Independent Test Set Performance in the Prediction of Early Relapse in Ovarian Cancer with Gene Expression Profiles

To the Editor: With great interest, we read the article by Hartmann et al. (1) investigating whether it is possible to apply gene expression patterns in order to discriminate between ovarian tumors with early and late relapse after platinum-paclitaxel combination chemotherapy. Among others, the authors claim to have derived a 14-gene predictive model with an independent test set accuracy of 86% and a positive predictive value of 95%.

However, after examination of the data analysis strategy of Hartmann et al., we noticed that the test set has been used to perform prior model selection and therefore cannot be called independent. Summarized and after data preprocessing, the authors constructed 100 support vector machine models each based on a set of genes with the highest signal-to-noise ratio derived from a random selection (70% of 51 training samples) of the training set. Subsequently, these 100 models were all tested on the (wrongfully called independent) test set (28 samples) and the top model with the fewest prediction errors was selected and reported.

Unfortunately, this selection implies that information from the test set was used to choose a model that optimally fits this particular test set but might perform worse on another and independently chosen test set. As a consequence, the reported performance indices might be overestimated and will probably be impossible to reproduce on new prospective data. In our experience, and due to the high-dimensional nature of microarray data, even the slightest use of a so-called independent test set (or the use of the left-out samples in cross-validation studies; ref. 2) within the model-building process will dramatically overestimate the performance of a classifier based on expression patterns. After model selection and in order to obtain a realistic estimate of the true performance, it is therefore imperative to test a new model on completely independent and prospective data (3).

In order to substantiate our claims, we implemented a similar data analysis scheme in MATLAB (Release 13, script can be obtained on request) based on 14-gene support vector machine models from LS-SVMlab (version 1.5, http://www.esat.kuleuven.ac.be/sista/lssvmlab/; refs. 4, 5). We subsequently applied our script to 10 randomly generated data sets, each subdivided in a training and test set (expression levels uniformly and independently drawn between 0 and 1) with the same dimensions and composition as reported by Hartmann et al. For a true independent test set, and since the random data does not contain any information about the process under study, one could expect an accuracy of ~50%. However, the 10 test set accuracies returned by our MATLAB script (one for each training + test set) ranged between 71.43% and 82.14% and were significantly (P = 0.002; sign test) different from 50%. Therefore, these results indicate that the procedure described by Hartmann et al. strongly overestimates the accuracy that can be expected on independent data. Also noteworthy was the observation that the accuracies on the test set (in this case, truly independent) indeed varied around a mean of ~50% if the model selection step was omitted. In the latter case, we considered all 1,000 models (100 models for each random data set) and not only the 10 models selected for their optimal performance on the test set.

Finally, we want to mention that Hartmann et al. stated that the reported accuracy of 86% was very unlikely to occur by chance alone. This was—similarly as above—assessed by comparing this result with a series of test set accuracies obtained through random models (in this case, generated by randomly permuting the outcome labels of the training set). However, this assessment only indicates that the reported accuracy is relatively better than the test set accuracies of the random models. Since our simulation showed that these values themselves are overestimated, this evaluation does not say anything about the validity of the absolute value of the reported accuracy of 86%. Nevertheless, this assessment seems to indicate that the expression patterns indeed contain information about the time of relapse after chemotherapy.

In summary, we believe that the magnitude of the performance indices of the 14-gene model derived by Hartmann et al. will not be confirmed on a truly independent test set. In our opinion, and in the absence of new prospective data to properly assess the current model, we believe that model training should be repeated using a method that refrains from exploiting any information from the test set. Only under these circumstances is it possible to correctly estimate the test set performance. Nowadays, the authors have the choice between a wide variety of suitable classification methods that have been specifically developed for expression data and that are publicly available [as an example, see Pochet et al. (3) and Tibshirani et al. (6)].

Frank De Smet
Nathalie L.M.M. Pochet
Bart L.R. De Moor
Department of Electrical Engineering, ESAT-SCD, K.U. Leuven, Leuven-Heverlee, Belgium

Toon Van Gorp
Dirk Timmerman
Ignace B. Vergote
Department of Obstetrics and Gynecology, Division of Gynecologic Oncology, University Hospitals, K.U. Leuven, Leuven, Belgium

Grant support: KUL Ph.D./postdoctoral grants; KUL-GOA AMBioRICS; FWO: G.0115.01, G.0407.02, G.0388.03; BFSPO-IUAP P5/22; EU-RTD: FP6-NoE Biopattern.

doi:10.1158/1078-0432.CCR-05-1216

References

1. Hartmann LC, Lu KH, Linette GP, et al. Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy. Clin Cancer Res 2005;11:2149–55.
2. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
3. Pochet NL, Janssens FA, De Smet F, Marchal K, Suykens JA, De Moor BL. M@CBETH: a microarray classification benchmarking tool [Epub 2005 May 12]. Bioinformatics 2005;21:3185–6.
4. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. Singapore: World Scientific; 2002.
5. Pochet N, De Smet F, Suykens JA, De Moor BL. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics 2004;20:3185–95.
6. Tibshirani R, HastieT, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567–72.

In response: De Smet et al. raise important issues about the validation of microarray studies and question the reproducibility of our findings. Clearly, the standard for this type of work is evolving. Previously accepted approaches to validation of microarray studies have included documenting the level of expression of the markers via reverse transcription-PCR, applying the markers in another set of samples, and running the samples on another array platform. The approach taken often depends on the availability of additional samples.

We had a total of 79 samples from women with advanced-stage ovarian cancer, of similar grade and histology, treated with the same two chemotherapeutics (paclitaxel and carboplatin) with sufficient follow-up to allow classification of response to treatment. To our knowledge, this is the largest and most homogeneous ovarian cancer population where the question of outcome after chemotherapy was under study by microarray technology.

De Smet et al. take issue with our use of the term "independent" to describe the second test set of 28 samples in which we tested our model that was developed in the discovery set. We described our use of the test set in Data Analysis and depicted the approach in Fig. 1. We used "independent" to mean a set of samples that were not employed in the discovery of the model itself. We used "validation" to mean the subsequent application of this model in a completely different set of samples, which we stated that we did not do. In fact, both in Abstract and Discussion, we stated that our observations require further validation. We agree with De Smet et al. that the term "independent" should probably be reserved, with "validation," to refer to completely distinct samples that were not employed in marker development.

Although it is difficult for us to comment on the De Smet et al. simulation study due to the lack of details, we do not dispute the suggestion that our reported prediction accuracy estimate of 86% is likely a positively biased estimate. We acknowledge (and state in Discussion) the need to test our top model on a data set that is external to the model selection process to obtain a less biased estimate of prediction accuracy. However, we would point out that a validation study in an external set would result in a single point estimate of prediction accuracy, the precision of which is highly dependent on the size of the validation set.

We agree that the permutation test does not validate the prediction accuracy. However, that is not its intended purpose. Its purpose is to assess how likely it is that the model could have been found by chance alone. This is a fairly standard method extensively used in these types of data sets and we believe that we have applied it fairly where, as depicted in Fig. 1, the test was external to the filtering and model selection steps.

In fact, De Smet et al.'s analysis corroborates our permutation test results. They apparently repeated our methodology on entirely random data and found a prediction accuracy in the mid 70% range. We believe that the 86% accuracy seen with our model (outside the range reported by De Smet et al.) reflects a set of markers that have biological relevance, which they acknowledge.

Without access to a similar set of samples for model validation, we opted to validate our platform by running the remaining RNA from our samples on Affymetrix U95 chips. We were able to identify corresponding Affymetrix U95 probe sets (methods available on request) for 11 of our top 14 markers. Of the 11 genes, 6 were reasonably well correlated ($r > 0.6$ includes *ID4*, *FARP1*, *BTF3*, *HIS1*, *sF3A3*, and *FLJ20241*) whereas 3 were poorly correlated ($r < 0.4$ includes *FLJ22269*, *PTPRS*, and *ZNF200*). The remaining two had intermediate correlation.

Another way to examine platform differences is to run an ANOVA model looking for differential effect of response by platform. Results of the ANOVA showed only two markers that had significant ($P < 0.05$) differential effect of response across platforms (HIS1 and PTPRS). Even these two were no longer significant after correcting for multiple testing artifacts. In addition, we had one TaqMan panel for ID4 but for only a limited subset of patient samples ($N = 20$). However, these data correlated extremely well ($r = 0.99$) with nearly the same mean, median, and SD. This high degree of correlation is surprising and is probably due to TaqMan reagents carefully designed to match the resequenced IMAGE clone in question.

Taken together, these platform validation results suggest that most of the markers in our data would show similar differential expression between response groups using another platform such as Affymetrix.

In summary, we share the goal of De Smet et al. to raise the bar of quality for validation of microarray work. Better standardization of terms such as "test set," "independent," and "validation set" will also improve the field.

Lynn C. Hartmann
Mayo Clinic College of Medicine,
Rochester, Minnesota

Andrew I. Damokosh
Clinical Biostatistics Oncology,
Bristol-Myers Squibb,
Wallingford, Connecticut

Sebastian Hoersch
Computational Biology,
Millenium Pharmaceuticals, Inc.,
Cambridge, Massachusetts