ELSEVIER

European Journal of Cancer 40 (2004) 1803–1806

European Journal of Cancer

www.ejconline.com

Editorial Comment

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

Elia Biganzoli [a], Patrizia Boracchi [b]

[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, Milan, Italy*

Received 15 April 2004; accepted 23 April 2004

Randomised clinical trials have shown an overall benefit for both adjuvant chemotherapy and endocrine therapy to treat breast cancer, independent of the patient's lymph node status. As adjuvant therapies have costs and side-effects associated with them, various prognostic criteria have been established to guide the indications for such treatments [1]. Particularly for patients at a low-risk of recurrence, a balance must be struck between the under-treatment of the few patients who are destined to relapse versus the unnecessary over-treatment of the majority who are already cured by loco-regional therapy. Illustrative of this problem are the disparate results of recent consensus conferences, which have attempted, using simple criteria, to define "low-risk" patients for whom adjuvant chemotherapy should not be recommended. The St. Gallen criterion [2] indicates tumours with oesterogen receptor (ER) and/or progesterone receptor (PgR) expression, and all of the following features: pT $\leqslant$ 2 cm, grade I, age $\geqslant$ 35 years should be considered in this low-risk category, whereas the National Institutes of Health (NIH) criterion [3] proposes tumours of pT $\leqslant$ 1 cm should be included in this group. Obviously, the few "low-risk" patients who would have relapsed after loco-regional therapy alone stand to benefit from adjuvant therapy, atleast potentially. However, the application of these criteria would result in systemic therapy being assigned to a large number of patients who do not really need it. By the same token, within the moderate/high-risk groups, patient outcome is also highly variable and depends on biological differences among the tumours.

There are strong expectations that biological tumour markers could improve prognostic assessments and help to better discriminate which subjects are likely to respond to the various tailored systemic treatments. These expectations have been increased following the introduction of techniques that can simultaneously evaluate the expression of large numbers of tumour genes. However, despite this continuous progress in the molecular biology of breast cancer, clinical decision-making still largely relies on pathological staging, grading and steroid receptor status. Consequently, there seems to be an increasing gap between the resources employed for basic and translational research and actual patient benefits and overall social gain.

In this issue of the *European Journal of Cancer*, Edén and colleagues [4] re-analyse the data of a previous study of van't Veer and colleagues [5]. The authors ask whether "good old" clinical markers, if optimally analysed, might have a similar discriminating power with regard to breast cancer prognosis as the microarray gene expression profilers. The question is a relevant one. Why have so few studies addressed this issue before? In the past, translational research focused on single biological markers which could putatively discriminate patients' prognosis or treatment response. Now, in the genomic era, our perspectives have changed completely, so that we now stress the limited prognostic power of single genes, whilst highlighting the need for a prognostic classification based an optimised, computer-aided analysis of many genes [5]. However, Edén and colleagues argue that conventional biomarkers could still be useful if analysed with advanced statistical methodologies, collecting information derived from these conventional markers to determine individual prognosis [6]. Perhaps the weak link in the approach of the Lund group is their faith in the reproducibility of conventional prognostic factors, since very few of these factors undergo quality assessment and control. Consequently, one might speculate that resources could be more profitably spent for the promotion of quality assurance programmes for conventional factors rather than for c-DNA microarray research.

The basic problem is whether the data from the original study of van't Veer and colleagues and the

E-mail address: biganzoli@institutotumori.mi.it (E. Biganzoli).

0959-8049/$ - see front matter
doi:10.1016/j.ejca.2004.04.026

re-analysis of Edén and colleagues are suitable for developing new prognostic classifications and/or comparing them with those already available. Quantitative issues, related to the development and assessment of prognostic classification based on genomic data, have been discussed elsewhere in [7]. Herein, further aspects related to the evaluation of the results of the above studies are presented. We point out the need to carefully consider the clinical relevance of the gene expression signature. In our Appendix below, we show the simple quantitative considerations underlying criteria, like those of St. Gallen and NIH, compared with the new ones. Studies such as those of van't Veer and colleagues and Edén and colleagues only provide statistical evidence of the accuracy of the criteria, calculated as sensitivity/specificity or related measures. Therefore, an assessment of their clinical value is needed in terms of the probability of relapse, as a function of the assigned class (positive/negative predictive values of the test). Overall, such issues support the need for cohort studies [8], planned *ad hoc*, following these initial studies. Validations should be subsequently performed in prospective studies with appropriate designs [9], targeted to provide answers in the clinical decision-making process. Studies like the MINDACT trial by the Breast International Group network, coordinated by the European Organisation for Research and Treatment of Cancer (EORTC) and funded by the European Commission (VI Framework Program) will hopefully address these aspects of genomic classification [10].

The need for integrating these initial exploratory studies addressing relevant biological issues (knowledge phase) with subsequent prospective clinical studies (decision phase) must be carefully considered to exploit biological knowledge in a clinical context. It is unlikely that the oncologist would apply a decision criterion without clearly understanding its biological basis, but this is the underlying risk of a blind classification based on multiple genes. This may be particularly relevant when trying to predict the response to therapy. Other exploratory studies have come to the same conclusions: focusing on individual patient prognosis, Huang and colleagues [11] stated that:

> (...)Genomic data will not replace traditional clinical factors, but will add substantial detail to this clinical information ...;

a different view from that of van't Veer and colleagues or similar reports.

According to Edén and colleagues, innovative prognostic studies on traditional factors could provide relevant benchmarks for assessing the real additional gain from the application of genomic techniques, if performed on suitable case series. This could lead to a high level of evidence. Quoting the World Health Organisation (WHO):

> Some of the claims for the medical benefits of genomics have undoubtedly been exaggerated, particularly with respect to the time-scale required for them to come to fruition. Because of these uncertainties, it is vital that genomic research is not pursued to the detriment of well-established methods of clinical practice and clinical and epidemiological research. Indeed, for its full exploitation it will need to be integrated into clinical research involving patients and into epidemiological studies in the community. It is crucially important that a balance is maintained in medical practice and research between genomics and these more conventional and well tried approaches.

A rapid increase in the number of studies on markers identified by means of high throughput genomic techniques, at considerable expense is likely. It would therefore be relevant to promote the application of suitable study designs and statistical methods for the reliable assessment of data collected on tumour markers, either genomic or "old", and a faster translation of basic research to medical decision-making. These goals can be most successfully met through the cooperation of clinicians, biomedical informaticians and biostatisticians. This is a challenge for the future of large European Union (EU) projects and Network of Excellences, like BIOINFOMED and BIOPATTERN.

Acknowledgements

This work was partially supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC) and EU (FP6-2002-IST-1 No. 508803). The authors thank N. Lama for his help and useful discussion.

Appendix A. Evaluation of clinical relevance: some quantitative considerations

The St. Gallen consensus meeting [2] defined risk categories for patients with node-negative breast cancer according to the leading criterion:

> (...) The panel agreed that a population of patients who have less then a 10% chance of relapse within 10 years, would not be candidates for receiving routine adjuvant systemic therapy. (...)

Patients within a "minimal-risk" category were expected to have at least a 90% disease-free survival (DFS) probability at 10 years of follow-up. From the perspective of diagnostic testing applied to prognosis, this approach is equivalent to setting the negative predictive value (NPV) of the classification (i.e. the conditional probability of being disease-free given a negative test result) at 0.9. A recent proposed validation of the St. Gallen criterion [12] estimated an 85% DFS at 8 years for the low-risk group, but a 73% DFS for those given a high-risk group classification (i.e. $1–0.73 = 0.27$ positive predictive value (PPV) of the test, which is the probability of developing the disease given a positive test result). Therefore, such a low

PPV implicitly supports the issue of over-treatment. However, following the same reasoning, it would appear that Keyomarsi and colleagues [13], reported values of more than 95% for both the NPV and the PPV for breast cancer death at 10 years. This is a different, but related, endpoint, referring to only one marker, cyclin E. This suggests, if confirmed, that cyclin E could be a very strong discriminator. Borg and colleagues also quoted the study Keyomarsi who showed the paradox that the expression of a single factor measured by Western blotting could outweigh the expression profiles of a thousand genes from a sophisticated data analysis [14].

The choice of a 10% misclassification threshold in the group of recurring patients, corresponding to 90% sensitivity in the studies of van't Veer and colleagues and Edén and colleagues does not directly follow the St. Gallen and NIH criteria. In fact, the NPV and PPV relate to the specificity and sensitivity, but cannot be calculated in such studies, since their design did not allow the incidence of disease relapse over follow-up time to be estimated. Therefore, in the present context, sensitivity/specificity, (or equivalently the area under the receiver operating curves (ROC), which can also be interpreted as an averaged sensitivity) are suitable for providing a first statistical evidence of the accuracy of prognostic classification criteria. Clinical relevance is more related to the NPV and PPV, as confirmed by the approach followed for the development of the St. Gallen and NIH criteria.

According to Simon and colleagues [7], an unbiased estimate of specificity of the criterion proposed by van't Veer and colleagues, fixing sensitivity at 0.9, was 0.59 (0.43–0.74 95% exact Confidence Interval, (CI)), i.e. one minus the false-positive fraction (18/44). This result has been confirmed in a subsequent study from the same research group [9], which showed (Table 2) a specificity of $32/55 = 0.58$ (0.44–0.71, 95% CI) on new node-negative patients, given an estimated sensitivity of 11/12, 0.92 (0.61–1, 95% CI). The same study provided an estimate of the NPV of approximately 0.90 and PPV of approximately 0.55 for the 10 years metastasis-free rate using gene-expression profiling data (Fig. 2 c) vs. 0.75 and 0.42 using the St. Gallen criterion (Fig. 3 b), respectively. The reliability of such estimates and the relevance of improvements in gene-expression profiling should be carefully assessed, both on a statistical and clinical basis. For the same reasons, the point estimates of the area under the ROC provided by Edén and colleagues [5] should be supplied with CI.

The use of sensitivity/specificity measures was previously advocated for the evaluation of the results of van de Vijver and colleagues [15], who replied that survival curves provide the most relevant information. Nevertheless, it should be pointed out that the Kaplan–Meier non-parametric estimator is quite variable when used for small sample sizes, therefore the calculation of PPV and NPV based on such curves, may have large CI. Moreover, the use of statistical significance for assessing the quality of the prognostic classification by testing for the difference between survival curves is meaningless from the viewpoint of clinical interpretation. All of the above considerations assume the Kaplan–Meier estimator was correctly adopted under uninformative censoring. However, the problem of events should be considered. These compete with the different endpoints considered (distant metastases, disease recurrence, death from breast cancer). Such problems call for different estimation procedures that are suitable for assessing competing risks.

Anyway, the application of the Kaplan–Meier estimator provided by Edén and colleagues [4] is misleading. In fact, in the original case series, patients are specifically selected when they recurred within 5 years of follow-up and compared with patients who were chosen because they had been followed for at least 5 years without disease recurrence. In the original study of van't Veer and colleagues [5], a correct estimation of the incidence of disease recurrence over 5 years of follow-up was not planned. Therefore, even if in the original sampling cohort no patients without recurrence were lost to follow-up before 5 years, (which is unlikely), the composition of this data-set appears to be artificially fixed to a ratio of patients who recurred or not recurred, that is likely different to the actual ratio in the studied population. Moreover, according to the plotted curves, the reader has the wrong perception of: (i) the absence of events after five years and (ii) a fictitious separation between curves. It should be noted that the exclusion of patients lost to follow-up before 5 years could lead to biased estimates and this is a well-known problem in the analysis of censored survival data. Other potential sources of bias have been already indicated in [16].

References

1. Boracchi P, Biganzoli E. Markers of prognosis and response to treatment: ready for clinical use in oncology? A biostatistician's viewpoint. *Int J Biol Markers* 2003, **18**, 65–69.
2. Goldhirsch A, Wood WC, Gelber RD, Coates AS, Thurlimann B, Senn HJ. Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. *J Clin Oncol* 2003, **21**, 3357–3365.
3. Eifel P, Axelson JA, Costa J, Crowley J, Curran Jr WJ, Deshler A, *et al.* National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. *J Natl Cancer Inst* 2001, **93**, 979–989.
4. Edén P, Ritz C, Rose C, Fernö M, Peterson C. "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer* 2004, this issue (doi:10.1016/j.ejca.2004.02.025).
5. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, **415**, 530–536.
6. Biganzoli E, Boracchi P, Coradini D, Grazia Daidone M, Marubini E. Prognosis in node-negative primary breast cancer: a

neural network analysis of risk profiles using routinely assessed factors. *Ann Oncol* 2003, **14**, 1484–1493.
7. 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**, 14–18.
8. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *New Engl J Med* 2002, **347**, 1999–2009.
9. Sargent D, Allegra C. Issues in clinical trial design for tumor marker studies. *Semin Oncol* 2002, **29**, 222–230.
10. Cardoso F. Microarray technology and its effect on breast cancer (re)classification and prediction of outcome. *Breast Cancer Res* 2003, **5**, 303–304.
11. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, *et al*. Gene expression predictors of breast cancer outcomes. *Lancet* 2003, **361**, 1590–1596.
12. Colomer R, Vinas G, Beltran M, Izquierdo A, Lluch A, Llombart-Cussac A, *et al*. Spanish Breast Cancer Research Group. Validation of the 2001 St. Gallen risk categories for node-negative breast cancer using a database from the Spanish Breast Cancer Research Group (GEICAM). *J Clin Oncol* 2004, **22**, 961–962.
13. Keyomarsi K, Tucker SL, Buchholz TA, Callister M, Ding Y, Hortobagyi GN, *et al*. Cyclin E and survival in patients with breast cancer. *New Engl J Med* 2002, **347**, 1566–1575, Erratum in: New Engl J Med 2003, **348**, 186.
14. Borg A, Ferno M, Peterson C. Predicting the future of breast cancer. *Nat Med* 2003, **9**, 16–18.
15. Helmbold P, Haerting J, Kolbl H. Gene-expression signatures in breast cancer. *New Engl J Med* 2003, **348**, 1715–1717, author reply 1715-7.
16. Caldas C, Aparicio SA. The molecular outlook. *Nature* 2002, **415**, 484–485.