Interpretation of microarray data in cancer: a statistical viewpoint

S. Michiels

Unit of Biostatistics and Epidemiology
Institut Gustave Roussy

Institut de cancérologie
GUSTAVE ROUSSY
VILLEJUIF - www.igr.fr

Expression profiling lexicon

Microarray

❖ based on one biological sample, a microarray produces the measurement of the expression of 10 to 40 thousand genes for a cost of 500-1000 $

Gene signature

❖ combination of genes with different expressions in tumors having a different outcome

❖ searching for "the signature" predicting the risk of distant metastasis of breast cancer patients within 5 years after diagnosis implies that there is a unique molecular fingerprint for this risk.

The curse of the dimension

❖ Retrospective series with a small number of patients and thousands of variables

❖ Issues in statistical power, interpretation and validation of results

Main objectives of microarray studies

1) to identify homogeneous subtypes of a disease on the basis of gene expression

❖ By use of cluster analysis

❖ Even in the case of random noise, the technique will produce a cluster tree

❖ Experts agree that clusterisation has been overused in the microarray field (Allison Nat Genet 2006; Dupuy, Simon JNCI 2007)

2) to find genes that are differentially expressed in tumours with different characteristics

❖ Example: between tumours from 34 breast cancer patients who developed a distant metastasis within 5 years after surgery and tumours from 44 patients who did not (van't Veer et al Nature 2002)

3) to develop a rule based on gene expression allowing the prediction of patient prognosis or of the response to a particular treatment
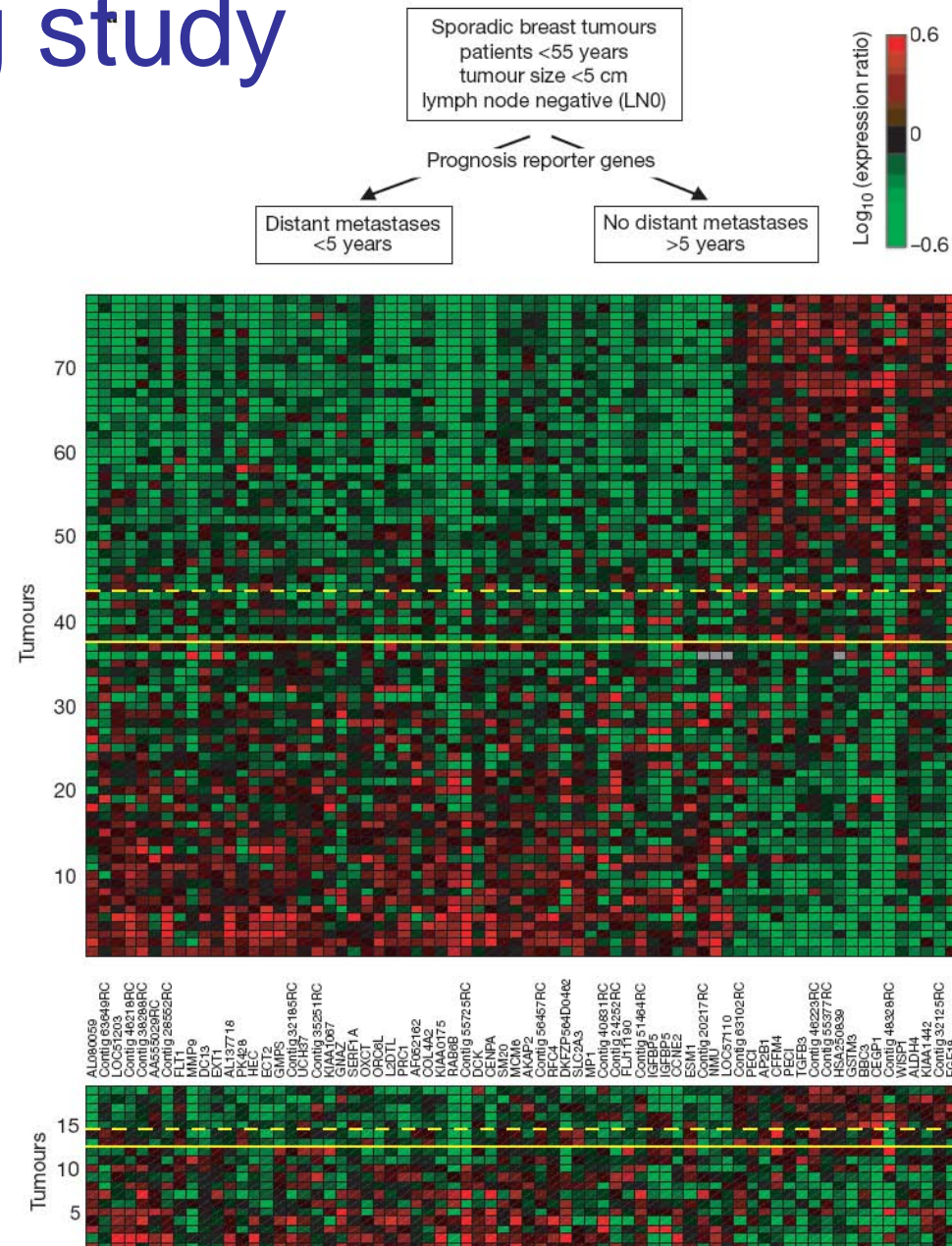
Pioneering study

Institut de cancérologie GUSTAVE ROUSSY
VILLEJUIF - www.igr.fr

van't Veer et al Nature 2002

❖ Training set: 78 pts, 34 with distant metastasis at 5 years

❖ 25 000 genes

❖ Ranked by correlation coefficient with binary metastatic status at 5 years

❖ « Molecular signature »

= top 70 prognostic genes

❖ Validation set: 19 pts

(Copyright Nature Publishing Group)

Multiple testing issue

False Discovery Rate (FDR)

❖ expected proportion of false positive genes among those declared as differentially expressed (Benjamini and Hochberg J R Stat Soc Ser B 1995)

At least four factors determine the FDR characteristics of a microarray study when comparing 2 groups of patients:

❖ the proportion of truly differentially expressed genes,

❖ the distribution of the true differences,

❖ measurement variability,

❖ sample size.

Institut de cancérologie
GUSTAVE ROUSSY
VILLEJUIF - www.igr.fr

FDR as a function of sample size

❖ Pilot experiment to estimate the proportion $p_1=1-p_0$ of truly differentially expressed genes and their non-zero effect size. Let's say $1-p_0=1\%$ and the non-zero effect sizes are equal to 1
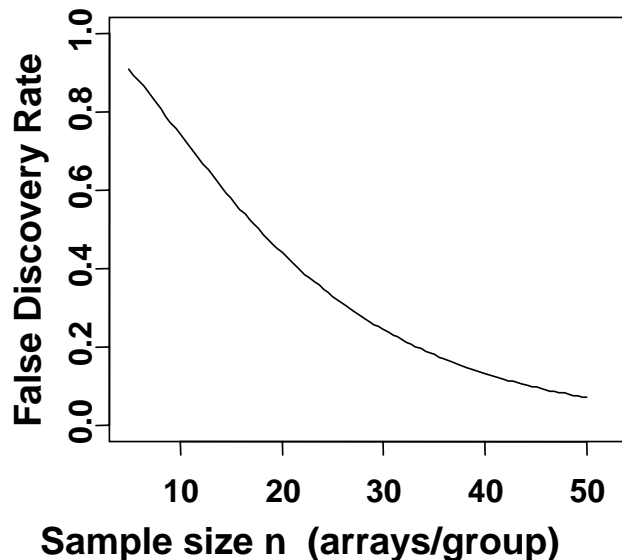
❖ After the experiment, declare the top 1% as "significant"

❖ With 5 pts per group, if 200 genes are truly differentially expressed among 20 000 genes, then if we take the 200 most discriminant genes between the 2 groups, we expect 182 false positives (FDR= 91%)

❖ With 56 pts per group: FDR=5%

$p_0=99\%$

Sample size n (arrays/group)

Pawitan, Michiels et al
Bioinformatics 2005

❖ When p_1 is smaller, or when the effect sizes are smaller, a larger sample size is needed to control for the FDR.

Prediction rules based on gene expression

Training – validation strategy
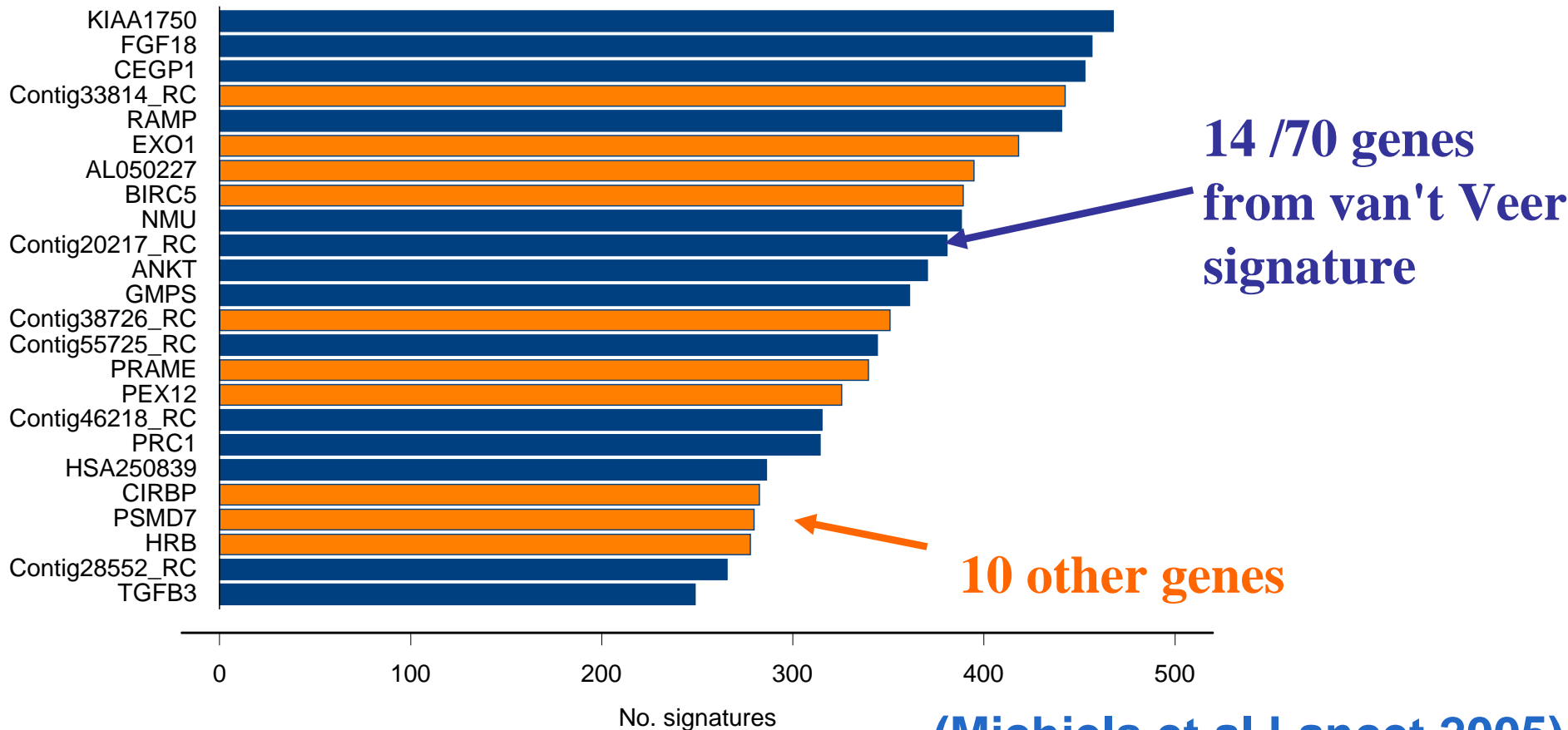
❖ Select the genes, equation and cut-off on a unique training set of patients

❖ Evaluate the performance of the prediction rule on a unique independent validation set of patients

Extension

❖ randomly assign patients between training and validation sets, and repeat (jackknife)

❖ Other possibilities: crossvalidation or bootstrap (Molinaro, Simon, Pfeiffer Bioinformatics 2005)

Instability of van't Veer's signature

Genes included in at least 250 out of 500 signatures for a training set size of 78

KIAA1750
FGF18
CEGP1
Contig33814_RC
RAMP
EXO1
AL050227
BIRC5
NMU
Contig20217_RC
ANKT
GMPS
Contig38726_RC
Contig55725_RC
PRAME
PEX12
Contig46218_RC
PRC1
HSA250839
CIRBP
PSMD7
HRB
Contig28552_RC
TGFB3

0   100   200   300   400   500

No. signatures

14 /70 genes from van't Veer signature

10 other genes

(Michiels et al Lancet 2005)

Proportion of misclassifications as a function of the training set size after repeated random sampling in van't Veer's study
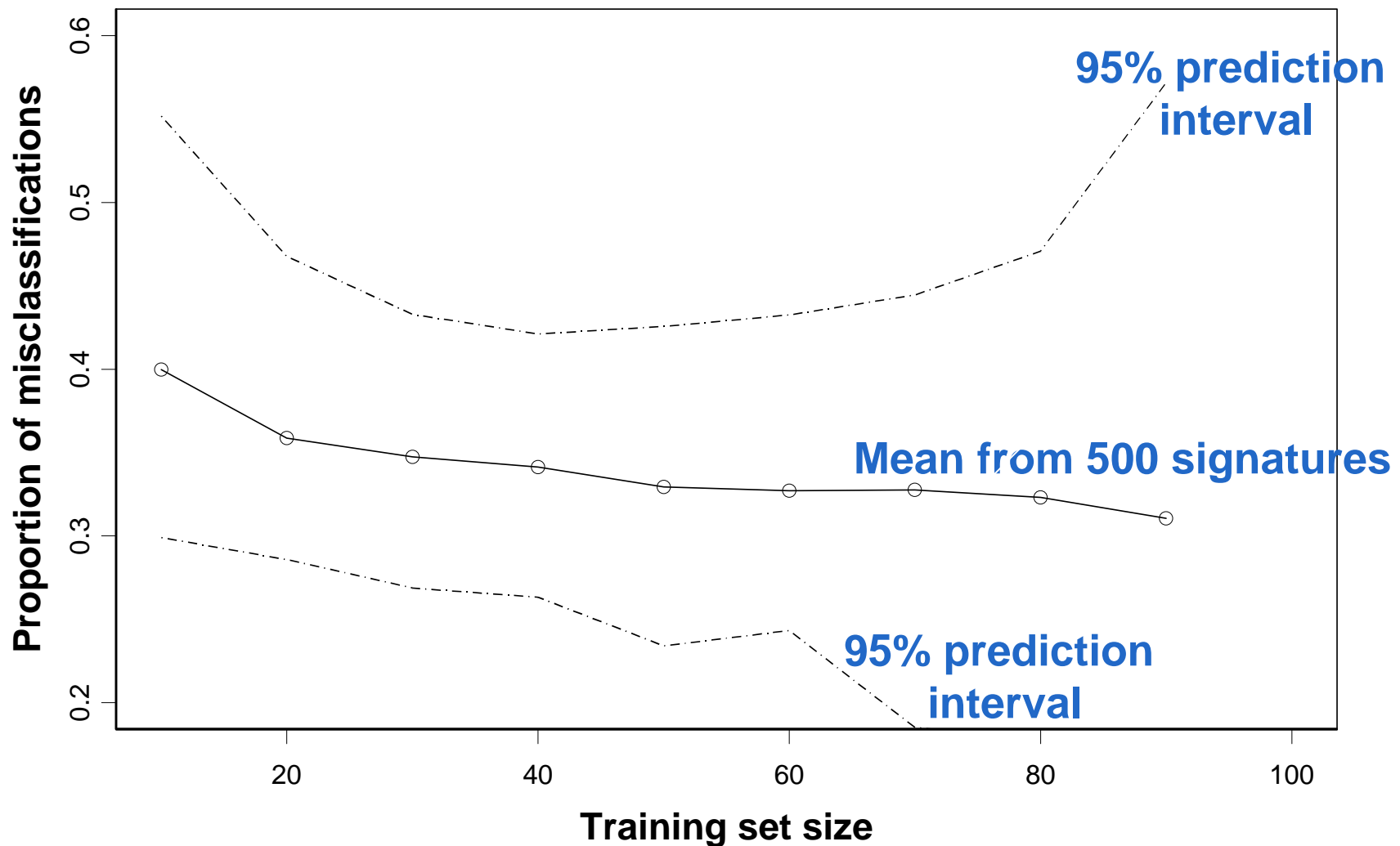
External validation

Definition of a validation study

- ❖ a study designed to confirm the results of a previous study, in order to reduce the play of chance and the potential for biases (Rahnsonoff Nat Rev Cancer 2004 and 2005)

Common mistakes in the oncology literature (Koscielny et al JCO 2005; Michiels, Hill NEJM 2007)

- ❖ To include part of the initial sample of patients in the validation study
- ❖ To include other type of patients in the validation study than in the initial sample
- ❖ To use another measurement technique (rt-PCR vs microarray)
- ❖ To change the prediction rule by adapting it to the new sample of patients through changing the list of genes, or the equation, or the cutoff

Same errors as those committed when studying just 1 tumour marker! (REMARK NCI-EORTC guidelines, McShane et al JNCI 2005)

On the other side of the Atlantic…

"The Recurrence Score assay … predicts the magnitude of chemotherapy benefit", Paik et al JCO 2006

NSABP B20 (**R**) → **Tamoxifen**

→ **Tamoxifen + chemotherapy**
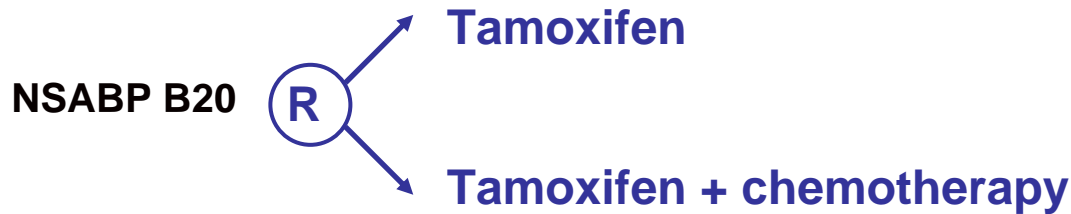
- ❖ In the development of the RS score, a large training set was used that included patients from the tamoxifen-only arm

- ❖ They observed that the RS score was a better predictor of recurrence-free survival in the tamoxifen-only arm as compared to the tamoxifen plus chemotherapy arm, and interpreted this result as a demonstration that the rule "predicts the magnitude of chemotherapy benefit'.

- ❖ A more obvious interpretation : **a prediction rule is optimal for the patients in the training set used for its construction !!** (Ioannidis Nat Clin Pract Onc. 2006; Michiels et al BJC 2007).

Validation of the pioneering study

Validation study 1 (Van de Vijver et al NEJM 2002)

❖ N=295 breast cancers (with and without nodal involvement), one single center

❖ Bias: inclusion of 61 pts of the pilot study

❖ Only 234 new pts

❖ Initial endpoint: distant metastasis-free survival status at 5 years

  - Se = 93% (CI$_{95\%}$ 81 to 99%)
  - Sp = 53% (CI$_{95\%}$ 44 to 61%)

Validation study 2 (TRANSBIG, Buyse et al JNCI 2006)

❖ N=307 breast cancers (without nodal involvement), multicentric study

❖ Distant metastasis-free survival status at 5-years:

  - Se = 90% (CI$_{95\%}$ 78 to 95%)
  - Sp = 42% (CI$_{95\%}$ 36 to 48%)

In Vitro Diagnostic Multivariate Assays (IVDMIA)

❖ Mammaprint from Agendia, Amsterdam is the first test to be approved as an IVDMIA by the FDA.

Class II Special Controls Guidance Document: Gene Expression Profiling Test System for Breast Cancer Prognosis : http://www.fda.gov/cdrh/oivd/guidance/1627.html

❖ In addition to a primary analysis using the results of your device, you should provide an analysis that demonstrates your device is "value added" and provides additional information concerning prognosis even after considering clinical data available to a physician. In breast cancer, there is information available from a variety of sources that provides prognostic value. (For example, the age of the patient, ER status, tumor size and grade, are routinely assessed.). You should provide information that demonstrates added prognostic value in comparison with routine information obtained in current clinical practice. A Cox regression model may be considered.

Added value of the gene signature?

What has been shown?

❖ Usual method : to include all known prognostic factors + the gene signature in the same Cox model

❖ Several gene signatures in early breast cancer have a very strong correlation with the tumour grade (Fan et al NEJM 2006)

❖ But: being a "significant" prognostic marker does not meaningfully describe a marker's ability to classify subjects (Pepe et al Amer J Epid 2004)

Specific measures of predictive accuracy

❖ Use of proportion of explained variation in Cox model, (Schemper, Henderson Biometrics 2000)

❖ For Mammaprint: all the data is available on the internet; for Recurrence Score: not…

Explained variation (EV) in validation 1

❖ N=234 pts (55 with distant metastases)

❖ Endpoint = distant metastasis-free survival, Cox model

(Dunkler, Michiels, Schemper. Eur J Cancer, 2007)

	EV $\pm$ SE*
Model without any factor	0%
Model with known prognostic factors: age, nodal involvement, oestrogen recepter status and tumour grade	16% $\pm$5%
Model with gene signature	12% $\pm$4%
Model with known prognostic factors + gene signature	19% $\pm$5%
Gain by adding the gene signature	**3% $\pm$5%**

*SE = standard error, estimated by 200 bootstrap samples

Conclusions

The search for gene signatures is based on the assumption that a clear distinction between tumours that will relapse and those that will not is possible using gene expression.

But

❖ The list of genes are highly unstable (do we care?)

❖ the actual performance of prediction rules using gene expressions is not as good as initially published

❖ The prediction rules using gene expression have not (yet?) provided a substantially and significantly improved prognostic classification when compared to conventional factors

Main reference for this presentation: **Michiels, Koscielny, Hill. BJC 2007**

Aristotle: clinical epidemiologist?

None of the arts theorise about **individual cases**. Medicine, for instance, does not theorise about what will help to cure Socrates or Callias, but only about **what will help to cure any or all of a given class of patients**. This alone is business: individual cases are so infinitely various that no systematic knowledge of them is possible.

Aristotle. Rhetoric. book I, chapter 2: 1356b

BMJ 2007;334:785