High dimensional data for supporting medical decision: are we meeting the challenge?

Elia Biganzoli, elia.biganzoli@unimi.it
Dept. of Medical Statistics and Biometry “G.A.Maccacaro”, Istituto Nazionale Tumori – Università degli Studi di Milano, Campus Cascina Rosa. Milano, Italy

Challenges in design, analysis and reporting of prognostic and predictive marker research – from single studies to an EBM based assessment

Organisers: Willi Sauerbrei and Doug Altman

Satellite Workshop to the Cochrane Colloquium, Freiburg, 8 October 2008
Omic study objectives

Three common types:

Class Discovery

Class Comparison

Class Prediction

Golub et al. 1999

*Molecular Classification of Cancer: Class Discovery and Class Prediction by gene expression Monitoring*
Traditional prognostic criteria in early N- breast cancer

• When they should receive adjuvant therapy?
• Balance between undertreatment of few patients vs overtreatment of the majority
• Definition of low risk patient group
  (...)The panel agreed that a population of patients who have less then a 10% chance of relapse i.e. 90% disease free survival probability (DFS) within 10 years, would not be candidates for receiving routine adjuvant systemic therapy. (...)
  • St.Gallen (Goldhirsch et al. JCO, 2003; Annals of Oncology, 2007):
    ER>0 and/or PgR>0 (no medullary, apocrine) and pT≤2 cm, grade I, age≥35 years, no extensive peritumoral vascular invasion, HER2/neu not overexpressed/amplified
  • NIH (Eifel et al. JNCI, 2001):
    pT≤1 cm
• Low accuracy: large number of patients overtreated
  high risk group: 70-80% DFS without adjuvant therapy
(1) Genomic prognostic criteria in early N- breast cancer

Need for improvement of prognostic accuracy

Ideally, without adjuvant therapy:

100% DFS probability for the low risk group
(a large group not needing to be treated)

0% DFS probability for the high risk group
(a small group needing to be treated)

cDNA microarray techniques for the simultaneous evaluation of the expression of tens thousand genes
(2) Genomic prognostic criteria in early N- breast cancer

Gene expression profiling for predicting clinical outcome
cDNA microarrays for the expression of 25000 genes on 78 patients
70 marker genes for identify low/high risk groups

Retrospective follow-up study on the low/high risk classificator on 70 genes (van de Vijver et al. NEJM, 2002)

90% 10 years DFS for the low risk group
45% 10 years DFS for the high risk group
(Is it fully satisfactory the prognostic accuracy gain?)
Issues on genomic prognostic criteria

Blind classification on 70 marker genes vs few established traditional prognostic factors.

Another classifier on the same data of van’t Veer et al. (2002) found 119 marker genes, showing similar performance but only 28 genes are shared.

(Lama N, Ambrogi F, Antolini L, Boracchi P and Biganzoli E, Some issues and perspectives in microarray data analysis in breast cancer: the need for an integrated research, 1st European Workshop on the Assessment of Diagnostic Performance, 7-9 July, Milano)


Another classifier on the same data but resorting to combined conventional biomarkers, showed similar performance of van’t Veer et al. (2002)

(Eden et al., Eur J Cancer, 2004)
Class Prediction Issues

Some issues and perspectives in microarray data analysis in breast cancer: the need for an integrated research

Nicolà Lama\textsuperscript{1}, Federico Ambrogi\textsuperscript{2}, Laura Antolini\textsuperscript{2}, Patrizia Boracchi\textsuperscript{1} and Elia Biganzoli\textsuperscript{2}

\textsuperscript{1}Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Via Venezian 1, 20133 Milano, Italy
\textsuperscript{2}Unità di Statistica Medica e Biometria, Istituto Nazionale per lo Studio e la Cura dei Tumori, Via Venezian 1, 20133 Milano, Italy

Michiels – Koscielny – Hill: Lancet 2005

Reanalysis of 7 studies that attempted to predict prognosis:
\begin{itemize}
  \item Instability of the set of genes included in the signature
  \item Five of the seven studies did not classify patients better than chance
  \item Studies with larger sample size are needed before gene expression profiling can be used in clinic.
\end{itemize}

It is unlikely that the oncologist would apply a decision criterion without clearly understanding its biological bases and its clinical value. This is the underlying risk of a black-box classification.
The instability of microarray data results is a major problem. It is hard to think that it could be solved on a pure statistical basis. Microarray studies have a major exploratory role.

Fig. E.1. Pooled MA-plot. $M$ versus $A$ plot for gene expression data pooled across the whole set of 24,481 genes for 78 arrays.
Challenges in Projecting Clustering Results Across Gene Expression-Profiling Datasets

Lara Lusa, Lisa M. McShane, James F. Reid, Loris De Cecco, Federico Ambrogi, Elia Biganzoli, Manuela Gariboldi, Marco A. Pierotti

Background
Gene expression microarray studies for several types of cancer have been reported to identify previously unknown subtypes of tumors. For breast cancer, a molecular classification consisting of five subtypes based on gene expression microarray data has been proposed. These subtypes have been reported to exist across several breast cancer microarray studies, and they have demonstrated some association with clinical outcome. A classification rule based on the method of centroids has been proposed for identifying the subtypes in new collections of breast cancer samples; the method is based on the similarity of the new profiles to the mean expression profile of the previously identified subtypes.

Methods
Previously identified centroids of five breast cancer subtypes were used to assign 99 breast cancer samples, including a subset of 65 estrogen receptor-positive (ER+) samples, to five breast cancer subtypes based on microarray data for the samples. The effect of mean centering the genes (i.e., transforming the expression of each gene so that its mean expression is equal to 0) on subtype assignment by method of centroids was assessed. Further studies of the effect of mean centering and of class prevalence in the test set on the accuracy of method of centroids classifications of ER status were carried out using training and test sets for which ER status had been independently determined by ligand-binding assay and for which the proportion of FR+ and FR− samples were systematically varied.

Results
When all 99 samples were considered, mean centering before application of the method of centroids appeared to be helpful for correctly assigning samples to subtypes, as evidenced by the expression of genes that had previously been used as markers to identify the subtypes. However, when only the 65 ER+ samples were considered for classification, many samples appeared to be misclassified, as evidenced by an unexpected distribution of ER+ samples among the resultant subtypes. When genes were mean centered before classification of samples for ER status, the accuracy of the ER subgroup assignments was highly dependent on the proportion of ER+ samples in the test set; this effect of subtype prevalence was not seen when gene expression data were not mean centered.

Conclusions
Simple corrections such as mean centering of genes aimed at microarray platform or batch effect correction can have undesirable consequences because patient population effects can easily be confused with these assay-related effects. Careful thought should be given to the comparability of the patient populations before attempting to force data comparability for purposes of assigning subtypes to independent subjects.

Fig. 1. Subtype prediction of data from the dataset of Sotiriou et al. (11) using method of centroids and gene expression of some genes representative of the subtypes. The dendrogram displays the results of hierarchical clustering of the complete dataset of Sotiriou et al. (11) (genes were median centered for the purpose of this clustering display). The distance metric used in the hierarchical clustering was one minus centered Pearson correlation, and linkage method was average linkage. The colored bars below the dendrogram represent the predicted subtype results obtained applying method of centroids mean centering the genes (Centered) and without mean centering the genes (Noncentered) in the Sotiriou et al. (11) dataset; the full dataset and a dataset restricted to estrogen receptor-positive (ER+) cases only were considered. The colors used to represent the subtypes are dark blue for luminal A, light blue for luminal B, pink for ERBB2+, red for basal and green for normal subtype. The expression of the genes was color coded using colors ranging from green (for low relative expression) to red (high relative expression). The first three genes shown in the bottom panel should be more expressed in the subtype that they represent (ERBB2 for ERBB2+ subtype, ESR1 for luminal A subtype, KRT5 for basal subtype). The set of proliferation genes (MYBL2, BUB1, TOP2A, and CENPF) should be highly expressed in basal and luminal B subtypes but not in the normal subtype. The gene expression of all the genes included in the intrinsic gene set is reported in Supplementary Fig. S1 (available online).
BIDIMENSIONAL ELECTROPHORESIS (2-DE)

Low reproducible: analytical and biological variability

Class comparison experiments

E.Biganzoli, Freiburg, 2008
Proteomics: Multiplex Cytokine Array

Assay Method
multiplex sandwich ELISA for the quantitative measurement of 9 cytokines in serum samples. Each well of the microplate provided is pre-spotted with cytokines/angiogenetic factors-specific antibodies. The enzyme substrate reaction produces a luminescent signal.

Study Design
In plate arrays we detected 9 cytokines on 150 samples from patients with diagnosis of myocardial infarction. Furthermore, on all samples we measured Il-6, IL-10, IL-12 (using a Gold Standard method). We compared the results for these three cytokines evaluated by the two assays.
Proteomics: Multiplex Cytokine Array

**Assay Method**
multiplex sandwich ELISA for the quantitative measurement of 9 cytokines in serum samples. Each well of the microplate provided is pre-spotted with cytokines/angiogenetic factors-specific antibodies. The enzyme substrate reaction produces a luminescent signal.

**Study Design**
In plate arrays we detected 9 cytokines on 150 samples from patients with diagnosis of myocardial infarction. Furthermore, on all samples we measured IF-6, IL-10, IL-12 (using a Gold Standard method). We compared the results for these three cytokines evaluated by the two assays.

**Results**
1) Available plate array imaging evidenced differences between plates, performed at a time distance of six months, possibly due to matrix effects.
4) The comparison with the Gold standard method showed relevant differences with even detectable/high values for those samples in which analytes resulted undetectable by the multiplex assay.

**Conclusion**
from our experience, the sensitivity of the system to matrix effect, the aspecific binding of the antibody, provided different results strictly associated to the plate array employed. This heavily affected the reproducibility of results.

E.Biganzoli, Freiburg, 2008
A different perspective in the development of prognostic criteria

Old and new markers for breast cancer prognosis: the need for integrated research on quantitative issues

Elia Biganzoli, Patrizia Boracchi

\(^{a}\) Istituto Nazionale per lo Studio e la Cura dei Tumori, Unità di Statistica Medica e Biometria, Via G. Venezian 1, I-20133, Milano, Italy

\(^{b}\) Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Milano, Italy

Received 15 April 2004; accepted 23 April 2004

Overall, the underlying issue in microarray studies is the lack of standard methods for design, data analysis, and performance assessment according to clinical aims. Achieving these goals requires cooperative efforts between multidisciplinary research networks.
A different perspective in the development of prognostic criteria


• Past research on single biomarkers with high prognostic power failed? (Cyclin-E still promising? Keyomarsi et al, NEJM, 2003)

• Combined use of clinical factors and biomarkers would improve prognostic accuracy

• Complex statistical models for exploring the biology of the disease (tumor bioprofiling and recurrence dynamics) to derive simple prognostic criteria for supporting medical decision (need to understand the clinical/biological reasons of therapy counselling)

E.Biganzoli, Freiburg, 2008
A different perspective in the development of prognostic criteria


• Past research on single biomarkers with high prognostic power failed? (Cyclin-E still promising? Keyomarsi et al, NEJM, 2003)

• Combined use of clinical factors and biomarkers would improve prognostic accuracy

• Complex statistical models for exploring the biology of the disease (tumor bioprofiling and recurrence dynamics) to derive simple prognostic criteria for supporting medical decision (need to understand the clinical/biological reasons of therapy counselling)

• The St.Gallen and NIH criteria may be too simple! (too low prognostic accuracy)
A changing perspective in the need of prognostic criteria: steroid receptors

(Goldhirsch et al. JCO, 2003)

Gene expression profiling support different biologic profiles according to steroid receptors. Clinical studies provide evidence that receptor-absent disease is different from that with even low levels of receptor expression.
A changing perspective in the need of prognostic criteria: steroid receptors

(Goldhirsch et al. JCO, 2003)

Gene expression profiling support different biologic profiles according to steroid receptors.
Clinical studies provide evidence that receptor-absent disease is different from that with even low levels of receptor expression.

Change of laboratory reporting strategies
(no longer arbitrary cutoff values!)

E.Biganzoli, Freiburg, 2008
A changing perspective in the need of prognostic criteria: steroid receptors

(Goldhirsch et al. JCO, 2003)

Gene expression profiling support different biologic profiles according to steroid receptors. Clinical studies provide evidence that receptor-absent disease is different from that with even low levels of receptor expression.

Change of laboratory reporting strategies (no longer arbitrary cutoff values!)

At the other extreme, patients with strong receptor expression may be effectively treated with endocrine therapy alone.
A changing perspective in the need of prognostic criteria: steroid receptors

(Goldhirsch et al. JCO, 2003)

Gene expression profiling support different biologic profiles according to steroid receptors.
Clinical studies provide evidence that receptor-absent disease is different from that with even low levels of receptor expression.
Change of laboratory reporting strategies (no longer arbitrary cutoff values!)
At the other extreme, patients with strong receptor expression may be effectively treated with endocrine therapy alone.

A gradation of levels between the two extremes (Need of risk profiling and predicting the response to therapy)
Flexible Modelling in survival data analysis

From knowledge…

- Exploratory study of disease dynamics on the basis of censored survival data.
  - Soft assumptions on covariate effects.
  - Non-linear, non-additive and time dependent continuous covariate effects.

...to decision

- Individual risk profiling from disease features for supporting clinical decision making
A two-phase perspective

• “All models are wrong, some, though are better than other”
• but...for doing what?

• Knowledge phase: exploratory modelling
  • Check or generation of biological or clinical hypotheses

• Decision phase: predictive modelling
  • Support medical decision
Knowledge vs Prediction

- Building a predictive model without previous exploratory studies is difficult
- Conducting both phases in the same study is difficult
- The assessment of prognostic factors in breast cancer is a typical example
  - many markers studied (often resorting to arbitrary cutoff values)
  - very few validated for clinical routine.
  - high costs and increased risks in the genomic era

E.Biganzoli, Freiburg, 2008
Knowledge from exploratory analysis

• Even for traditional clinical factors measured on a continuous scale is a recent acquisition.
• Example: age and breast cancer recurrence

![Graphs showing hazard ratio vs age](British Journal of Cancer (1999) 79(11/12), 1752–1760)

![Graphs showing hazard ratio vs age](Journal of Clinical Oncology, Vol 18, No 14 (July), 2000: pp 2702-2709)

E. Biganzoli, Freiburg, 2008
Knowledge from exploratory analysis


Peaks in breast cancer recurrence dynamics and the “tumour dormancy” hypothesis
Flexible estimation of competing risk (Regression Splines)

Joint modelling of cause-specific hazard functions with cubic splines: an application to a large series of breast cancer patients

Patrizia Boscacci, Elio Biganzoli, Ettore Marebiani

*Institute of Statistics, Modena and Reggio Emilia, Università degli Studi di Modena, Italy

†Institute of Statistics, Modena and Reggio Emilia, Istituto Nazionale per la Studio e la Cura dei Tumori, Via G. Venezian 1, 20133, Milan, Italy

Received 1 November 2000; received in revised form 1 February 2002

Fig. 2. Estimated marginal cause-specific hazards (CSH) for in-breast tumor recurrences (IBTR) and distant metastases (DM). Spline estimates for the selected model with one knot at 40 months are reported as continuous lines (IBTR-S40 and DM-S40). Spline estimates for the model with one knot at 22 months, corresponding to a local minimum of AIC, are reported as dotted lines (IBTR-S22 and DM-S22). Bootstrap pointwise 99% confidence intervals are also reported as dashed lines.

Fig. 3. Estimated marginal cause-specific hazards (CSH) for in-breast tumor recurrences (IBTR) and distant metastases (DM). Spline estimates are reported as continuous lines (IBTR-S40 and DM-S40 for the selected spline model with one knot at 40 months; IBTR-Q and DM-Q for the spline model with four knots located at the percentiles of the unmasked time distribution, respectively). Non-parametric kernel estimates are reported as dashed lines (IBTR-K and DM-K, respectively).
Research article

Recurrence dynamics does not depend on the recurrence site
Romano Demicheli¹, Elia Biganzoli²,³, Patrizia Boracchi³, Marco Greco⁴ and Michael W Retsky⁵

This article is online at: http://breast-cancer-research.com/content/10/5/R83

---

<table>
<thead>
<tr>
<th>degrees of freedom</th>
<th>AIC (No interaction)</th>
<th>AIC (Interaction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>234.9</td>
<td>212.2</td>
</tr>
<tr>
<td>5</td>
<td>224.8</td>
<td><strong>208.4</strong></td>
</tr>
<tr>
<td>6</td>
<td>225.2</td>
<td>221.2</td>
</tr>
<tr>
<td>7</td>
<td>227.4</td>
<td>218.8</td>
</tr>
<tr>
<td>8</td>
<td>228.1</td>
<td>224.5</td>
</tr>
<tr>
<td>9</td>
<td>230.2</td>
<td>231.2</td>
</tr>
<tr>
<td>10</td>
<td>230.9</td>
<td>235.6</td>
</tr>
</tbody>
</table>

---

E.Biganzoli, Freiburg, 2008
Knowledge from exploratory analysis

Log-logistic Regression Models for Survival Data

By STEVE BENNETT
University of Reading, UK

Modelling covariate effects on failure dynamics

E.Biganzoli, Freiburg, 2008
Knowledge from exploratory analysis

Partial Logistic ANN (PLANN) modelling continuous interactions (time, covariates)

FEED FORWARD NEURAL NETWORKS FOR THE ANALYSIS OF CENSORED SURVIVAL DATA: A PARTIAL LOGISTIC REGRESSION APPROACH

ELIA BIGANZOLI1, PATRIZIA BORACCHI2, LUIGI MARIANI1 AND ETTORE MARUBINI1,2

1Dipartimento di Statistica Medica e Biomimetica, Istituto Nazionale per lo Studio e la Cura dei Tumori, Via Venezian, 1, 20133 Milano, Italy
2Istituto di Statistica Medica e Biomimetica, Università degli Studi di Milano, Via Venezian 1, 20133 Milano, Italy

E. Biganzoli, Freiburg, 2008
Women with primary invasive breast cancer (N=633), who underwent surgery between January 1983 to December 1992. (archieved at the Pathology dept. of the University of Ferrara)

**Tumour markers** *(by IHC)*:

- ER
- PgR
- Ki-67/MIB1
- c-ErbB-2/NEU
- p53

Aim: clustering of tumours according biomarkers, for the search of underlying cancer subtypes and their clinical features.

Study of conditional independencies among biomarkers

---

Conditional independence relations among biological markers: a case study on breast cancer biomarkers

F. M. Stefanini\(^1\) and E. Biganzoli\(^2\)

\(^1\) Dipartimento di Statistica ‘G. Parenti’ Università degli Studi di Firenze, viale Morgagni 59, 50134 Firenze, Italy, stefanini@ds.unifi.it

\(^2\) Unità di Statistica Medica e Biometria, Istituto Nazionale per la Cura e lo Studio dei Tumori, via Venezian 1, 20133 Milano, Italy, biganzoli@istitutotumori.mi.it
Conclusions
Integration of exploratory studies (knowledge phase) with predictive ones to support clinical decision
Use of biological knowledge in a clinical context (biological bases need to be understood for medical decision making)
Conclusions

Integration of exploratory studies (knowledge phase) with predictive ones to support clinical decision.

Use of biological knowledge in a clinical context (biological bases need to be understood for medical decision making).

According to Huang et al. (Lancet, 2003):

“Genomic data will not replace traditional clinical factors but will add substantial detail”
Conclusions

Integration of exploratory studies (knowledge phase) with prospective ones to support clinical decision

Exploitation of biological knowledge in a clinical context (biological basis for the decision need to be understood)

According to Huang et al. (Lancet, 2003):

“Genomic data will not replace traditional clinical factors but will add substantial detail”

Need for innovative prognostic studies to integrate traditional factors with bioprofiles, allowing for simple but more accurate decision rules.
Conclusions

Integration of exploratory studies (knowledge phase) with prospective ones to support clinical decision
Exploitation of biological knowledge in a clinical context (biological basis for the decision need to be understood)
According to Huang et al. (Lancet, 2003): “Genomic data will not replace traditional clinical factors but will add substantial detail”
Need for innovative prognostic studies on traditional factors to better use them, allowing for simple but more accurate decision rules.
Need for the cooperation of clinicians, biologists, bioinformaticians and biostatisticians, looking for shared semantics.

E.Biganzoli, Freiburg, 2008
ADDITIONAL REFERENCES