Annals of Oncology 23: 2997–3006, 2012
doi:10.1093/annonc/mds586

Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement†

S. Guiu[1], S. Michiels[2], F. André[3]*, J. Cortes[4], C. Denkert[5], A. Di Leo[6], B. T. Hennessy[7], T. Sorlie[8], C. Sotiriou[9], N. Turner[10], M. Van de Vijver[11], G. Viale[12], S. Loi[13]* & J. S. Reis-Filho[14]

[1]Department of Medical Oncology, Georges-François Leclerc Cancer Center, Dijon, France; [2]Department of Biostatistics and Epidemiology, Jules Bordet Institute, Brussels, Belgium; [3]Department of Medical Oncology, Gustave Roussy Institute, Villejuif, France; [4]Department of Oncology, Vall d'Hebron Institute of Oncology, Barcelona, Spain; [5]Institute of Pathology, Charité University Medicine, Berlin, Germany; [6]Medical Oncology Unit, Hospital of Prato, Istituto Toscani Tumori, Prato, Italy; [7]Department of Medical Oncology, Beaumont Hospital, Dublin, Ireland; [8]Department of Genetics, Institute for Cancer Research, Oslo University Hospital, Norwegian Radium Hospital, Oslo, Norway; [9]Centre des Tumeurs, Jules Bordet Institute, Brussels, Belgium; [10]Institute of Cancer Research, Royal Marsden Foundation Trust, London, UK; [11]Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands; [12]European Institute of Oncology, University of Milan, Milan, Italy; [13]Department of Translational Research, Jules Bordet Institute, Brussels, Belgium; [14]Breakthrough Breast Cancer Research, Institute of Cancer Research, London, UK

Received 16 July 2012; accepted 8 October 2012

The 2012 IMPAKT task force investigated the medical usefulness of current methods for the classification of breast cancer into the 'intrinsic' molecular subtypes (luminal A, luminal B, basal-like and HER2). A panel of breast cancer and/or gene expression profiling experts evaluated the analytical validity, clinical validity and clinical utility of two approaches for molecular subtyping of breast cancer: the prediction analysis of microarray (PAM)50 assay and an immunohistochemical (IHC) surrogate panel including oestrogen receptor (ER), HER2 and Ki67. The panel found the currently available evidence on the analytical validity and clinical utility of Ki67 based on a 14% cut-off and PAM50 to be inadequate. The majority of the working group members found the available evidence on the analytical validity, clinical validity and clinical utility of ER/HER2 to be convincing. The panel concluded that breast cancer classification into molecular subtypes based on the IHC assessment of ER, HER2 and Ki67 with a 14% cut-off and on the PAM50 test does not provide sufficiently robust information to modify systemic treatment decisions, and recommended the use IHC for ER and HER2 for the identification of clinically relevant subtypes of breast cancers. Methods for breast cancer classification into molecular subtypes should, however, be incorporated into clinical trial design.

Key words: breast cancer, Ki-67, molecular classification, molecular subclasses, PAM50

introduction

For many decades, the classical breast cancer classification systems were solely based on the histological appearances of breast cancers. Numerous classification systems with limited agreement have been developed [1]. The 2003 World Health Organisation classification recognises 18 distinct histological types of invasive breast cancer [1, 2]; however the diagnostic criteria for the characterisation of each entity are rather subjective, and information on histological subtype has a limited impact on therapeutic decision making. In fact, the current stratification of breast cancers into clinically meaningful subgroups is based on prognostic clinicopathological parameters other that type, including histological grade, presence of lymph node metastasis and lympho-vascular invasion. Furthermore, predictive biomarkers, such as expression of oestrogen receptor (ER), progesterone receptor (PR), and the assessment of HER2 status, have proven to be clinically useful [3–6].

Over the past decade, microarray-based gene expression studies brought to the forefront of breast cancer research and clinical practice the fact that breast cancer comprises a heterogeneous group of diseases that have different distinct molecular features. Based on hierarchical clustering, Perou et al. [7] initially identified four breast cancer 'intrinsic' subtypes (basal-like, HER2-enriched, luminal and normal breast-like), which were shown to display gene expression patterns. Subsequent studies have led to the sub-stratification of luminal breast cancers into luminal A and luminal B, and shown that this classification system is of prognostic

*Correspondence to: Dr F. André, Department of Medical Oncology, INSERM Unit U981, Institut Gustave Roussy, Rue C Desmoulins, 39, 94805 Villejuif, France. Tel: +33-1-42-11-43-71; Fax: +33-1-42-11-52-74; E-mail: fandre@igr.fr; Dr S. Loi, Breast Cancer Translational Research Laboratory, J.C. Heuson, Institut Jules Bordet, Boulevard de Waterloo, 125, 1000 Brussels. Belgium. Tel: +32-2-541-34-57; Fax: +32-2-541-33-39; E-mail: sherene.loi@bordet.be

†These recommendations were presented at the IMPAKT 2012 Breast cancer conference in Brussels, Belgium.

significance [8–10]. Furthermore, additional class discovery studies with larger datasets resulted in the identification of additional molecular subtypes, including interferon-rich, claudin-low and molecular apocrine [8, 11–16]. Owing to the limitations of hierarchical clustering for the classification of individual samples, the proponents of the microarray-based breast cancer classification developed single sample predictors (SSPs), which enable the subtyping of a single tumour based on microarray gene expression profiling (GEP) [10]. The SSP has been further refined [8, 12] and a classifier using 50 genes was developed to identify the four major intrinsic subtypes, namely luminal A, luminal B, HER2-enriched and basal-like (Table 1) and named prediction analysis of microarray (PAM) 50 [12]. This classifier was subsequently converted into a quantitative real-time PCR (qRT-PCR) and can be carried out with RNA extracted from formalin-fixed paraffin-embedded (FFPE) samples, thereby making it applicable on archival material. In conjunction with the development of the qRT-PCR version of the PAM50 assay, a prognostic model named risk of relapse score (ROR-S), was devised based on the molecular subtypes [12]. An assay in development by NanoString Technologies, based on the PAM50 gene expression signature, provides a subtype classification as well as a prognostic score (referred to as the ROR-S) that predicts the probability of cancer recurrence over 10 years and is treated by the other IMPAKT task force (a companion manuscript concurrently submitted to *Annals of Oncology*). This PAM50 Breast Cancer Intrinsic Classifier™ is currently commercially available from ARUP laboratories (www.aruplab.com).

The cost, complexity and initial requirement of fresh frozen tissues for GEP have limited its use in clinical practice and led to the development of immuno-histochemical (IHC) surrogate definitions for the identification of the molecular subtypes of breast cancer (Table 1), given the similarities of the molecular subtypes as defined by GEP. Although in the original GEP-defined molecular classification, there was a great degree of agreement between HER2-enriched tumours as defined by GEP and HER2-positive tumours as defined by IHC and *in situ* hybridisation (ISH), in later versions of the classification, the concordance between the two subtypes is more limited. Up to 31%–59% of cases with HER2 positivity as defined by IHC and/or ISH are classified as an 'intrinsic' subtype other than HER2 enriched [12, 13, 17, 18]. The majority of basal-like breast cancers (~80%) have been shown to be of triple-negative phenotype (i.e. negative for ER, PR and HER2), however between 1%–3% of ER positive tumours have been shown to display a basal-like phenotype [7, 9, 10]. Luminal breast cancers are characterised by the expression of ER-associated genes [7] and can be sub-stratified in at least two groups based on the expression levels of proliferation-related genes, including MKI67, the transcript of the proliferation marker Ki67. Cheang et al. [19] established a Ki67 cut-off of 14% to distinguish luminal B from luminal A breast cancers by comparing GEP and IHC data; this cut-off was shown to have a sensitivity of 72% and a specificity of 77% to identify luminal B tumours [19].

In 2011, the subtypes as defined by GEP (i.e. luminal A, luminal B, HER2-enriched and basal-like) were included in the 2011 St Gallen International Expert Consensus (Table 1), and an approximation based on IHC surrogates was proposed [20]. The assessment of Ki67 and a 14% cut-off for the distinction between luminal A and luminal B were recommended by St Gallen, despite the limited evidence available in support of their use in clinical practice. In addition, discrepancies between GEP- and IHC-defined subtypes also have an impact on the translatability of the subtypes in clinical practice.

Table 1. Molecular subtypes of breast cancer

Intrinsic subtypes (GEP)	IHC classification (St Gallen)	Agreement IHC/GEP
Luminal A	'Luminal A' ER and/or PR positive HER2 negative Ki-67<14%	73%–100%
Luminal B	'Luminal B (HER2 negative)' ER and/or PR positive HER2 negative Ki-67 ≥14% 'Luminal B (HER2 positive)' ER and/or PR positive Any Ki-67 HER2 over-expressed or amplified	73%–100%
HER2-enriched	'HER2 positive (non-luminal)' HER2 over-expressed or amplified ER and PR absent	41%–69%
Basal-like	'Triple negative' ER and PR absent HER2 negative	80%

GEP, gene expression profiling; IHC, immuno-histochemical; ER, oestrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2.

Given the importance of the sub-classification of breast cancers into clinically useful groups and the controversies surrounding the GEP- and IHC-defined breast cancer taxonomy, we sought to evaluate the medical usefulness of (i) the 'intrinsic' molecular subtypes (determined by PAM50 assay [12] and representing the 'intrinsic' gene list initially described by Perou et al.) and (ii) the IHC definition of molecular subtypes, based on the St Gallen Consensus recommendation [20], to determine the most appropriate molecular subtype's classification for today's daily practice.

methods

molecular subtypes

The IHC definition of molecular subtypes is based on the St Gallen Consensus 2011 [20]: this approach uses IHC definition of ER and PR, the detection of HER2 IHC overexpression and/or *HER2* gene amplification as defined by ISH and Ki-67 labelling index, with a pre-defined cut-off 14% [19]. For ER and HER2 international guidelines have already been published [5, 6], while the efforts to standardise Ki-67 have not been completed as yet [21]. Therefore, for this work, we evaluated ER/HER2 and Ki-67 separately. We did not assess

Table 2. PUBMED search

Search terms immuno-histochemical classification	1) BREAST NEOPLASMS **AND** 2) Immuno-histochemical **OR** immunohistochemistry **OR** immunocytochemistry **OR** immunochemistry **AND** 3) Molecular subtypes **OR** molecular subtype intrinsic subtypes **OR** intrinsic subtype **OR** 'luminal A' **OR** 'luminal B' **OR** PAM50 **OR** molecular classes
Search terms PAM50	1) BREAST NEOPLASM **OR** BREAST NEOPLASMS **OR** BREAST CANCER **AND** 2) Molecular portraits **OR** tumor subtypes **OR** tumor subtyping **OR** molecular characterization **OR** molecular subtyping **OR** molecular subtyping **OR** tumor subclasses **OR** molecular subtypes **OR** molecular subtype **OR** intrinsic molecular subtypes **OR** intrinsic molecular subtype **OR** INTRINSIC SUBTYPE **OR** INTRINSIC SUBTYPES **OR** INTRINSIC GENE LIST **OR** INTRINSIC GENE LISTS **OR** PAM50 **OR** PAM50 intrinsic subtyping **OR** PAM50-based intrinsic subtype **AND** 3) Profiling **OR** SINGLE SAMPLE PREDICTOR **OR** SINGLE SAMPLE PREDICTORS **OR** 50-GENE **OR** 50-GENE SUBTYPE PREDICTOR **OR** subtype predictor **OR** gene predictor **OR** hierarchical clustering **OR** hierarchical cluster **OR** microarrays **OR** gene expression profiling **OR** microarray profiling **OR** transcriptome **OR** gene expression arrays **OR** gene expression microarrays **OR** transcriptome **OR** microarrays **OR** RT-PCR **OR** reverse transcription polymerase chain reaction **OR** real-time polymerase chain reaction **OR** qRT-PCR
Limits (exclusion)	1) EDITORIAL 2) NEWS 3) CASE REPORTS 4) *IN VITRO* 5) ANIMALS
Language	ENGLISH only
Period	2000 to February 2012

the IHC4 test [22], which also assesses ER, PR, HER2 and Ki67, as this test does not ascribe cancers into molecular subtypes and was considered to beyond the remit of this taskforce. PAM50 representing the 'intrinsic' gene list [12] is currently under development by NanoString Technologies with a quantitative reverse transcription polymerase chain reaction (qRT-PCR) carried out on FFPE. Table 1 summarises the two molecular classifications, which were evaluated.

literature search

On February 2012, we searched in the MEDLINE database using the terms summarised in Table 2. Initial results were reviewed and cross-referencing was carried out among the identified studies to ensure that all eligible studies were captured. The main eligibility criterion was studies on the breast cancer molecular classification applied in the clinical setting. We deemed some studies as not eligible for this evaluation, namely (i) studies without distinction between luminal A and luminal B tumours; and (ii) studies where luminal B tumours were defined exclusively ER and/ or PR positive and HER2 positive as defined by IHC/fluorescence ISH.

evaluation method and procedure

This project was undertaken by a working group composed of oncologists, pathologists, scientists and biostatistician with expertise in the field of breast cancer and/or GEP. The working group adopted a Delphi process which was coordinated by a medical oncology doctorate scholar. The Delphi process is a structured communication technique; joining an expert panel to answer to a pre-defined question. The experts answer questionnaires blinded to the responses provided by other members of the group in at least two rounds. After each round, a facilitator provides a summary of the experts' forecasts from the previous round as well as the reasons they provided for their judgments. Thus, experts are encouraged to revise their earlier answers in light of the replies of other members. Finally, the process is stopped after a consensus is reached.

To assess the quality of the studies, evaluation was carried out based on the general principles of the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative [23]. This initiative provided rigorous evidence-based criteria for evaluating genetic tests and other genomic applications for clinical and public health practice. The EGAPP initiative was not specifically designed to evaluate genomic applications in the oncologic field.

Each study (where appropriate) was evaluated for:

(i) The test's ability to accurately and reliably measure the molecular phenotype of interest, the expression of mRNA by breast cancer tumour cells, as well as assay reproducibility, robustness (e.g. resistance to small changes in pre-analytical or analytical variables), and quality control (analytical validity);
(ii) The test's ability to identify accurately and reliably or predict a relevant breast cancer survival end point 5–10 years after surgery (clinical validity);
(iii) The evidence that using a test to guide management in patients with diagnosed early-stage breast cancer will significantly improve health-related outcomes. It was assessed by investigating the balance of benefits (reduced adverse events due to low risk women avoiding chemotherapy) and harms (cancer recurrence that might have been prevented) (clinical utility).

Table 3. Parameters used to evaluate the eligible studies

Analytical validity	Clinical validity	Clinical utility
Data source • Number of samples • Source of samples	Data source • Cohort studies • Case–control studies • Case series	Data source • Meta-analysis • Randomised trials • Case–control studies • Case series
Reproducibility • Intra-laboratory validation • Inter-laboratory validation • Effect of time	Eligibility criteria	End points • Primary • Secondary
Blinded testing	Sample size and demographics	Data collection • Prospective • Retrospective
Specimen • FFPE • Frozen tissue • Fresh tissue	Point estimates of prognostic value • Sensitivity • Specificity • Hazard ratio	Treatment used
Report of test failures Report of indeterminate results	Study population Power calculation	Randomisation Independence of the test • Multivariate model • Comparison with current standards

FFPE, formalin-fixed paraffin embedded.

Different parameters were used in the evaluation of analytical validity, clinical validity and clinical utility (Table 3). To evaluate the ability of a molecular classification to predict recurrence risk accurately, we evaluated univariate and multivariate survival models that were reported for each classification. If more than one model was carried out on the same dataset to test different end points, we considered the one that was specified as the primary end point except if several end points were pre-specified. If more than one multivariate model was reported in the same paper but on different datasets, both models were considered.

The members of the working group were asked to review and grade all the available literature blinded to the score provided by other experts. Grading was based on the criteria provided in Table 3. Each member was asked to provide a qualitative evaluation of the evidence in support of the analytical validity, clinical validity and clinical utility for each molecular classification. Qualitative scores ranged from inadequate, adequate to convincing (Table 4). Next, all responses are gathered, tabulated, and shared with the task force members for discussion to reach a final consensus.

results

Following the literature search criteria employed, 13 articles were deemed eligible and evaluated for IHC classification, including 11 with distinction between luminal A/B based on Ki-67 and only 8 with a cut-off of 14% (Figure 1). Thirty-six publications were evaluated for PAM50 with only two studies using the PAM50 NanoString test (Figure 1). A complete publication list of all articles reviewed and evaluated is provided in Supplement 1.

analytical validity

Details on some of the parameters related to the assessment of analytical validity of the tests were reported in eight and two papers on the IHC and GEP molecular classification of breast cancers (Table 5). The source of tissue samples was mainly from retrospective studies, which were variable in size.

IHC classification

Eight studies provided analytical validity information for IHC classification based on Ki67 14%. Among 144 patients with luminal A ($n = 84$) or B ($n = 60$) tumours (PAM50) and ER +/HER2− (IHC), Cheang et al. [19] determined a Ki-67 cut-off of 14% to distinguish luminal A and B tumors with a sensitivity of 72% (95% CI 59% to 82%) and a specificity of 77% (95% CI 67% to 85%). When 17 borderline tumours were excluded (difference between Spearman rank correlation coefficients for luminal A and B centroids < 0.1), the same cut-off was found with a sensitivity of 77% (95% CI 64% to 87%) and a specificity of 78% (95% CI 68% to 87%). Independent validation of the 14% cut-off for the sub-stratification of ER-positive breast cancers into luminal A and luminal B has yet to be carried out.

It should be noted that assessment of IHC markers was carried out by a central laboratory in only four studies [24–27]. An independent review by two pathologists was reported in only two publications, although no agreement rate was provided [26, 28]. IHC staining was carried out on tissue microarrays slides in 50% of the studies, while 50% was applied on full-face slides. Six retrospective studies reported a sampling failure rate for IHC biomarkers of 2%–11.5% (insufficient tumour material or fixation with a fixative other than buffered formalin) [19, 24–27, 29].

PAM50

The 'intrinsic' molecular subtypes have been initially described by Perou et al. [7, 9] through hierarchical clustering. In 2003, Sorlie et al. [10] reported these distinct subtypes (mainly luminal and basal-like tumours) with distinct outcomes in three independent datasets comprising different patient populations whose GEP was determined by using different microarray platforms (extended Norway/Stanford cohort, $n = 122$; Van't Veer et al., $n = 97$ and West et al., $n = 49$). The overlapping 'intrinsic' genes of these three cohorts were used to develop a breast cancer class predictor using PAM with an agreement ranging from 79% to 89% when compared with the results of the hierarchical clustering applied for Van't Veer and West datasets, respectively. Of note, some genes used as predictors were used to define the test set groups. Several studies (with <100 tumour samples) subsequently reported on similar classifications (where luminal and basal-like tumors were the most robust and distinct subtypes), including distinct

Table 4. Criteria used to assess the quality of evidence for the analytical validity, clinical validity and clinical utility (EGAPP criteria)

	Analytical validity	Clinical validity	Clinical utility
Convincing	Studies that provide confident estimates of analytic sensitivity and specificity using intended sample types from representative populations	Well-designed and conducted studies in representative population(s) that measure the strength of association between a genotype and a specific and well-defined disease	Well-designed and conducted studies in representative population(s) that assess specified health outcomes
	Collaborative study using a large panel of well-characterized samples or well-designed peer-reviewed study (inter-laboratory comparison and validation study) that is generalisable and has an appropriate number and distribution of challenges	High-quality well-designed longitudinal cohort study or systematic review/meta-analysis of well-designed longitudinal cohort studies with homogeneity	Systematic review/meta-analysis of randomised, controlled trials showing consistency in results or at least one large randomised, controlled trial
Adequate	Two or more collaborative studies using a large panel of well-characterized samples or well-designed peer-reviewed studies (inter-laboratory comparisons and validation studies) are consistent but not generalisable or that lack the appropriate number and/or distribution of challenges	Systematic review of lower quality studies or review of well-designed longitudinal cohort or case–control studies with heterogeneity	Systematic review of randomised, controlled trials showing heterogeneity or controlled trial(s) without randomisation or systematic review of cohort or case–control studies with consistent results
Inadequate	Lower quality studies or combinations of higher quality studies that show important unexplained inconsistencies	Single well-designed cohort or case–control study or case series or non-consecutive cases	Systematic review of cohort, case–control studies or controlled trials without randomisation showing heterogeneity or single cohort or case–control study or case series

EGAPP, Evaluation of Genomic Applications in Practice and Prevention.

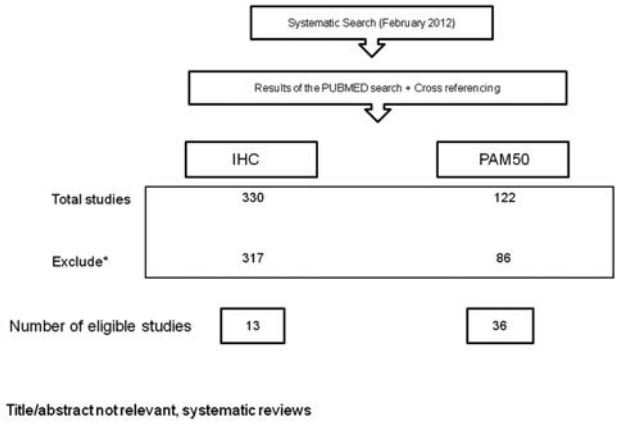

Figure 1. Flowchart of eligible studies.

ethnic populations and inflammatory breast cancers [30–35]. In a larger population-based Swedish cohort ($n = 412$), Calza et al. [36] described a concordance rate of 77.5% between the centroid prediction (using Norway/Stanford data as the training set) and the k-means clustering carried out internally within the Swedish cohort (the highest rates of discordant assignments were between luminal A and B and luminal B and HER2-enriched subtypes). In addition, Perreard et al. [37] showed that a minimised 'intrinsic' gene set could be used in a qRT-PCR assay to recapitulate the microarray classification with a concordance rate of 93%. Parker et al. [12] applied an expanded 'intrinsic' gene set of 1900 common genes in four

Table 5. Publications that provided information on analytical validity

	Ki-67 (14%)	PAM50 (qRT-PCR)
No. of papers	8	2
Reproducibility	2	0
Blinded testing	0	1
Reporting test failures/indeterminate results	6	2

qRT-PCR, quantitative reverse-transcriptase-polymerase chain reaction.

previous studies [8–10, 37] to the study of 122 breast cancers, and developed a 50-gene subtype predictor (minimised gene set using the qRT-PCR data for genes that passed FFPE performance criteria [38]).

Several publications have investigated the limitations of the use of SSPs for the classification of breast cancers into the 'intrinsic' subtype classes. PAM50 was compared with two others SSPs (Sorlie et al., 2003 and Hu et al., 2006) in a combined analysis that assessed the performance of SSPs in four datasets ($n = 832$ patients) [17]. The three SSPs were shown not to assign consistently the same patients into the molecular subtypes (fair-to-substantial agreement between every pair of SSPs in each cohort was recorded, $\kappa = 0.238$–0.740). Nevertheless, all SSPs identified molecular subtypes with similar but not equivalent survival. Molecular classification with each SSP remained an independent prognostic factor in multivariate analysis. Only the proportion

of basal-like tumours was similar between the three SSPs (agreement, $\kappa > 0.812$). Mackay et al. [39] evaluated the interobserver reproducibility between five researchers to assign a molecular subtype in three publicly available datasets ($n = 779$) using five distinct 'intrinsic' gene lists [39]. Substantial interobserver agreement was consistently observed in all datasets for four molecular subtypes (luminal, basal, HER2-enriched and normal-like) for 70.8%–76.1% of the samples (κ scores from 0.635 to 0.701). Only basal-like and HER2-enriched molecular subtypes were reproducibly identified with an almost perfect agreement ($\kappa \geq 0.81$). Haibe-Kains et al. [40] compared the three SSPs in 36 publicly available datasets ($n = 5715$) and tested their robustness using the 'prediction strength statistic'. PAM50 had low prediction strength of 0.59, 0.36 and 0.25 for 3, 4 and 5 subtypes, respectively. In addition, the three SSPs were fair to moderate concordance with each other (58%–68%; $\kappa = 0.45$–0.58). Elloumi et al. [41] highlighted that genomic classification by PAM50 could be altered by normal tissue contamination (typically 30%–50% in samples) and caused misclassification of a given tumour, mainly from more aggressive to less aggressive subtypes as the content of normal/non-neoplastic cells increased.

Limited information is available on the analytic validity of the PAM50 NanoString test. Nielsen et al. [42] reported a failure rate of 21% in a large retrospective study involving 991 patients, which was mainly due to RNA of suboptimal quality. No details of intra-or inter-laboratory comparisons were reported in this study. Sixty-seven percent of samples were obtained for qRT-PCR in a retrospective study conducted from the NCIC.CTG MA.5 randomised trial. No further explanation for the exclusion of 33% of samples was provided in this study [43].

clinical validity/utility

Tumour stage and treatment modalities were heterogeneous in the patient population included in studies investigating the IHC and GEP breast cancer molecular classification.

Table 6 summarises the multivariate survival models developed for each molecular classification. A total of 15 multivariate models were evaluated across the two molecular classifications. In 10 models (66%), the molecular classification was significantly associated ($P < 0.05$) with prognosis, but in almost no cases, an appropriate likelihood ratio test was applied. End points were disease-free survival ($n = 4$), loco-regional free survival ($n = 3$), breast cancer-specific survival ($n = 3$), overall survival (OS; $n = 2$), relapse-free survival ($n = 2$) and distant metastasis-free survival ($n = 1$). There was strong variability between the factors included in the multivariate models.

IHC classification

In seven of the eight studies, the source of information was retrospective in nature. In two of seven studies, the IHC biomarkers were carried out on retrospectively collected samples from patients that were enrolled in a prospective randomised trial [24, 27]. Cheang et al. [19] tested his 14% Ki-67cut-off in a retrospective analysis of an independent

Table 6. Evaluable multivariate models

	IHC classification	PAM50 NanoString test
Number of unique patients	13 085	1262
Number of multivariate models	13	2
Adequate documentations of multivariable regressions		
Molecular classification is significant ($P < 0.05$)	9	1[a]
Added value demonstrated using the likelihood ratio test	0	1
Adjustment factors (%)		
Histological subtype	0	0
Tumour size	92	100
Nodal status	62	100
Histological grade	62	50
ER	0	50
HER2	8	50
Ki-67	8	0
Age	85	100
Lympho-vascular invasion	77	50
Treatment	38	50

[a]In one extra model, an interaction between PAM50 and treatment was evaluated.

population of 2598 HR+ patients with different adjuvant modalities (no treatment, tamoxifen alone, tamoxifen and chemotherapy with anthracyclines or CMF). Luminal B (33%) and luminal-HER2-positive (8%) breast cancers were statistically significantly associated with poor 10-years recurrence-free and disease-specific survival in all adjuvant systemic treatment categories, and similar results were seen in lymph node-positive or lymph node-negative patients treated with tamoxifen alone. Similarly, Hugh et al. [24] analysed 1326 patients with lymph node-positive disease, constituting 91% of those who were previously treated in the BCIRG001 randomised trial and who all received adjuvant chemotherapy (TAC versus FAC), and observed a significantly longer 3-year disease-free survival for patients with luminal A breast cancers ($n = 212$) compared with those with luminal B tumours ($n = 808$). A retrospective study of 1006 non-consecutive Korean patients, including 53% luminal A and 22% luminal B tumours, reported a significantly better 5-year disease-free and OS for patients with luminal A cancers as compared with those with luminal B disease [26]. No adjustment for the type of adjuvant treatment was provided in that publication [26]. Five-year OS and disease-free survival were significantly better in patients with luminal A cancers from another study conducted from 50% ($n = 298$) of patients with high-risk breast cancer included in the phase III HE10/97 randomised trial [27].

Risks of local and regional relapse was significantly higher in patients with HER2, basal-like and luminal B tumours than in those with luminal A breast cancers after mastectomy ($n = 2085$ patients) [25]. After breast conserving surgery and adjuvant radiotherapy, risk of regional recurrence was only higher in patients with HER2 and basal-like disease on a multivariate survival analysis [25]. In 1691 consecutive women with small invasive node-negative breast tumour, higher risk of loco-regional relapse was only found for HER2 and basal-like

tumours [28]. A consistent finding in multiple studies [24, 26, 27] was a significantly poorer outcome for HER2-positive and basal-like tumours when compared with luminal A cancers, even in small breast cancers (size ≤1 cm), which are often considered as of good prognosis [28].

One study investigated the predictive value of the IHC classification in determining patients who would benefit taxanes [24]. Hugh et al. [24] reported that TAC was superior to FAC only in patients with luminal B tumours, even if combined chemo-endocrine therapy containing tamoxifen was administered. However, no interaction test between IHC classification and TAC versus FAC was provided in this study. The 3-year disease-free survival rates were 89.4% and 82%, respectively (HR = 0.71; 95% CI 0.53–0.95; $P = 0.02$, multivariate model). An exploratory analysis conducted by Cancello et al. [29] in 199 young patients (<35 years) with luminal B breast cancers reported a statistically shorter disease-free survival when received tamoxifen or LH-RH analogue alone versus the combination of the two drugs.

It should be noted that in all studies, no patients received adjuvant trastuzumab, as they preceded the approval of this humanised monoclonal antibody for the management of patients with early-stage breast cancer. Table 6 summarises the design of the different multivariate survival models included in the studies evaluated ($n = 13$).

PAM50

Two studies using the PAM50 SSP found that molecular subtypes were prognostic in two large, N0, untreated cohorts of patients ($n = 761$ [12] and $n = 1260$ [40]). These data were confirmed in a multivariate survival model adjusted by tumour size, node status, grade and ER status in one of the studies [12]. The same study reported a predictive value of the 'intrinsic' subtypes in 133 patients treated with neoadjuvant chemotherapy in a multivariate model (with or without histological grade) [12].

In 120 patients treated with neoadjuvant chemotherapy, the authors described higher pCR rates for basal-like and HER2-enriched tumours but molecular subtypes added no predictive value to ER and HER2 status in multivariate model [44].

PAM50 NanoString test was evaluated in two studies, including >1200 patients [42, 43]. The first study included both lymph node-positive (65%) and negative ER+ breast cancers from patients uniformly treated with only adjuvant tamoxifen [42]. It should be noted that his cohort was the same as the one partly described by Cheang et al. [19]. Luminal A breast cancers assigned by the PAM50 assay had significantly better 5 and 10-years disease-specific survival than luminal B, HER2-enriched or basal-like breast cancers. In multivariate model, the 'intrinsic' subtypes remained independent prognostic factor, as did tumour size and node status, particularly during the first 5 years. A model of IHC subtype (incorporating data on Ki-67 and HER2) and tumour size (IHC-T) was constructed following the same process used for development of ROR scores [42]. This model failed to show an added prognostic value using c-index to Adjuvant! Online (AOL) in node-negative or node-positive disease [42].

The second study included retrospectively samples of 476 of 716 pre-menopausal patients with node-positive breast cancer who were randomised between anthracycline (CEF) versus non-anthracycline (CMF) adjuvant chemotherapy in the NCIC.CTG MA.5 trial [43].'Intrinsic' subtypes were significantly associated with 5-years relapse-free and OS in the entire cohort uniformly treated with chemotherapy. Multivariate analyses were used with treatments, intrinsic subtypes (PAM50) and their interaction as covariates, and adjusted for age, number of positive lymph nodes, ER level, type of surgery and tumour size. A benefit in relapse-free survival and OS with anthracycline regimen was seen only for HER2-enriched tumours ($P = 0.03$, interaction test), but this differential benefit did not appear different from the one using the FISH/IHC-based HER2 status. A non-significant statistical trend was observed in patients treated with an anthracycline-based regimen, where patients with luminal B cancers appeared to have an improved survival, whereas in patients with basal-like and luminal A disease, anthracycline-based therapy was associated with a numerical short survival.

Table 6 summarises the different multivariate survival models included in the studies evaluated ($n = 2$).

Limitations of this study

It should be noted that the analyses of analytical validity, clinical validity and clinical utility provided in this study have limitations that are applicable to exercises of this nature. First, new data may have emerged between the literature review and the publication of the results, which could have had an impact on the interpretation of the levels of analytical and clinical validity and clinical utility of the tests investigated. Second, the analyses of analytical validity, clinical validity and clinical utility were carried out using the EGAPP criteria; it is possible that if other sets of rules had been employed, slightly different conclusions would have been rendered. Third, although the task force was composed of breast cancer and/ or GEP experts, we cannot rule out the possibility that a different panel of experts with different expertise would have come to slightly different conclusions.

IMPAKT 2012 Working Group Statement

analytical validity

According to the EGAPP criteria, the majority of the working group members found the available evidence on the analytical validity of ER/HER2 to be convincing (Figure 2A). The panel found the available evidence on the analytical validity of Ki-67 and PAM50 to be inadequate and the panel acknowledges that further data are required.

clinical validity

According to the EGAPP criteria, the majority of the working group members found the available evidence on the clinical validity of ER/HER2 to be convincing and on the clinical validity of Ki-67 and PAM50 to be adequate (Figure 2B). The panel found that currently molecular subtypes should be defined based on pathological assessment of ER and HER2 (IHC ± ISH) which have already largely demonstrated their

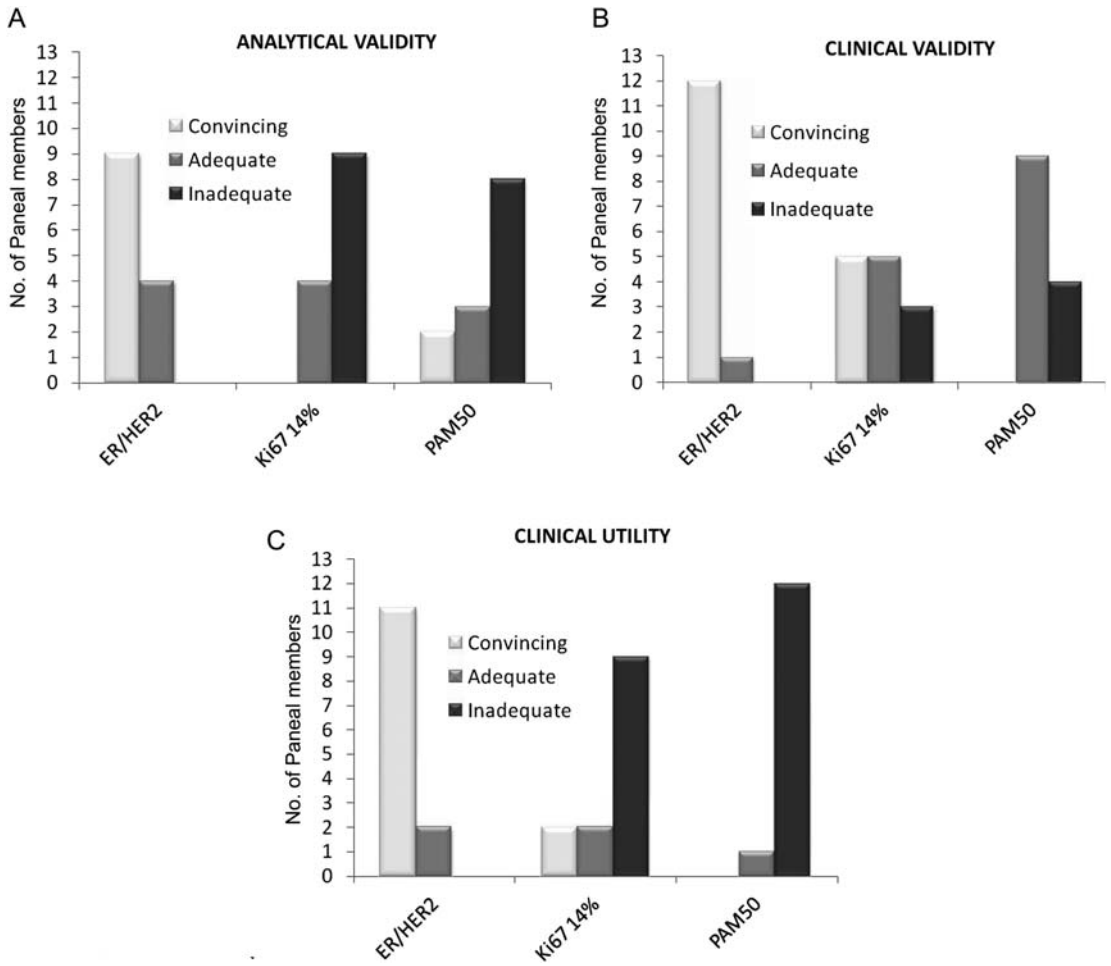

Figure 2. Panel assessment of the evidence on analytical validity, clinical validity and clinical utility of PAM50 and an immuno-histochemical surrogate panel for the identification of the molecular subtypes of breast cancer, according to the EGAPP criteria. (A) Quality of evidence for analytical validity (answers of the panel); (B) quality of evidence for clinical validity (answers of the panel) and (C) quality of evidence for clinical utility (answers of the panel).

prognostic and predictive values (i.e. HER2 as defined by HER2 IHC overexpression and/ or HER2 gene amplification; luminal as defined by HER2 negativity and ER and/or PR expression; basal-like as defined by triple-negative IHC phenotype). Neither Ki-67 nor PAM50 reached this level of evidence, however both approaches were shown to lead to the identification of subsets of ER-positive breast cancers that have distinct outcomes, and may consist in promising ways to distinguish luminal A from luminal B tumours. It was a consensus in the panel that further data on the clinical validity of the tests are still required. These include the development of standardised test (particularly for Ki-67) [21] and the evaluation of both Ki67 and PAM50 in a larger number of patients from randomised prospective trials.

clinical utility

The majority of the working group members found the available evidence on the clinical utility of ER/HER2 to be convincing. The panel found the available evidence on the clinical utility of Ki67 and PAM50 to be inadequate (Figure 2C). For current clinical practice, the panel recommends to use IHC for ER and HER2 for the identification of clinically relevant subgroups of breast cancers, however no convincing data for the use of Ki67 with a cut-off of 14% for the subdivision of luminal tumours into luminal A and luminal B. The panel does not support the use of PAM50 for current clinical decisions in regards to systemic therapy, especially in cases of discordance with IHC. For example, a HER2-enriched tumour that is not HER2 positive by IHC or FISH should be managed as HER2 negative, as there is no evidence that these patients would benefit from anti-HER2 therapies. In a patient with an ER-positive/HER2-negative with high-risk clinical features breast cancer that is classified as luminal A by PAM50, chemotherapy should not be ruled out for consideration of systemic chemotherapy in the adjuvant setting.

Although patients with luminal B breast cancers are often described as having a poorer outcome than those with luminal A tumours, the panel does not support the notion that luminal A and B should be used to inform clinical decision in regards to the use of adjuvant chemotherapy due to the absence of convincing data on clinical utility both in lymph node-positive or -negative disease (for Ki67 and PAM50). To date, the group believes that decisions on assigning chemotherapy should be made using the clinically available tools until more robust data

on the value of molecular subtyping of breast cancers are available.

All the members of the working group agreed on the need to incorporate the molecular subtypes based on PAM50 or other molecular models in the design of clinical trials. Ideally tests that could be carried out using FFPE samples would be preferable in the context of clinical trials. PAM50 and IHC data could be accrued concurrently to build more comprehensive databases of cases with both types of data. To provide information on clinical utility, Ki67 expression should be subjected to a central review.

conclusions and future directions

Despite the progress made over the past 12 years on the understanding of the molecular heterogeneity of breast cancers, the classification into molecular subtypes based on the IHC assessment of ER, PR, HER2 and Ki67 with a single 14% cut-off or PAM50 NanoString GEP test does not provide sufficiently robust information to modify treatment decisions based on their results. In addition, discrepant results between molecular subtypes defined by PAM50 on one hand and by IHC on the other hand, are frequently reported. In these cases, there is no evidence to support the definition of one subtype classification as having a higher predictive ability. Currently, clinical variables including tumour size, nodal status, histological grade, ER, PR and HER2 status remain the current 'gold standard' for systemic therapy decision making. Further studies addressing the clinical utility of the IHC and PAM50 classification and investigating the optimal therapy for patients with discrepant results are warranted.

The breast cancer taxonomy including the five 'intrinsic' molecular subtypes (luminal A, luminal B, basal-like, HER2-enriched and normal-like) is a working model, and that additional molecular subtypes with distinct repertoires of molecular aberrations and clinical behaviour will be identified. In fact, recent studies have led to the identification of claudin-low [11] and molecular apocrine subtypes [14], whose clinical and biological significance remain to be fully elucidated. Furthermore, unsupervised analysis of triple-negative breast cancers led to the identification of six distinct subtypes displaying unique gene expression and ontologies, including two basal-like, an immunomodulatory, a mesenchymal, a mesenchymal stem-like and a luminal androgen receptor subtypes [45]. Taken together, these observations demonstrate that the final molecular taxonomy of breast cancer is likely to be more complex that initially envisaged [7]. With the development of high-throughput technologies that allow for the complete characterisation of the repertoire of somatic genetic aberrations and the complete transcriptomic features of human cancers, it is likely that the molecular taxonomy of breast cancers will evolve further.

acknowledgements

The authors thank ESMO and the Breast International Group (BIG) for their help and support in coordinating the working group activities. The authors are also grateful to Torsten Nielsen for his input in the activities of the task force.

disclosures

- Angelo Di Leo has received Honoraria from Genomic Health.
- Christos Sotiriou is a named inventor of patent of the genomic grade index (GGI).
- Giuseppe Viale has received Honoraria from consulting from DAKO.
- Carsten Denkert is a shareholder and has received honoraria and research funding from Sividon.
- All remaining authors have declared no conflicts of interest.

references

1. Weigelt B, Reis-Filho JS. Histological and molecular types of breast cancer: is there a unifying taxonomy? Nat Rev Clin Oncol 2009; 6: 718–730.
2. Ellis P. WHO Classification of Tumours. Pathology and Genetics of Tumours of the Breast and Female Genital OrgansTumours, 2003.
3. Davies C, Godwin J, Gray R et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 2011; 378: 771–784.
4. Yin W, Jiang Y, Shen Z et al. Trastuzumab in the adjuvant treatment of HER2-positive early breast cancer patients: a meta-analysis of published randomized controlled trials. PLoS One 2011; 6: e21030.
5. Hammond ME, Hayes DF, Dowsett M et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Arch Pathol Lab Med 2010; 134: 907–922.
6. Wolff AC, Hammond ME, Schwartz JN et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 2007; 25: 118–145.
7. Perou CM, Sorlie T, Eisen MB et al. Molecular portraits of human breast tumours. Nature 2000; 406: 747–752.
8. Hu Z, Fan C, Oh DS et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 2006; 7: 96.
9. Sorlie T, Perou CM, Tibshirani R et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001; 98: 10869–10874.
10. Sorlie T, Tibshirani R, Parker J et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2003; 100: 8418–8423.
11. Herschkowitz JI, Simin K, Weigman VJ et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol 2007; 8: R76.
12. Parker JS, Mullins M, Cheang MC et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009; 27: 1160–1167.
13. Prat A, Parker JS, Karginova O et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 2010; 12: R68.
14. Farmer P, Bonnefoi H, Becette V et al. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene 2005; 24: 4660–4671.
15. Doane AS, Danso M, Lal P et al. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene 2006; 25: 3994–4008.
16. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet 2011; 378: 1812–1823.
17. Weigelt B, Mackay A, A'Hern R et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol 2010; 11: 339–349.
18. de Ronde JJ, Hannemann J, Halfwerk H et al. Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response. Breast Cancer Res Treat 2010; 119: 119–126.

19. Cheang MC, Chia SK, Voduc D et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst 2009; 101: 736–750.
20. Goldhirsch A, Wood WC, Coates AS et al. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol 2011; 22: 1736–1747.
21. Dowsett M, Nielsen TO, A'Hern R et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. J Natl Cancer Inst 2011; 103: 1656–1664.
22. Cuzick J, Dowsett M, Pineda S et al. Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. J Clin Oncol 2011; 29: 4273–4278.
23. Teutsch SM, Bradley LA, Palomaki GE et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. Genet Med 2009; 11: 3–14.
24. Hugh J, Hanson J, Cheang MC et al. Breast cancer subtypes and response to docetaxel in node-positive breast cancer: use of an immunohistochemical definition in the BCIRG 001 trial. J Clin Oncol 2009; 27: 1168–1176.
25. Voduc KD, Cheang MC, Tyldesley S et al. Breast cancer subtypes and the risk of local and regional relapse. J Clin Oncol 2010; 28: 1684–1691.
26. Park S, Koo JS, Kim MS et al. Characteristics and outcomes according to molecular subtypes of breast cancer as classified by a panel of four biomarkers using immunohistochemistry. Breast 2011; 21: 50–57.
27. Skarlos P, Christodoulou C, Kalogeras KT et al. Triple-negative phenotype is of adverse prognostic value in patients treated with dose-dense sequential adjuvant chemotherapy: a translational research analysis in the context of a Hellenic Cooperative Oncology Group (HeCOG) randomized phase III trial. Cancer Chemother Pharmacol 2012; 69: 533–546.
28. Cancello G, Maisonneuve P, Rotmensz N et al. Prognosis in women with small (T1mic,T1a,T1b) node-negative operable breast cancer by immunohistochemically selected subtypes. Breast Cancer Res Treat 2011; 127: 713–720.
29. Cancello G, Maisonneuve P, Rotmensz N et al. Prognosis and adjuvant treatment effects in selected breast cancer subtypes of very young women (<35 years) with operable breast cancer. Ann Oncol 2010; 21: 1974–1981.
30. Van Laere SJ, Van den Eynden GG, Van der Auwera I et al. Identification of cell-of-origin breast tumor subtypes in inflammatory breast cancer by gene expression profiling. Breast Cancer Res Treat 2006; 95: 243–255.
31. Han W, Nicolau M, Noh DY, Jeffrey SS. Characterization of molecular subtypes of Korean breast cancer: an ethnically and clinically distinct population. Int J Oncol 2010; 37: 51–59.
32. Sotiriou C, Neo SY, McShane LM et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci USA 2003; 100: 10393–10398.
33. Sorlie T, Wang Y, Xiao C et al. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. BMC Genomics 2006; 7: 127.
34. Yu K, Lee CH, Tan PH, Tan P. Conservation of breast cancer molecular subtypes and transcriptional patterns of tumor progression across distinct ethnic populations. Clin Cancer Res 2004; 10: 5508–5517.
35. Bertucci F, Finetti P, Rougemont J et al. Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. Cancer Res 2005; 65: 2170–2178.
36. Calza S, Hall P, Auer G et al. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. Breast Cancer Res 2006; 8: R34.
37. Perreard L, Fan C, Quackenbush JF et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res 2006; 8: R23.
38. Mullins M, Perreard L, Quackenbush JF et al. Agreement in breast cancer classification between microarray and quantitative reverse transcription PCR from fresh-frozen and formalin-fixed, paraffin-embedded tissues. Clin Chem 2007; 53: 1273–1279.
39. Mackay A, Weigelt B, Grigoriadis A et al. Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. J Natl Cancer Inst 2011; 103: 662–673.
40. Haibe-Kains B, Desmedt C, Loi S et al. A three-gene model to robustly identify breast cancer molecular subtypes. J Natl Cancer Inst 2011; 104: 311–325.
41. Elloumi F, Hu Z, Li Y et al. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med Genomics 2011; 4: 54.
42. Nielsen TO, Parker JS, Leung S et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin Cancer Res 2010; 16: 5222–5232.
43. Cheang MC, Voduc KD, Tu D et al. Responsiveness of intrinsic subtypes to adjuvant anthracycline substitution in the NCIC.CTG MA.5 randomized trial. Clin Cancer Res 2012; 18: 2402–2412.
44. Esserman LJ, Berry DA, Cheang MC et al. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). Breast Cancer Res Treat 2011; 132: 1049–1062.
45. Lehmann BD, Bauer JA, Chen X et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 2011; 121: 2750–2767.