Discovery and Validation of Breast Cancer Early Detection Biomarkers in Preclinical Samples

Christopher I. Li

Abstract Despite the widespread use of mammography for breast cancer screening, breast cancer remains the most common cause of cancer-related mortality among women worldwide. The identification of blood-based biomarkers useful for the early detection of breast cancer could have a major impact on reducing breast cancer disease burden by identifying cancers early when they are most treatable. We conducted a series of large-scale proteomic discovery and validation studies using preclinical samples from the Women's Health Initiative Observational Study prospective cohort. Of the 503 proteins quantified in experiments conducted on samples from ER$^+$ breast cancer patients and matched controls, 57 differentiated cases from controls. The seven candidates were assessed in an independent validation set with a commercially available ELISA assay. We confirmed that one of these candidates, epidermal growth factor receptor (EGFR), was elevated in cases versus controls. Compared to women in the lowest EGFR quartile, those in the highest quartile has a 2.90-fold ($p=0.0005$) increased risk of developing breast cancer. An interaction with use of menopausal hormone therapy was observed such that among current estrogen plus progestin users, those in the highest EGFR quartile had a 9.04-fold ($p=0.0004$) increased risk of developing breast cancer. While the performance of EGFR in terms of sensitivity and specificity is insufficient for it to be used on its own clinically, the formal validation of EGFR suggests that there may indeed be changes in the plasma proteome prior to the clinical diagnosis of breast cancer that are detectable and of potential clinical utility.

C. I. Li ($\boxtimes$)
Fred Hutchinson Cancer Research Center,
1100 Fairview Avenue North, M4-C308,
Seattle, WA 98109-1024, USA
e-mail: cili@fhcrc.org

Introduction

There is an urgent need for breast cancer early detection biomarkers given that none are currently available and given the considerable public health importance of breast cancer. At present, the best available tool for the early detection of breast cancer is mammography. Randomized trials have shown that annual or biennial mammography reduces mortality 30% among women 50–69 years of age [1]. However, mammography is limited by its 75–90% sensitivity and 90–95% specificity, with optimal performance only among 50- to 69-year-olds. For example, its positive predictive value (proportion of those with a positive test who have the disease) is 60–80% among 50- to 69 year-olds, but is only 20% among those <50 years of age. Despite improvements in technology and the widespread use of mammography, breast cancer still remains the second leading cause of cancer mortality in women in the USA [2] and the leading cause of cancer mortality in women worldwide [3]. There are several reasons for this, including the sensitivity and specificity of mammography, lack of uniform access to screening services, and the variable utility of mammography to detect different types of tumors. It is well established that mammography is better able to detect certain types of breast cancer (such as ductal carcinomas) than other types (such as poor prognosis estrogen receptor (ER)-negative tumors) [4–9]. Considering ER status, interval-detected tumors are 1.8- to 2.6-fold more likely to be ER$^-$ compared to screen-detected tumors [7, 9].

Continued improvements in our ability to detect breast cancer early offer the promise of further reducing the

burden of this disease as breast cancer detected at an earlier stage is much more curable than is metastatic disease. The current 5-year survival rate for localized breast cancer in the USA is 98%, but is only 27% for metastatic disease [2]. Thus, there is a great public health need for improvements in our abilities to detect breast cancers of all types earlier, and proteomics is a means for identifying early detection biomarkers in plasma.

We recently completed a unique large-scale study aimed at discovering breast cancer early detection biomarkers. Unlike the vast majority of biomarker discovery studies, this case–control study nested in the Women's Health Initiative Observational Study [10, 11] used preclinical samples where the plasma specimens assessed were all obtained 0–17 months prior to diagnosis among breast cancer cases. The proteomic discovery platform used involved applying mass spectrometry to extensively fractionated case and control plasma pools using an approach that has been previously described [12]. Though low-throughput, a major advantage of this approach is that it allows the plasma proteome to be interrogated in considerable depth, which is important given the enormous dynamic range of plasma protein concentrations. In addition, protein concentrations can be quantitated and proteins can be identified by searching mass spectra against the human International Protein Index database (version 3.13). Given the heterogeneity of breast cancer and the likelihood that breast cancer early detection biomarkers may be specific to certain breast cancer types, we evaluated markers specific to breast cancer overall and to several breast cancer subtypes, including: estrogen receptor-positive (ER$^+$), triple negative (TN, ER$^-$/PR$^-$/HER2$^-$), and HER2-overexpressing (H2E, ER$^-$/HER2$^+$) breast cancer.

We have completed and published an initial set of validation experiments on a small subset of promising candidates specific to ER$^+$ disease that have an available ELISA assay [13]. Thus far, we have only attempted to validate markers specific for ER$^+$ disease because of the rarity of TN and H2E breast cancers and the lack of a sufficiently sized validation set for either tumor subtype. Very few of our ER$^+$ candidates had an available ELISA, including none of our top-ranked candidates. Nevertheless, we proceeded with the validation of the seven candidates specific for ER$^+$ breast cancer with an available ELISA assay since a completely independent validation set of ER$^+$ cases and matched controls was available to us. The results of this study have been published and are summarized below [13].

Results

The workflow of this study is summarized in Fig. 1 and included a discovery phase on pooled samples and two
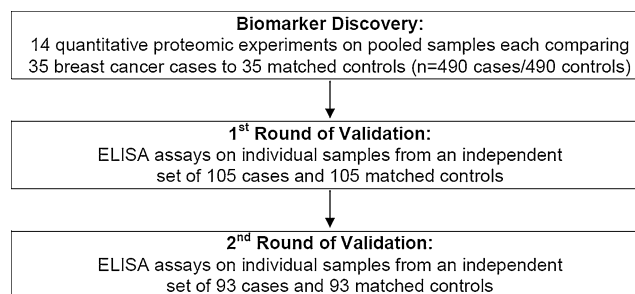

Fig. 1 Study design

rounds of validation on independent sets of individual samples. The discovery phase involved 14 quantitative proteomic experiments. In each experiment, a pooled set of plasma from 35 breast cancer cases was compared to plasma pooled from 35 matched cancer-free controls. Case pools were made uniform with respect to ER status, PR status, and histology. Twelve of these experiments were limited to ER$^+$ breast cancer and two were limited to ER$^-$ disease. Cases and controls were ascertained from the WHI OS, and the characteristics of cases and controls in our discovery and independent validation sets are shown in Table 1. Controls were matched to cases on age, race/ethnicity, study blood draw date (± 1 year), and clinical center of enrollment. Briefly, cases were somewhat more likely to be current users of combined estrogen plus progestin (E+P) menopausal hormone therapy and to be overweight/obese (body mass index ≥ 25.0 kg/m^2) and less likely to have had a hysterectomy compared to controls

We quantified a total of 503 proteins in our discovery experiments on ER$^+$ breast cancer and 57 met predefined statistical criteria of having a fold change ≥1.15, a p value <0.10, and were quantified in at least 2 of our 12 experiments. More stringent statistical criteria were not applied to our discovery data in order to make our candidate lists more inclusive. This is because we planned subsequent rigorous validation experiments to identify false positives, but we did not want to discard potential true positive candidates that more stringent criteria would have excluded. Of these 57 candidates, seven had a commercially available ELISA assay and thus could be validated in a straightforward manner. These seven candidates were: epidermal growth factor receptor (EGFR), fibronectin 1 (FN1), insulin growth factor binding protein 1 (IGFBP1), lactotransferrin (LTF), protein NOV homolog (NOV), trefoil factor 3 (TFF3), and von Willebrand factor (VWF). So while discovery experiments were conducted on pooled samples, validation ELISA assays were performed on individual specimens.

Results from our discovery and first round of validation experiments on 105 cases and 105 controls not used in our discovery experiments are shown in Table 2. EGFR levels

Table 1 Characteristics of breast cancer cases and controls used for biomarker discovery and validation

Characteristic	Discovery				Validation			
	Cases (n=490)		Controls (n=490)		Cases (n=198)		Controls (n=198)	
	n	%	n	%	n	%	n	%
Age at blood draw (years)								
50–59	127	25.9	127	25.9	41	20.7	41	20.7
60–69	237	48.4	237	48.4	91	46.0	91	46.0
70–79	126	25.7	126	25.7	66	33.3	66	33.3
Race/ethnicity								
Non-Hispanic white	435	88.8	435	88.8	179	90.4	179	90.4
African American	22	4.5	22	4.5	8	4.0	8	4.0
Hispanic white	14	2.9	14	2.9	3	1.5	3	1.5
Asian/Pacific Islander	14	2.9	14	2.9	6	3.0	6	3.0
Other	5	1.0	5	1.0	2	1.0	2	1.0
Use of menopausal hormone therapy at blood draw								
Never user	141	28.8	173	35.3	51	25.8	69	34.8
Former user	60	12.2	78	15.9	23	11.6	20	10.1
Current unopposed estrogen user	131	26.7	136	27.8	49	24.7	53	26.8
Current estrogen and progestin user	158	32.2	103	21.0	75	37.9	56	28.3
First degree family history of breast cancer								
No	378	77.5	393	80.2	157	79.3	151	76.6
Yes	110	22.5	97	19.8	41	20.7	46	23.4
Missing	2		0		0		1	
Body mass index, kg/m2								
<25.0	199	41.2	224	46.3	80	40.4	92	46.7
25.0–29.9	166	34.4	156	32.2	74	37.4	62	31.5
>30.0	118	24.4	104	21.5	44	22.2	43	21.8
Missing	7		6	1.2	0		1	
Hysterectomy								
No	309	63.1	291	59.5	132	67.0	121	61.1
Yes	181	36.9	198	40.5	65	33.0	77	38.9
Missing	0		1		1			

This table was previously published in Pitteri et al. [13]

differentiated cases from controls [odds ratio (OR)=1.68, $p=0.0017$], but none of the other six proteins differed between cases and controls. However, for the most part, the magnitude and direction of the risk estimates were similar in the discovery and validation experiments for each marker. To further confirm this finding, we conducted a second round of validation on another independent set of 93 cases and 93 controls from the WHI OS and again confirmed that EGFR was elevated in cases compared to controls (Table 3). Combining the two validation sets, the OR for EGFR was 1.44, with a highly statistically significant p value of 0.0008.

It has previously been shown that the use of menopausal hormone therapy impacts a significant portion of the serum proteome [12]. When we stratified our validation results

according to the use of menopausal hormone therapy (never/former users, current unopposed estrogen users, and current E+P users), there was evidence that EGFR only differentiated cases from controls among women who were current E+P users (Table 3). Specifically, the OR among current E+P users was 2.41 ($p=0.0001$), but among never/former users of hormone therapy and current users, unopposed estrogen EGFR was not associated with risk (OR=1.05, $p=0.78$; OR=1.40, $p=0.13$, respectively). The p value comparing the odds ratios among current E+P users to never/former users was 0.0019, and the p value comparing current E+P users to current E users was 0.12.

We also assessed risk according to EGFR quartile. Across all cases and controls, women in the highest EGFR quartile had a 2.90-fold ($p=0.005$) increased risk of

Table 2 Results from the first round of ELISA-based validations on 105 case/control pairs

Protein name	Abbreviation	Discovery results			Validation results		
		Fold change[a]	95% CI	p value	Odds ratio[b]	95% CI	p value
Epidermal growth factor receptor	EGFR	1.17	0.98–1.40	0.070	1.68	1.20–2.24	0.002
Fibronectin 1	FN1	0.77	0.64–0.93	0.020	1.26	0.95–1.61	0.11
Protein NOV homolog	NOV	1.54	1.11–2.14	0.029	1.20	0.89–1.64	0.22
Insulin-like growth factor binding protein 1	IGFBP1	0.85	0.77–0.93	0.003	1.15	0.86–1.51	0.35
von Willebrand factor	VWF	0.68	0.46–1.00	0.048	1.11	0.83–1.49	0.49
Trefoil factor 3	TFF3	1.28	1.08–1.51	0.010	1.11	0.83–1.50	0.47
Lactoferrin	LTF	1.80	1.19–2.73	0.025	0.98	0.75–1.28	0.88

This table was previously published in Pitteri et al. [13]

[a] Fold change refers to geometric mean of case-to-control ratio across samples in the subset

[b] Odds ratios were computed using logistic regression and were adjusted for age, time between blood draw and breast cancer diagnosis, race/ethnicity, body mass index, and first-degree family history of breast cancer

developing breast cancer compared to those in the lowest quartile. This risk was substantially higher among current E+P users, where those in the highest EGFR quartile had a 9.04-fold ($p=0.0004$) increased risk (Table 4). The receiver operator curve for EGFR among current E+P users has an area under the curve of 0.7. At 80% specificity, EGFR's sensitivity as a single marker is 56%, and at 90% specificity, its sensitivity is 31% (Fig. 2).

Discussion

We convincingly demonstrated in two separate validation sets completely independent from our discovery set that EGFR levels are elevated in preclinical plasma of E+P users who went on to be diagnosed with breast cancer compared to controls. In this setting, at 80% specificity, EGFR's sensitivity was 56% and at 90% specificity the sensitivity was 31%. In comparison, PSA, which is

clinically used to screen men for prostate cancer, has 40.5% sensitivity at 81.1% specificity and 20.5% sensitivity at 93.8% [14]. So while comparable in performance to PSA, EGFR cannot be viewed as a clinically useful breast cancer early detection biomarker on its own given that it only appears to be predictive among E+P users. It is important to note though that while this finding was statistically significant, due to the relatively small numbers of cases and controls who were E+P users, the 95% confidence intervals for our risk estimates were somewhat wide.

Nevertheless, our finding is still important in two respects. First, no prior studies have validated even a single breast cancer early detection biomarker in preclinical specimens to the degree we have here, validating EGFR in two completely independent validation sets. This suggests that detectable changes in the plasma proteome may indeed be present preclinically for diseases such as ER$^+$ breast cancer, which are relatively small tumors compared to other cancers, such as ovary cancer, where

Table 3 First and second round validation results for EGFR

ELISA round	Overall			Menopausal hormone therapy use								
				Never/former users			Current E users			Current E+P users		
	OR[a]	95% CI	p value	OR[a]	95% CI	p value	OR[a]	95% CI	p value	OR[a]	95% CI	p value
EGFR												
1st round ELISA	1.68	1.20–2.24	0.002	1.09	0.71–1.66	0.71	1.73	0.86–3.42	0.12	4.27	1.71–12.05	0.002
2nd round ELISA	1.30	0.96–1.81	0.09	0.96	0.58–1.61	0.89	1.15	0.57–2.41	0.67	2.03	1.13–4.03	0.02
1st and 2nd rounds combined	1.44	1.16–1.79	0.0008	1.05	0.76–1.43	0.78	1.40	0.90–2.27	0.13	2.41	1.56–3.99	0.0001

This table was previously published in Pitteri et al. [13]

[a] ORs were computed using logistic regression and were adjusted for age, time between blood draw and breast cancer diagnosis, race/ethnicity, body mass index, and first-degree family history of breast cancer

Table 4 Quartile distributions of EGFR validation results among all case/control sets and among current estrogen and progestin users

EGFR quartile[a]	Controls		Cases		Odds ratio[b]	95% CI	p value
	n	%	n	%			
All cases and controls							
(−2.5, −0.63)	50	25.3	29	14.6	1.00	Reference	
(−0.63, 0.02)	49	24.7	50	25.3	1.79	0.98–3.32	0.06
(0.02, 0.66)	49	24.7	41	20.7	1.47	0.78–2.76	0.23
(0.66, 2.48)	50	25.3	78	39.4	2.90	1.60–5.32	0.0005
Current estrogen and progestin users							
(−1.87, −0.63)	14	25.0	6	8.0	1.00	Reference	
(−0.63, −0.08)	14	25.0	12	16.0	2.37	0.68–8.93	0.19
(−0.08, 0.46)	14	25.0	15	20.0	2.95	0.85–11.25	0.10
(0.46, 1.85)	14	25.0	42	56.0	9.04	2.78–33.21	0.0004

This table was previously published in Pitteri SJ, et al. *Cancer Research* 2010 (13)

[a] Units are standardized to give controls a mean of 0 and a standard deviation of 1. Quartiles computed from the control distribution

[b] Odds ratios were computed using logistic regression and were adjusted for age, time, and first-degree family history of breast cancer

biomarker discovery and validation has thus far been much more successful. Second, consideration of other exposures in biomarker discovery studies is likely critical given that while overall EGFR was not found to distinguish cases from controls among E+P users it is highly statistically significant. So consideration of factors like use of hormone therapy, which has been shown to have a major impact on the plasma proteome [12], is critical in this type of work. Future work aimed at discovering and validating breast cancer early detection biomarkers is needed, and our EGFR

Fig. 2 EGFR receiver operator curve for cases versus control among all current E+P users. This figure was previously published in Pitteri et al. [13]

finding warrants further replication. The primary limitation of this study was the lack of readily available means to validate the numerous other candidates we discovered, most of which were much more compelling candidates based on their discovery odds ratios and *p* values. It is not surprising that the other six markers did not validate given their comparatively high FDRs (ranging from 0.66 to 1.00) compared to EGFR and to several of our other much higher ranked candidates.

With respect to the biology of EGFR, EGFR is a cell surface tyrosine kinase receptor and is a member of the *ERBB* protoncogene family which also includes HER2. EGFR is a transmembrane protein, and the peptides we identified by mass spectrometry in our discovery IPAS experiments were all located on the extracellular region, suggesting shedding of the extracellular domain by cells. Binding of EGFR by various ligands can result in increased uncontrolled proliferation of cancer cells, and EGFR is overexpressed in 20–81% of breast tumors [15–17]. Several studies have measured blood levels of EGFR in relation to breast cancer, though overall, the results are quite inconsistent. It is difficult to directly compare the results of these studies to ours since none involved measurements of EGFR in the preclinical period prior to a breast cancer patient's diagnosis. EGFR levels have been reported to be higher in normal individuals than patients with primary breast cancer [18] and metastatic breast cancer [19, 20], with another finding no differences in EGFR levels between patients with metastatic breast cancer patients and healthy women [21]. Again, none of these reports used preclinical specimens from breast cancer patients, so the differences observed could be the result of factors related to the
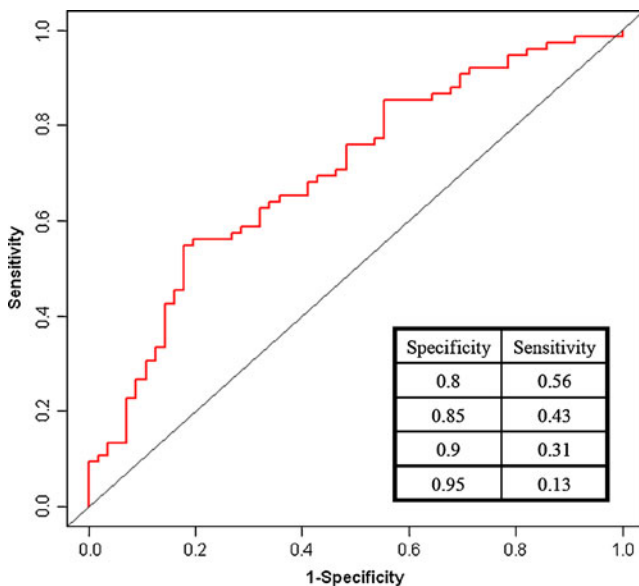
diagnosis and treatment of breast cancer. With respect to survival, some studies suggest that lower levels of EGFR in patients with metastatic breast cancer are associated with shorter overall survival [20], particularly in patients with ER$^+$ tumors [19], whereas others have found no association between EGFR serum levels and overall [22, 23] or progression-free survival [22]. Among women with hormone receptor-positive disease, serum EGFR levels decreased significantly after 1 and 3 months of letrozole therapy versus pretreatment conditions [21]. Given the inconsistency across studies and several critical differences in their respective designs, the role of serum/plasma levels of EGFR with respect to breast cancer is quite unclear. While EGFR is involved in hormonal pathways relevant to breast cancer, there is no clear explanation at this point for why EGFR may be useful for the early detection of breast cancer only among E+P users, but not among either E users or never/former users. Thus, further investigations of EGFR as a potential breast cancer early detection biomarker are warranted.

Conclusions

Our study demonstrates that there may indeed changes in the plasma proteome prior to the clinical diagnosis of breast cancer that are detectable and of potential clinical utility. Ongoing efforts to follow up on promising candidates and formally validate them are warranted as confirmed breast cancer biomarkers could have several important clinical uses. These include as a companion to regular mammographic screening either to inform decision making regarding timing of subsequent screening (early recall in 6 months, next mammogram in 1 year, or next mammogram in 2 years) or to aid in the detection of cancers missed by mammography.

Conflict of interest statement Dr. Li has no financial, professional, or personal conflicts relevant to the content of this paper.

References

1. Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S (1993) Report of the International Workshop on Screening for Breast cancer. J Natl Cancer Inst 85:1644–1656
2. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ (2009) Cancer statistics, 2009. CA Cancer J Clin 59:225–249
3. Hortobagyi GN, de la Garza SJ, Pritchard K et al (2005) The global breast cancer burden: variations in epidemiology and survival. Clin Breast Cancer 6:391–401
4. Dixon JM, Anderson TJ, Page DL, Lee D, Duffy SW (1982) Infiltrating lobular carcinoma of the breast. Histopathology 6:149–161
5. Yeatman TJ, Cantor AB, Smith TJ et al (1995) Tumor biology of infiltrating lobular carcinoma. Implications for management. Ann Surg 222:549–559

6. Silverstein MJ, Lewinsky BS, Waisman JR et al (1994) Infiltrating lobular carcinoma. Is it different from infiltrating duct carcinoma? Cancer 73:1673–1677

7. Porter PL, El-Bastawissi AY, Mandelson MT et al (1999) Breast tumor characteristics as predictors of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst 91:2020–2028

8. Krecke KN, Gisvold JJ (1993) Invasive lobular carcinoma of the breast: mammographic findings and extent of disease at diagnosis in 184 patients. AJR Am J Roentgenol 161:957–960

9. Collett K, Stefansson IM, Eide J et al (2005) A basal epithelial phenotype is more frequent in interval breast cancers compared with screen detected tumors. Cancer Epidemiol Biomark Prev 14:1108–1112

10. The Women's Health Initiative Study Group (1998) Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials 19:61–109

11. Hays J, Hunt JR, Hubbell FA et al (2003) The Women's Health Initiative recruitment methods and results. Ann Epidemiol 13: S18–S77

12. Katayama H, Paczesny S, Prentice R et al (2009) Application of serum proteomics to the Women's Health Initiative conjugated equine estrogens trial reveals a multitude of effects relevant to clinical findings. Genome Med 1:47

13. Pitteri SJ, Amon LM, Buson TB, Zhang Y, Johnson MM, Chin A, Kennedy J, Wong C-H, Zhang Q, Wang H, Lampe PD, Prentice RL, McIntosh MW, Hanash SM, Li CI (2010) Detection of elevated plasma levels of epidermal growth factor receptor prior to breast cancer diagnosis among hormone therapy users. Cancer Res 70:8598–8606

14. Thompson IM, Ankerst DP, Chi C et al (2005) Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower. JAMA 294:66–70

15. Abd El-Rehim DM, Pinder SE, Paish CE et al (2004) Expression and co-expression of the members of the epidermal growth factor receptor (EGFR) family in invasive breast carcinoma. Br J Cancer 91:1532–1542

16. Tsutsui S, Kataoka A, Ohno S, Murakami S, Kinoshita J, Hachitanda Y (2002) Prognostic and predictive value of epidermal growth factor receptor in recurrent breast cancer. Clin Cancer Res 8:3454–3460

17. Hudelist G, Singer CF, Manavi M, Pischinger K, Kubista E, Czerwenka K (2003) Co-expression of ErbB-family members in human breast cancer: Her-2/neu is the preferred dimerization candidate in nodal-positive tumors. Breast Cancer Res Treat 80:353–361

18. Asgeirsson KS, Agrawal A, Allen C et al (2007) Serum epidermal growth factor receptor and HER2 expression in primary and metastatic breast cancer patients. Breast Cancer Res 9:R75

19. Muller V, Witzel I, Pantel K et al (2006) Prognostic and predictive impact of soluble epidermal growth factor receptor (sEGFR) protein in the serum of patients treated with chemotherapy for metastatic breast cancer. Anticancer Res 26:1479–1487

20. Souder C, Leitzel K, Ali SM et al (2006) Serum epidermal growth factor receptor/HER-2 predicts poor survival in patients with metastatic breast cancer. Cancer 107:2337–2345

21. Lafky JM, Baron AT, Cora EM et al (2005) Serum soluble epidermal growth factor receptor concentrations decrease in postmenopausal metastatic breast cancer patients treated with letrozole. Cancer Res 65:3059–3062

22. Hudelist G, Kostler WJ, Gschwantler-Kaulich D et al (2006) Serum EGFR levels and efficacy of trastuzumab-based therapy in patients with metastatic breast cancer. Eur J Cancer 42:186–192

23. Witzel I, Thomssen C, Krenkel S et al (2006) Clinical utility of determination of HER-2/neu and EGFR fragments in serum of patients with metastatic breast cancer. Int J Biol Markers 21:131–140