between subsets of variables, virtually eliminating (principal component regression, lasso) or “shrinking” (ridge regression, support vector machine, elastic net) nonessential features (48). A detailed discussion of these methods is beyond the scope of this review and has been addressed comprehensively elsewhere (49).

**Less consistency across classifiers developed by different methods.** Classifiers developed with different methods on the same data set often result in similar predictive performance, but exhibit small overlap in the features selected (Fig. 2). By reducing the dimensions of data, the models represent only one from multiple possible solutions (also called local optima). Whereas, we believe there is a global optimum (the best of all possible solutions), finding this solution is often computationally intractable. Sometimes, the aggregation of multiple models into a consensus classifier may result in improved predictive performance (Fig. 2). Ultimately, how the knowledgeable oncologist incorporates these different models for decision-making remains a challenging and still open question.

**Internal validation performance assessment.** Cross-validation and bootstrapping are the most widely used internal validation methods. These methods are conducted by developing a classifier on a subset of samples and then testing the resultant classifier on held-out samples (48). Investigators must be aware that cross-validation and bootstrap methods are aimed to estimate prediction error and do not preclude testing the model in an external dataset (external-validation). Internal validation methods are also used to improve the robustness of the model against noise inherent in the data.

**External validation and clinical use.** Ultimately, as the clinical use of the model is highly dependent upon its ability to correctly predict an endpoint in an external dataset, particular attention must be paid to performance metrics. A typical measure of the performance for binary classifiers (i.e., predicting binary outcome such as tumor recurrence) is based upon the receiver-operating characteristic curve (ROC), which illustrates the variation of true-positive and false-positive rates along the variation of the discrimination threshold (50). In the case of predicting a continuous endpoint, root-mean squared error (RMSE) or $R^2$ are frequently computed. When assessing time-to-event or survival endpoints the most commonly used metrics are concordance index and time-dependent AUC (4, 51, 52).

Oncologists should be aware that Kaplan–Meier curves and log-rank comparisons estimate differences in hazard across predicted risk groups but do not assess predictive performance. The most striking example of this disconnect between discrimination and prediction is recent work that showed random signatures in breast cancer are associated with outcome (53). The discerning clinician tasked with assessing the validity of a particular signature should demand reporting of additional statistical performance metrics. An illustrative comparison of ROC-area under the curve (AUC), Kaplan–Meier estimates and heatmap results is in Fig. 3. In addition, the medical use of any molecular
model must be formally addressed with regard to clinicopathologic covariates or scores currently used in the clinic.

**Issues with the effective translation of the classifiers into the clinic**

Translating modern classifiers to the bedside requires not only robust preprocessing and analytic methods, but also the infrastructure for incorporating high-dimensional molecular data into the clinical and translational research.

A first critical issue related to the clinical environment of the assays used to generate the molecular data. Currently, a unique biopsy is conducted on one tumor site per patient at a single time point over the course of the disease. However, a growing body of evidence shows that the molecular data derived from “single site, single time point” biopsies are highly context-specific and may provide a biased representation of the disease state (54–56). Indeed, these data can vary depending on biopsy location given intratumoral heterogeneity (54) and the discordance between the primary site and metastases (56). The time at which a sample is obtained is also a source of variation because the relevance of an oncogenic driver may change along the sequence of antineoplastic treatments (55) or during the natural history of the disease itself (56). Multiple assessments are usually not conducted at bedside due to technical, safety, and ethical constraints. However, emerging technologies, such as circulating tumor cells (57), circulating DNA (58), or next generation functional imaging could allow for dynamic sampling.

Second, mechanisms must be in place for appropriate clinical evaluation of classifiers to concretely translate them to the bedside. Although retrospective analysis on completed prospective trials can be conducted in certain circumstances, true prospective validation remains the gold standard (59). Several prospective clinical trial designs are optimized to validate biomarkers: biomarker-stratified design, enrichment design, biomarker strategy design, multiple hypothesis design, and maker based strategy design (60–65). Unfortunately, these designs were developed for the phase I–II context and are thus not powered to capture the complexity of cancer biology, and do not support inference analyses in large populations. Recently introduced adaptive signature designs address this caveat, using randomized phase III trial designs to develop and validate classifiers (66, 67).

The recent scandal at Duke University (Durham, NC) surrounding the use of microarrays to drive clinical trials highlights the need for peer access to both the data and methods used to generate complex biomarkers. As discussed in the recent National Academies Report, “Toward Precision Medicine” (1), before a prospective clinical trial incorporating molecular classifiers is begun, the analytic process by which these classifiers are generated must be reproducible and transparent. Furthermore, all developmental assumptions about clinical endpoints and parameterization of the models should be made explicit, including explicit annotations of treatment and inclusion criteria that are distributed with the data in standard format like MIAME (68). Taking a cue from the world of software development, genomics researchers are progressively incorporating improved operating principles such as version control for scientific source code in platforms favoring open access such as GitHub, and displaying enriched content in dedicated sites such as COSMIC, cBioPortal, GenePattern.

Furthermore, making these data and methods truly accessible has the potential to dramatically increase the efficiency of research by enabling the identification of new avenues of research and avoiding the futile reproduction of strategies that may not work. To this end, dedicated compute spaces allowing for transparent and reproducible collaboration could enable the access of models in real-time independent of the cycle time of peer-reviewed journals. Synapse is one such system that integrates unique data URLs, provenance, cloud computing, and markup to provide a cohesive communication of a data-intensive study (69). At a fundamental level, adequate community assessment of published models mandates that authors make data and code available (60). Publishers are also beginning to introduce new ways to share data such as with Nature: Scientific Data to improve data transparency, citation, and curation.

Ideally, published models should provide simple interfaces where clinicians can quickly obtain predictions for individual patients on the basis of their molecular and clinical features in a manner similar to Adjuvant! Online (70). Although molecular classifiers are more than that a simple combination of genes and specific issues need to be addressed when translating these tools at bedside for a unique patient setting (Fig. 4). In that regard, the modern oncologist must come to terms with the ever-changing nature of genomic science, with the ultimate risk of being left behind.

**Conclusion**

Many investigators believe that effective analysis of high-dimensional molecular data is the key to controlling cancer. It is a widespread assumption that molecular classifiers represent the foundation of individualized oncologic care. To optimize the translation of these tools into the clinic, the oncology community needs to be aware of the strengths and limitations of the specific bioinformatics and statistical methods used in the development of classifiers. Unless molecular classifiers are truly able to discern key clinical issues, the promise of precision medicine will remain elusive and the clinical impact will remain limited. The drive toward precision medicine depends on cross-disciplinary training necessary for the next generation of oncologists to lead these research endeavors. Developing competencies in cancer biology, biostatistics, computational biology, molecular biology, and computer science in addition to patient care will be crucial to training the next leaders of the field.
Bioinformatic Procedures for Molecular Classifiers

Figure 4. Challenges in the translation at bedside of a validated molecular classifier. Described are the steps taken by the modern oncologist when obtaining a prediction of a validated classifier for a single patient. Passing through these steps, particular problems and their potential solutions are highlighted in red boxes. To embrace precision medicine, the modern oncologist needs to develop or access competencies in molecular biology and in computational biology, in addition to clinical oncology.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors' Contributions
Conception and design: C. Ferté, A.D. Trister, E. Huang, B. Besse
Development of methodology: C. Ferté, A.D. Trister, J. Guinney, J.-C. Soria
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C. Ferté, E. Huang
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): C. Ferté, A.D. Trister, B.M. Bot, J. Guinney, F. André, J.-C. Soria
Writing, review, and/or revision of the manuscript: C. Ferté, A.D. Trister, E. Huang, B.M. Bot, J. Guinney, F. Commo, S. Sieberts, F. André, B. Besse, J.-C. Soria, S.H. Friend

Grant Support
This work was supported by a grant 1154CA149237 from The Integrative Cancer Biology Program of The National Cancer Institute, NIH. C. Ferté is a recipient of a fellowship from The French Cancer Research Association (ARC; 2011) and of a Young Investigator Award from The Conquer Cancer Foundation of The American Society of Clinical Oncology (2012).

Received December 28, 2012; revised May 15, 2013; accepted May 31, 2013; published OnlineFirst June 18, 2013.

References


Published OnlineFirst June 18, 2013; DOI: 10.1158/1078-0432.CCR-12-3937

Downloaded from clincancerres.aacrjournals.org on May 31, 2017. © 2013 American Association for Cancer Research.