Identification and external validation of gene expression signature as potential diagnostic biomarker for endometriosis

Anke Schulz[1], Sebastian Voss[2]

[1]Genomics & Biomarker Statistics, Bayer AG, Berlin, Germany, [2]Chrestos Concept GmbH & Co. KG, Essen, Germany

INTRODUCTION

Endometriosis is a chronic disease, defined by the presence of endometrial tissue outside the uterus, resulting in pain and difficulties to conceive. As a surgical procedure (laparoscopy) is currently the standard way to achieve a definite diagnosis, an aim of this study was to find a diagnostic biomarker panel based on genome wide RNA expression in the eutopic endometrium.

All N=90 women in the study underwent a laparoscopy, in which tissue samples were collected and endometriosis was confirmed for n=67 subjects. The RNA expression data from the tissue samples was used to construct a diagnostic classifier for the presence of endometriosis. The classifier was then applied to a comparable publicly available data set shared by Tamaresis et al. (2014).

DATA PREPARATION

- RNA expression in the eutopic endometrium was measured by **Affymetrix HG-U133 Plus 2.0** chips.
- Robust Multi-Array Average (**RMA**) was used for chip normalization (preprocessing in order to extract the signal from the noisy raw data and transform all samples to a comparable scale).
- Probe sets were summarized on gene level using a customary chip description file published by the Molecular and Behavioral Neuroscience Institute (University of Michigan). → **20141 features / variables**

STATISTICAL METHODS

Pre-analysis and data visualization

- Univariate moderated t-tests according to Efron (2001) were performed for a **feature-wise comparison** of subjects with and without endometriosis.
- Results were visualized by means of a volcano plot and a histogram of the p-values.

Classifier construction

- Features that were considered in the classifier construction were preselected based on the univariate moderated t-tests (**preselection of 200 features** with smallest p-values).
- The diagnostic classifier was constructed based on a **penalized logistic regression** model with the endometriosis diagnosis as the binary dependent variable y and the intensities of the preselected features as the independent variables x, i.e.

$$P(\, y = \text{endometriosis} \mid x\,) = \left(1 + \exp(-\beta_0 - \, x^T\beta)\right)^{-1}.$$

- The final feature selection and parameter estimation in the logistic model was performed using the **elastic net**, i.e. the intercept β_0 and the effect vector β were estimated as

$$\underset{\beta_0,\beta}{\operatorname{argmax}} \left[l(\beta_0, \beta | y, x) - \lambda[(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1] \right],$$

where l denotes the log-likelihood of the standard logistic regression model. The optimal shrinkage parameter λ was chosen based on a 10-fold cross-validation, while the weight parameter $\alpha \in [0; 1]$ was set to 0.5.

Performance estimation (for new data)

- Classifier performance was evaluated by applying a **5-fold cross-validation** (stratified by diagnosis) to the described construction procedure (repeated 500 times).
- Results of the cross-validation (CV) were summarized by averaging the ROC-curves and the AUC over all CV-folds. A 95% confidence interval of this cross-validated AUC was determined with an influence curve based approach as described in LeDell et al. (2012).

RESULTS

- Feature-wise comparison indicated slight differences in RNA expression between subjects with and without endometriosis **(Figure 1)**.
- Final classifier was based on RNA expression intensities of 45 genes.
- Performance evaluation of the classifier resulted in a cross-validated AUC of 57.9%, indicating weak discriminatory power of RNA expression in the eutopic endometrium for the diagnosis of endometriosis **(Figure 2)**.

Figure 1: Volcano plot (left) and p-value distribution (right) of feature-wise comparison of subjects with and without endometriosis by univariate moderated t-tests.

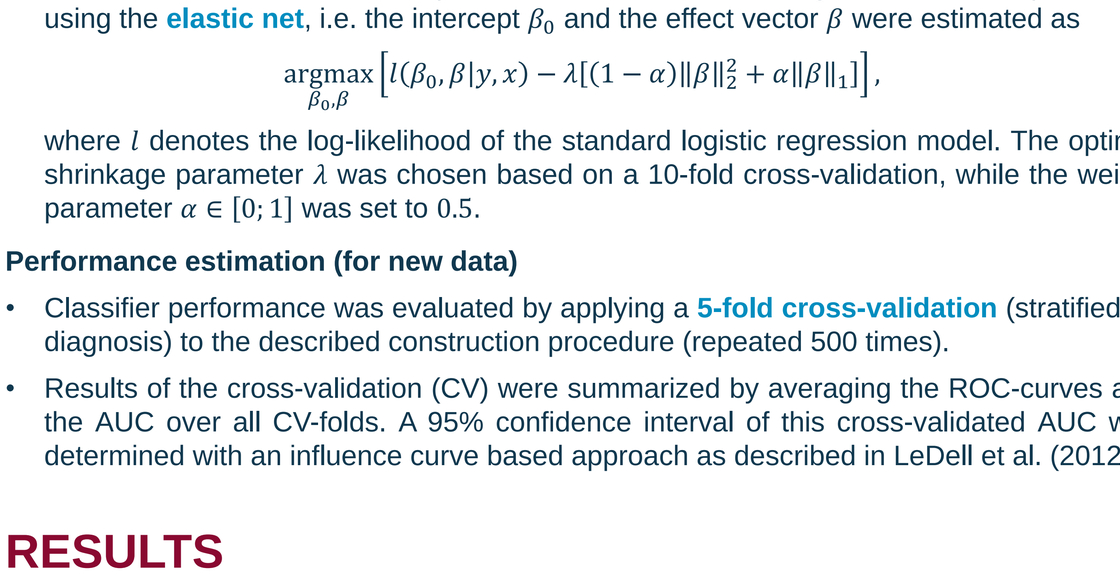
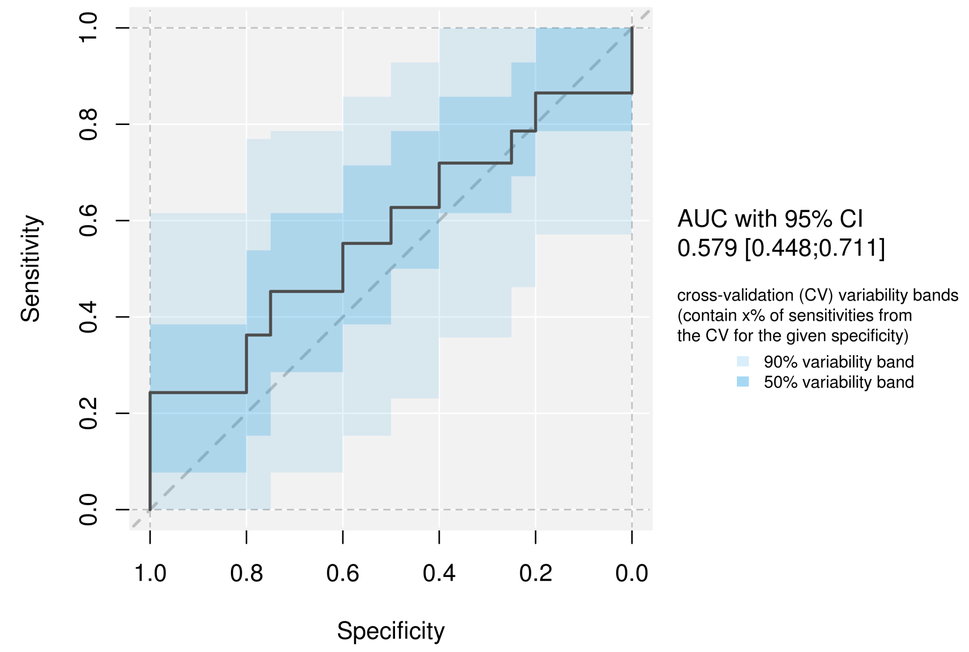


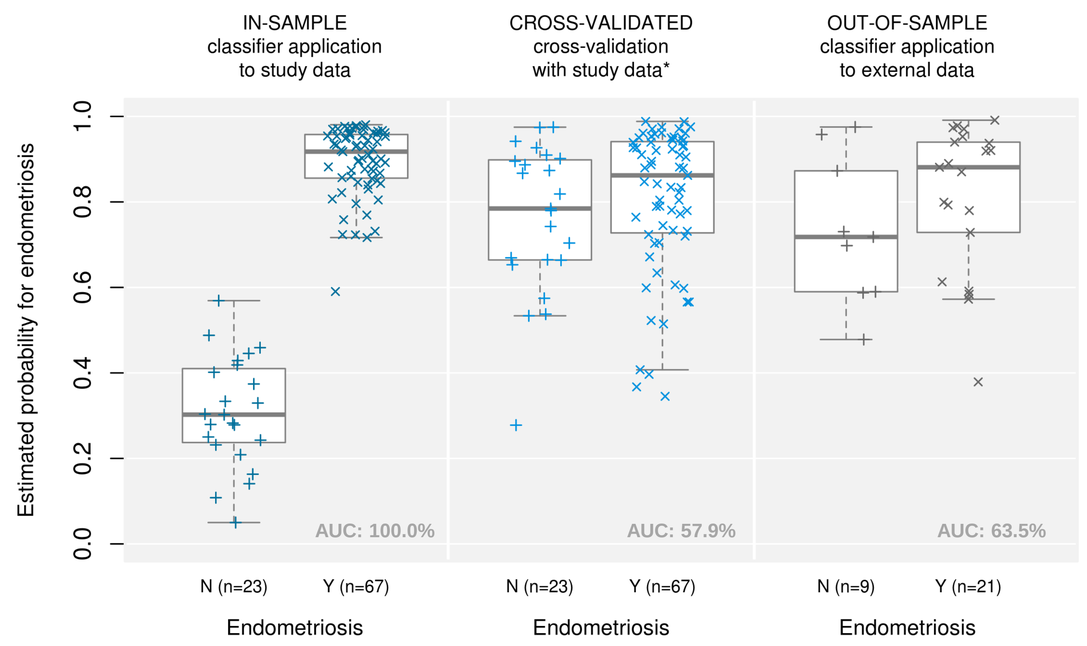
Figure 2: 5-fold cross-validated ROC curve of RNA expression classifier for the diagnosis of endometriosis (500 repeats).



VALIDATION WITH EXTERNAL DATA

- The classifier was applied to a comparable published data set shared on the GEO data base by Tamaresis et al. (2014).
- According to **Figure 3**, the out-of-sample classification performance in this independent data set is in line with results from the cross-validation of the study data (see also **Table 1** for details with regard to specific probability cutoffs)

Figure 3: Probabilities for the presence of endometriosis as estimated by the classifier



*) Mean estimated probabilities per subject over all cross-validation repeats

Table 1: Detailed classifier performance for various probability cutoffs

Cutoff criterion	Cutoff	Dataset	Sens.	Spec.	Acc.
Sensitivity ≥ 80%	0.687	Study data (cross-validated)	80.6%	34.8%	68.9%
		External data (out-of-sample)	76.2%	33.3%	63.3%
Closest topleft	0.787	Study data (cross-validated)	64.2%	52.2%	61.1%
		External data (out-of-sample)	66.7%	66.7%	66.7%
Youden index	0.910	Study data (cross-validated)	40.3%	82.6%	51.1%
		External data (out-of-sample)	42.9%	77.8%	53.3%
Specificity ≥ 80%	0.910	Study data (cross-validated)	40.3%	82.6%	51.1%
		External data (out-of-sample)	42.9%	77.8%	53.3%

CONCLUSION

- Differences in RNA expression in the eutopic endometrium between subjects with and without endometriosis are not sufficient for a reliable classification, indicating only **limited benefit** for the support of the diagnosis.
- This analysis emphasizes the **value of cross-validation techniques** when dealing with the classification of high-dimensional data. The estimation of the classifier performance by the cross-validation in the study data is confirmed by the external data, while the simple application of the classifier to the study data is clearly over-optimistic.

References

1. LeDell, E., Petersen, M. and van der Laan, M.: "Computationally Efficient Confidence Intervals for Cross-validated Area Under the ROC Curve Estimates." *U.C. Berkeley Division of Biostatistics Working Paper Series*, 304 (2012).
2. Efron, B., Tibshirani, R., Storey, J., and Tusher, V.: “Empirical Bayes Analysis of a Microarray Experiment.” *Journal of the American Statistical Association*, 96 (2001):1151–1160.
3. Tamaresis, J.S., Irwin, J.C., Goldfien, G.A., Rabban, J.T., Burney R.O., Nezhat, C., DePaolo, L.V. and Giudice, L.C.: "Molecular Classification of Endometriosis and Disease Stage Using High Dimensional Genomic Data." *Endocrinology* ,155 (2014): 4986-4999.

Presented at the PSI conference 2018 – Breaking Boundaries in Drug Development, 4. - 6. June 2018, Amsterdam, Netherlands