Comparison of six breast cancer classifiers using qPCR

Bioinformatics, 35(18), 2019, 3412–3420
doi: 10.1093/bioinformatics/btz103
Advance Access Publication Date: 13 February 2019
Original Paper

OXFORD

Systems biology

Comparison of six breast cancer classifiers using qPCR

Evi Berchtold[1,*], Martina Vetter[2], Melanie Gündert[3,†], Gergely Csaba[1], Christine Fathke[2], Susanne E. Ulbrich[3,‡], Christoph Thomssen[2], Ralf Zimmer[1,*,§] and Eva J. Kantelhardt[2,§]

[1]Department of Informatics, Institute of Bioinformatics, Ludwig-Maximilians-Universität München, München 80333, Germany, [2]Department of Gynecology, Institute of Clinical Epidemiology, Martin-Luther-Universität, Halle an der Saale 06120, Germany and [3]Physiology Weihenstephan, Technical University of Munich, Freising 85354, Germany

*To whom correspondence should be addressed.
[†]Present address: Institute of Diabetes Research, Helmholtz Zentrum München - German Research Center for Environmental Health, Munich-Neuherberg 85764, Germany
[‡]Present address: ETH Zurich, Animal Physiology, Institute of Agricultural Sciences, Zurich 8092, Switzerland
[§]The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Russell Schwartz

Received on July 30, 2018; revised on January 10, 2019; editorial decision on February 1, 2019; accepted on February 11, 2019

Abstract

Motivation: Several gene expression-based risk scores and subtype classifiers for breast cancer were developed to distinguish high- and low-risk patients. Evaluating the performance of these classifiers helps to decide which classifiers should be used in clinical practice for personal therapeutic recommendations. So far, studies that compared multiple classifiers in large independent patient cohorts mostly used microarray measurements. qPCR-based classifiers were not included in the comparison or had to be adapted to the different experimental platforms.

Results: We used a prospective study of 726 early breast cancer patients from seven certified German breast cancer centers. Patients were treated according to national guidelines and the expressions of 94 selected genes were measured by the mid-throughput qPCR platform Fluidigm. Clinical and pathological data including outcome over five years is available. Using these data, we could compare the performance of six classifiers (scmgene and research versions of PAM50, ROR-S, recurrence score, EndoPredict and GGI). Similar to other studies, we found a similar or even higher concordance between most of the classifiers and most were also able to differentiate high- and low-risk patients. The classifiers that were originally developed for microarray data still performed similarly using the Fluidigm data. Therefore, Fluidigm can be used to measure the gene expressions needed by several classifiers for a large cohort with little effort. In addition, we provide an interactive report of the results, which enables a transparent, in-depth comparison of classifiers and their prediction of individual patients.

Availability and implementation: https://services.bio.ifi.lmu.de/pia/.

Contact: berchtold@bio.ifi.lmu.de or zimmer@bio.ifi.lmu.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Breast cancer is a diverse disease for which several treatment options are available, depending on the specific type of the tumor (Dai *et al.*, 2015). Traditionally, histological factors such as hormone receptor (estrogen receptor (ER) and progesterone receptor (PgR)), human epidermal growth factor receptor 2 (HER2) status or histological tumor grade and clinical features are used to decide on the most suitable treatment.

Over the last decades several gene expression-based risk scores and subtype classifiers have been developed. These tools measure the gene expression of a small subset of genes and use these data to predict either the tumor subtype or a risk score that indicates the probability of recurrence. Several of these classifiers have been developed to commercial assays and are partially used in clinical practice. In the last years, there were two large prospective randomized clinical trials that analyzed the survival of patients who received treatment according to the classification of Mammaprint (70 genes; Cardoso *et al.*, 2016) and the Oncotype recurrence score (RS) (21 genes; Sparano *et al.*, 2015, 2018).

In 2011, Venet *et al.* (2011) reported that gene sets that are completely unrelated to breast cancer or even random gene sets can yield significant *P*-values for the prediction of risk of recurrence for breast cancer patients. Given this observation it seems hazardous to simply report a significant *P*-value on some cohort when presenting a new classifier, as is routinely done. Instead, the new classifier should be compared with existing classifiers to show that it has some advantage, e.g. improved performance, robustness or applicability. Furthermore, the already published classifiers need to be evaluated systematically on independent test sets that were not used in the development of any classifiers. In the last years, such studies have been published (Buus *et al.*, 2016; Fan *et al.*, 2006; Haibe-Kains *et al.*, 2008; Kelly *et al.*, 2012; Lundberg *et al.*, 2017; Martin *et al.*, 2016; Pelaez-Garcia *et al.*, 2017; Prat *et al.*, 2012a, b; Sestak *et al.*, 2018; Varga *et al.*, 2013), but most of these studies analyze quite small cohorts, only compare two classifiers or are based on microarray measurements, even though many of the available classifiers have been developed for qPCR measurements of the gene expression. Supplementary Table S1 shows a brief overview of these studies.

The Fluidigm Dynamic Array IFC qPCR platform (Spurgeon *et al.*, 2008) can help to decrease the cost of measuring the gene expression of many genes, as needed for breast cancer classifiers. For most classifiers, the gene expression of several genes is measured by qPCR. Traditional qPCR platforms require that each combination of patient sample and primers of the genes are pipetted together individually to be measured. This results in *patients*genes**2 pipetting steps. The Fluidigm IFC platform has a system of fluid lines and valves that automatically distribute the RNA samples and primers to the individual reaction chambers without mixing them. So only *patients* + *genes* pipetting steps are needed to measure hundreds of genes for hundreds of patients.

We have used the Fluidigm IFC platform to measure the expression of 94 genes for a large cohort of 726 patients. We selected the 94 genes such that they cover six different breast cancer classifiers: scmgene (Haibe-Kains *et al.*, 2012), the research versions of PAM50 and the corresponding risk of recurrence score (ROR-S; Bernard *et al.*, 2009), EndoPredict (and its variant EPclin that incorporates clinical variables; Filipits *et al.*, 2011), Genomic Grade Index (GGI; Toussaint *et al.*, 2009) and RS (Paik *et al.*, 2004). Thus, we can compare the prognostic power of these classifiers on an independent routine cohort on which none of the classifiers was trained and show that the Fluidigm IFC platform can be used to measure the gene expression of the many genes needed for such a comparison study.

2 Materials and Methods

2.1 PiA cohort

Within the multicenter prospective PiA study (NCT 01592825) tumor tissue samples of consecutively diagnosed breast cancer patients from 7 German certified breast centers (Hospital Fürth, St. Elisabeth and St. Barbara Hospital Halle, St. Bernward Hospital Hildesheim, Helios Hospital Hildesheim, Medical Office Uleer Hildesheim, Hospital Martha-Maria Dölau Halle and Asklepius Harzkliniken Goslar, see Supplementary Material for more information) were collected at Martin-Luther-University, Halle-Wittenberg between 2009 and 2011. Female patients with operable, non-metastasized breast cancer independent of lymph node status were included. The study was approved by the ethics committee of the Martin-Luther-University Halle-Wittenberg and each patient gave informed consent. A total of 726 fresh frozen samples of primary tumor tissue were investigated using Fluidigm IFC platform (Spurgeon *et al.*, 2008). Tumor specimens were fresh frozen after surgery and stored at −80°C until further use. A total of 264 patients were not included as only formalin-fixed paraffin embedded material was obtained since tumors were, for example, too small to separate frozen tissue and 210 patients were assigned for neoadjuvant chemotherapy. Tumor content was verified histologically. Clinical and pathological parameters were obtained for each patient and documented using SPSS 24 (SPSS Inc., Chicago, IL, USA). TNM staging system was used (Sobin *et al.*, 2011). Information on therapy applied was not available. Patient information was anonymized prior to analysis. Receptor defined breast cancer subtypes were determined according to the St. Gallen classification (Goldhirsch *et al.*, 2013), cutoff [ER = 1%, PgR = 20% or IRS≥3]. Due to missing Ki-67 values, we used histopathological grading to assess cell proliferation (Von Minckwitz *et al.*, 2012).

The following system was applied to define histopathological subtypes:

- Luminal A-like: ER-positive, PgR-positive, HER2-negative, grade 1 or 2.
- Luminal B-like (HER2-negative): ER-positive, PgR-negative, HER2-negative or grade 3.
- Luminal B-like (HER2-positive): ER-positive, HER2-positive, any grades.
- HER2-positive (non-luminal-like): ER-negative, PgR-negative, HER2-positive, any grade.
- Triple-negative breast cancer (TNBC, Basal-like): ER-negative, PgR-negative, HER2-negative, any grade.

An overview of the clinical and histopathological characteristics of the patients and tumors is shown in Table 1. Most of the tumors (610 of 726) are ER-positive and only a small subset (104) is HER2-positive. The majority of the tumors had histological grade 2, and lymph nodes were not affected.

The standardized definitions for efficacy endpoints (STEEP) criteria were used as endpoint definitions (Hudis *et al.*, 2007). The primary endpoint of this study was overall survival (OS). Person time equaled the time from the date of diagnosis to the date of event or to the date of last contact. Women without event were right-censored at the last visit to the clinic.

2.2 Gene expression measurement

Expressions of 94 genes were measured using the Fluidigm qPCR platform. This amounts to 726×94 = 68.244 qPCR reactions. To ensure that the measurements of the Fluidigm platform are of

Table 1. Clinical characteristics of the PiA cohort, grouped by histopathological subtype

	All	Luminal A-like	Luminal B-like (HER2-negative)	Luminal B-like (HER2-positive)	HER2-positive (non-luminal-like)	Triple-negative breast cancer (TNBC, Basal-like)	Not classified
No. patients	726	378	163	69	34	74	8
Grade							
1	76	67	4	3	0	0	2
2	447	311	59	40	12	22	3
3	203	0	100	26	22	52	3
Size							
<1	42	22	9	2	4	5	0
1–2	302	176	69	24	13	16	4
2–5	341	161	77	37	16	46	4
>5	41	19	8	6	1	7	0
Nodal status							
0	450	239	102	41	21	42	5
1	201	108	48	16	8	20	1
2	47	22	5	7	4	7	2
3	28	9	8	5	1	5	0
Age							
Avrg	62.62	62.46	64.89	59.19	61.32	63.11	54.25
Min	22	22	29	28	31	25	30
Max	90	89	90	86	81	88	75
Survival							
Alive	630	348	136	58	28	53	7
Deceased	96	30	27	11	6	21	1

Note: Patients who do not fall in any category described in Section 2.1 are shown in the last column.

good quality and comparable across chips, for all samples five genes were also measured on the CFX384 qPCR platform, so that the results could be compared. This platform uses 384-well plates, so that qPCR measurements for one gene can be done in parallel for 384 samples. Due to technical problems, SNPs or too low mRNA abundance some gene expression values are missing. Some of the classifiers are not able to make a prediction for a patient with missing values. We circumvent this by substituting the missing value if the missing gene(s) do not influence the prediction. Nevertheless, a number of cases have to be excluded from the analysis of the classifier. For more information on the Fluidigm gene expression measurement and missing values, see Supplementary Material.

On one Fluidigm IFC chip 96 genes can be measured by qPCR for 96 samples. Thus, the 726 patients have been measured on several chips that need to be normalized to make them comparable. There are three sources of bias when several Fluidigm chips are measured: the amount of cDNA can differ between samples (within a chip and between chips), there can be variation between the chips, e.g. due to different efficiency of the PCR reactions and there can be differences in the pre-amplification of the cDNA that is necessary for the Fluidigm platform. To correct for variation between chips, so-called inter plate calibrator samples are measured on each chip. The difference between cDNA amounts of individual samples can be diminished, by using the expression of genes that are expected to be constant between samples, e.g. housekeeping genes. Most classifiers already include housekeeping genes for normalization purposes so that no additional genes have to be measured. The cDNA has to be pre-amplified before it is loaded on the Fluidigm IFC chip. Amplification for all 96 primers at once can generate problems, so that we splitted the set of primers in 2 subsets that are amplified individually. For this we tried several different batches and used the division that yielded most successful amplifications. However, there can be differences between the efficiencies of the pre-amplification reactions. This can be corrected as one can assume that the median of all measurements of each chip and pre-amplification mix is the same. For more information on the individual normalization steps, see Supplementary Material.

2.3 Classification

The genefu R package (Gendoo *et al.*, 2016; R Core Team, 2016) was used to calculate the research versions of PAM50, scmgene, ROR-S and RS. The PAM50 classifier can be applied in two ways: the published centroids can be used directly for the prediction, or the centroids are first trained on the given dataset and then used to predict the subtypes (both using default parameters). As a high $C(t)$ value indicates low gene expression whereas a high microarray intensity indicates high gene expression, the $C(t)$ values were not used directly for these microarray-based methods, instead the difference to the maximal PCR cycle $C(t)_{\max}$ was used. For GGI, EndoPredict and EPclin the formulas from the corresponding papers were re-implemented and the published cutoffs were used for EndoPredict and EPclin. For GGI no published cutoff is available, so that we used the median to divide the cohort in two equally sized groups. All classifiers were applied to the complete cohort.

2.4 Performance and concordance of predictions

To assess the performance of the predictions, we generated Kaplan–Meier plots and calculated the concordance index (c-index) for each classifier. The c-index corresponds to the probability that for a pair of randomly chosen samples, the sample with the higher risk score experiences an event before the other sample.

As we are able to calculate several classifiers for the same cohort, we compared their predictions by calculating Spearman's correlations and Cramer's V which quantifies the association between two nominal predictions. It ranges between 0 and 1, with values above

0.5 indicating a strong association. We compared subtype classifiers (PAM50 and scmgene) and risk scores separately, to account for the different number of predicted groups.

Moreover, we used multivariate Cox regression to create a combined predictor that uses the risk scores of the different classifiers as input. For this, only risk scores that return a numeric risk score were used (excluding PAM50 and scmgene) and their scores were scaled, so that scores yielding a low risk prediction (i.e. having a score below the corresponding cutoff) are mapped to 0–0.5 and high risk scores to 0.5–1. Most risk scores are not able to return a score if one of the measurements is missing due to technical errors during the measurement. In this case, the combined risk score is also not able to return a score. As this is more probable when more genes are used, the combined risk score cannot return a score for many patients. To nevertheless return a score for these patients, we trained multiple models, excluding each risk score in turn. For the final prediction, we used the model that uses all risk scores, and only used one of the restricted models if the complete model does not return a risk score. To evaluate the performance of this combined risk score, a 5-fold cross validation was used to prevent overfitting.

2.5 Robustness of classifications

Like all measurements, also gene expression measurements are subject to noise. As most subtype classifiers use a combination of many genes, the impact of noisy measurements is reduced, as no single gene influences the prediction too strongly. To assess the impact of noise on the prediction, we simulated noisy measurements and checked how often the prediction changed due to small changes in the gene expression data. For this, we repeatedly sampled for each measurement a noise term from a normal distribution centered around zero and added it to the measurement. Then we checked for each classifier, whether the same subtype or risk group (high or low) was predicted for the real and modified measurement. Robust classifiers should be able to make the same prediction for the real and modified measurements with simulated noise in most cases.

A similar approach allows us to estimate the probability that a single noisy measurement results in a false prediction for a given patient. For this we calculate for each gene contained in the classifier the minimal difference of the gene expression value that would result in a different prediction. For classifiers with simple formulas this can be calculated directly, while it can be sampled by calculating the score with a growing noise term for more complex classifiers. Given a background noise distribution (e.g. a normal distribution with mean zero) the probability of observing at least as much noise can be calculated. These probability values can help to identify gene expression measurements for which already small (i.e. highly probable) deviations have an effect on the prediction. For these measurements replicate measurements can then be considered to reduce the impact of random noise and improve the quality of the prediction.

2.6 Interactive report

In addition to the results presented in this paper, we provide a website (https://services.bio.ifi.lmu.de/pia/) that contains an interactive report of the results. The overview page contains all the main results: the clinical and pathological characteristics table, performance table, coherence plot and Cramer's V table and additionally an overview of all features for all patients. In the clinical characteristics table for large enough patient groups with similar characteristics the performance results for this sub-cohort can be analyzed. Moreover, for each entry in the performance table the corresponding Kaplan–Meier plot can be shown in a popup window, to evaluate the performance in more detail. The survival endpoint used in the Kaplan–Meier plot can be selected to directly compare the influence of the different survival endpoints. Furthermore, a page comparing two classifiers is linked to the corresponding entry of the Cramer's V table. This comparison page shows both Kaplan–Meier plots side by side, so that they can be compared directly. Furthermore, a contingency table shows how many patients are classified with a given combination of classifications of the two selected classifiers. This table is again linked to a list of the corresponding patients, with all available clinical features, classifications and survival information. This way, one can analyze the patients who were classified discordantly in full detail. The patient overview table is linked to a details view for each individual patient. This view not only shows the available features of this patient, but also for each classifier an overview of the corresponding gene expression measurements and how they relate to the distribution of the gene expression measurements of the whole cohort, or the subsets that experienced an event or not. Furthermore, the minimal difference in gene expression to change the prediction and the corresponding probability to experience this difference due to random noise is shown for each gene contained in the classifier. Such a detailed view on individual patients can greatly help to understand individual predictions and the influence of the contained genes.

3 Results

3.1 Comparability of Fluidigm chips

With appropriate normalization the different Fluidigm chips should be comparable. To test this, the CFX and Fluidigm measurements were compared for the five genes that were also measured on the CFX platform. Figure 1 shows the comparison of the $C(t)$ values of the two platforms for the reference gene RPLP0. The different Fluidigm chips are highlighted by different colors and there is only some bias for chips 1 and 2. For the first three chips, the sample amounts differed slightly as they were not done in one batch with the other chips. This variation is normally corrected for by the housekeeping normalization that was not applied for this comparison due to the small number of genes on the CFX platform. The concordance between the two measurements is quite good with only few outliers. The $C(t)$ values are shifted between the different qPCR platforms as they are using different amounts of cDNA and the cDNA is pre-amplified for the Fluidigm platform. But in general, the two platforms agree very well, so that the Fluidigm platform seems to be suitable for its use in gene expression profiling also of large cohorts using multiple chips.

3.2 Survival analysis

For the PiA study endpoints were assessed five years after the end of recruitment. Median observation time of patients alive was 5.13 years. We analyzed the OS ($n = 97$ events), invasive disease-free survival ($n = 122$ events), distant disease-free survival ($n = 117$ events) and recurrence-free interval (RFI = 67 events), all defined according to STEEP criteria (Hudis *et al.*, 2007). In this paper, we focus on OS, the results for the other endpoints can be found in the interactive report. The survival data were used to calculate different measures for the performance of the risk scores: hazard ratios, log-rank P-values and the c-index.

Table 2 shows these measures for all risk scores. The corresponding Kaplan–Meier plots are available in the Supplementary Material and the interactive report. All risk scores yield significant P-values, hazard ratios well above 1 and a c-index above 0.5. Values above

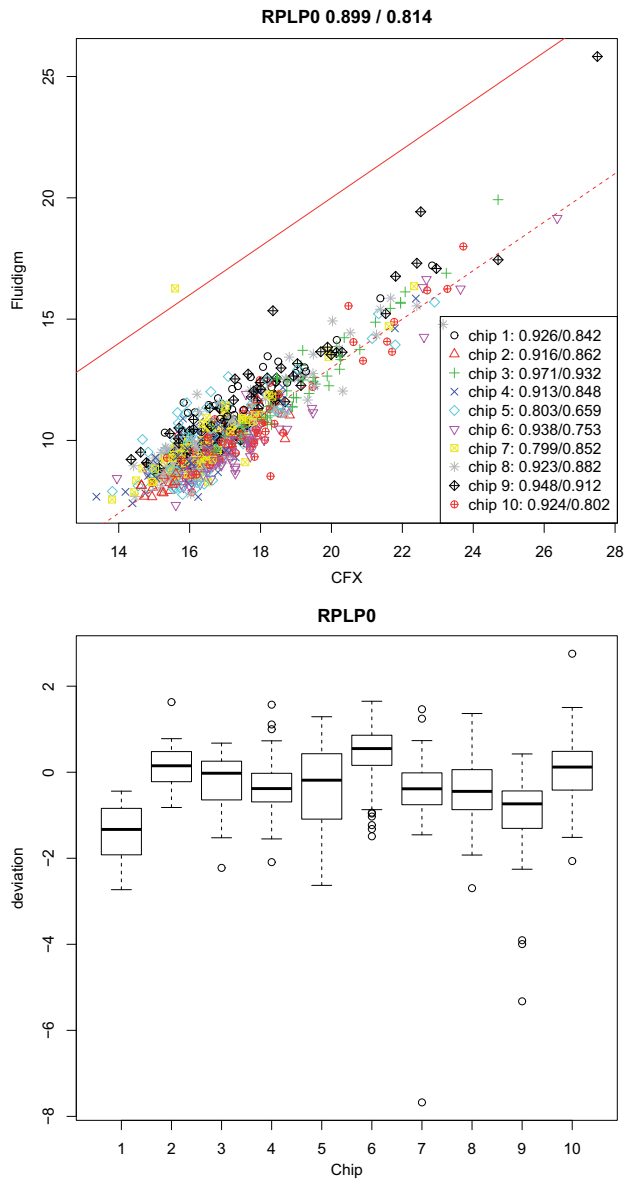

Fig. 1. Comparison of C(t) values for the RPLP0 gene for 726 samples measured on 10 Fluidigm chips and the CFX platform. On the top the C(t) values are scattered against each other. The correlation between the two measurements was calculated using all measurements (first number) and excluding the outliers (below 5% quantile or above 95% quantile, second number). The correlations are given for each chip separately (see legend) and for all chips combined (see title). There is a shift in the absolute $C(t)$ values due to the different cDNA concentrations and the pre-amplification, but there is a clear correlation between the two measurements and no apparent bias between the Fluidigm chips. The bottom plot shows the deviations between the Fluidigm and CFX measurements for each Fluidigm chip separately. (Color version of this figure is available at *Bioinformatics* online.)

0.7 are often considered to indicate good prognostic ability for the c-index. For the endpoint OS, only EPclin yields a c-index above 0.7 whereas RS, EndoPredict and ROR-S have scores slightly below 0.7. For the RFI, however, all risk scores yield c-index scores above 0.7. Interestingly, PAM50 yields a very high hazard ratio and low *P*-value for the RFI endpoint. For most endpoints, EPclin performs best: it yields both the lowest *P*-value and the highest c-index. For the OS endpoint, of the 292 patients in the low-risk group of EPclin, only 17 had an event, while 75 of the 395 patients from the high-risk group had an event after five years. For GGI on the other hand, 28 of the 363 low-risk patients and 68 of the 363 high-risk patients experienced an event. Additionally, for each classifier we performed a multivariate Cox regression incorporating clinical features (Supplementary Material). All classifiers except scmgene contributed additional prognostic information. The combined risk score, derived from the multivariate Cox regression performs even slightly better than EPclin, with a lower *P*-value, higher hazard ratio and comparable c-index. However, the effect is moderate, given the increased number of measurements needed.

Gene expression-based classifiers are especially interesting for patients whose histopathological features are neither clearly associated with low or high risk. We therefore repeated the analysis limiting the cohort to 370 patients with intermediate risk according to histopathological features (ER+/PgR+/HER2− patients with grade 2). In this sub-cohort, ROR-S and GGI perform slightly better than the other risk scores (Supplementary Material). Generally, the *P*-values are higher for all risk scores as these patients cannot be classified into low and high risk as easily as the other patients.

For the two subtype classifiers PAM50, scmgene and the histopathological classification, the values for the Luminal A (low risk) subtypes are shown. While for PAM50 the Luminal A patients have significantly better prognosis, for scmgene the logrank *P*-value is only 0.001 and also its hazard ratio of 1.48 is by far the lowest of all classifiers. The histopathological classification that does not take any gene expression measurements into account preforms similarly well as the other classifiers.

3.3 Concordance of classifications

Figure 2 shows the predictions of all classifiers, as well as some clinical characteristics for all patients. Each row corresponds to one classifier/characteristic and each column corresponds to one patient. The patients are ordered in the same way in all rows (according to the histopathological subtype), so that the predictions/characteristics can be compared for each patient. Both variants of PAM50 [using the published model (PAM50) or training a new model (PAM50 new)] yielded similar results. The main difference is that the newly trained model only returns four subtypes, so that the normal-like subtype is missing. The predicted subtypes are in many cases the same as the histopathological subtype, only for HER2 and Luminal B subtype patients, the two classifications differ. The predictions of scmgene that only uses three genes to predict the subtype differ in many cases from the prediction of PAM50. Especially the normal-like patients are predicted to be basal according to scmgene, while the newly trained PAM50 classifies them as Luminal A. These patients are assigned a low risk score by all other methods and they are ER-positive and HER2-negative according to the immunohistological measurements. Also, only 2 of the 19 patients had an event within five years, so these are likely false predictions of scmgene.

All the risk scores predict predominantly low risk scores for the patients who had Luminal A or normal-like subtypes, and high risk scores for the basal and HER2 subtypes according to PAM50. Their predictions differ most for the Luminal B patients. Here, GGI and EPclin predict high scores for most patients, whereas EndoPredict and RS yield mostly low scores. The RS did not return a risk score for many patients, as it uses 21 genes, and cannot return a result if a measurement for any of these genes is missing.

Table 3 shows the correlation of the risk scores and the Cramer's V statistic for the subtype classifiers. All risk scores correspond quite well to each other, with Spearman's rank correlation values about 0.7–0.9. Additionally, the Cramer's V statistic for the risk score's classifications into low- and high-risk patients is given in

Table 2. Logrank *P*-values, hazard ratios (HR) and concordance index (*c*-index) for the different risk scores

	OS					RFI				
Risk score	logrank *P*	HR	*c*-index	No. event	No. no event	logrank *P*	HR	*c*-index	No. event	No. no event
Recurrence score	2.818e−7	3.47	0.66 (0.54–0.76)	62/20	220/272	9.406e−9	4.36	0.73 (0.58–0.84)	46/10	236/285
EndoPredict	4.546e−6	3.75	0.69 (0.57–0.78)	80/12	366/230	2.715e−7	11.17	0.78 (0.65–0.87)	59/3	387/239
EPclin	1.20e−6	3.41	0.72 (0.61–0.81)	75/17	320/275	7.159e−8	7.30	0.80 (0.67–0.89)	56/6	339/286
GGI	9.19e−6	2.61	0.64 (0.52–0.73)	68/28	295/335	2.871e−7	4.29	0.70 (0.57–0.81)	53/13	310/350
ROR-S	3.03e−6	3.43	0.68 (0.57–0.77)	88/8	430/200	8.360e−6	7.15	0.75 (0.62–0.85)	62/4	456/204
Combination	8.644e−7	4.49	0.72 (0.61–0.81)	82/9	367/230	7.77e−7	10.46	0.70 (0.59–0.79)	57/4	392/235
PAM50	1.678e−5	3.82	–	72/24	331/299	3.054e−12	11.25	–	59/7	344/316
scmgene	0.001	1.48	–	53/12	313/183	1.086e−2	1.80	–	31/7	335/188
histopathological	7.184–6	2.45	–	66/30	274/348	2.566e−9	6.57	–	51/15	289/363

Note: Additionally, the number of patients with high/low risk score with and without an event is given. On the left the results for the overall survival (OS) endpoint and on the right for the recurrence-free interval (RFI) are shown. For the concordance index, the lower and upper bound of the 95% confidence interval is given in brackets. For the subtype classifiers PAM50, scmgene and the histopathological classification we used the values for the Luminal A (low risk) subtype to make them comparable to the binary predictions of the risk scores. For all risk scores, the low- and high-risk patients differ significantly in their survival, but overall, EPclin performed best.

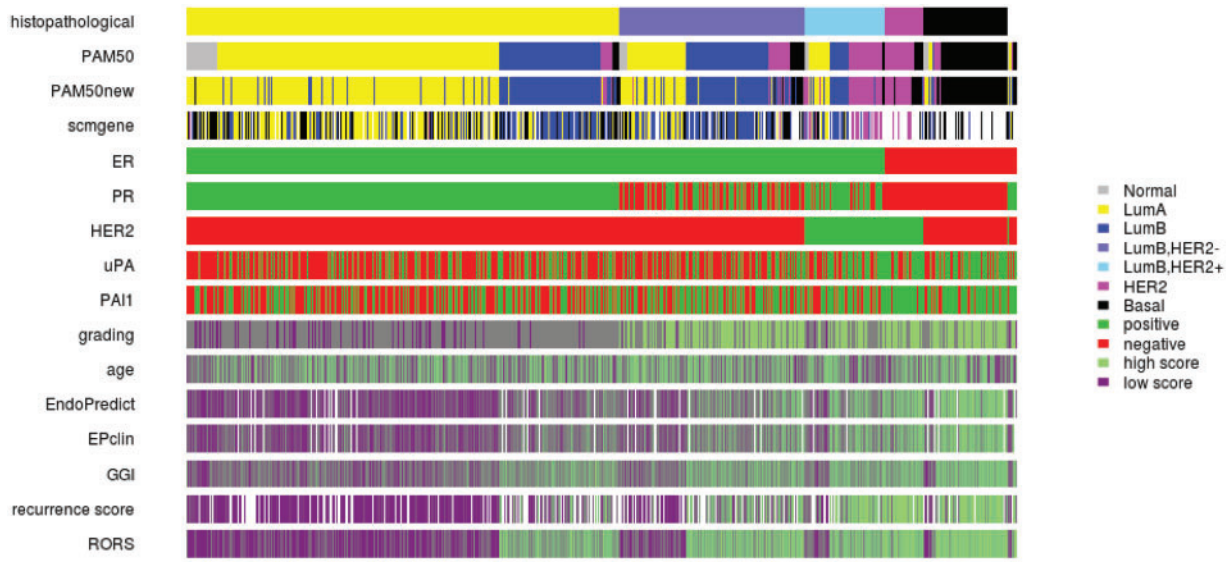

Fig. 2. Overview of classification results and clinical variables for all patients. The first four rows correspond to subtype classifications, the next 7 rows are clinical characteristics, and the remaining rows are risk scores. A continuous scale between green and purple is used for numeric values such as the risk scores or age and grading and different colors for the categorial attributes. The different subtype classifications are mapped to each other by using prior knowledge (e.g. slightly different names for the Luminal A subtype by PAM50, scmgene or the histopathological classification) or by maximizing the overlap to the histopathological classification (for the newly trained PAM50). (Color version of this figure is available at *Bioinformatics* online.)

Supplementary Material. The concordance of the subtype classifiers was inferior to the risk scores. Only the published and newly trained PAM50 classifiers corresponded well to each other, while scmgene only yielded Cramer's V statistics of 0.484 and 0.486. Note that in the literature Cramer's V values between 0.36 and 0.49 are considered a substantial relation while values above 0.5 indicate a strong relation. We also compared the subtype classifier's predictions to the clinical histopathological subtypes. The newly trained PAM50 had the highest correspondence with these clinical subtypes, yielding a Cramer's V value of 0.58, whereas scmgene again yielded the least correspondence with a Cramer's V value of 0.419.

3.4 Robustness to noise

To analyze the robustness of the classifiers to experimental noise, we simulated 100 datasets where we added a small noise term to each measurement, and compared the resulting prediction to the predictions without noise. Figure 3a shows for each classifier how many patients were misclassified how often in the 100 runs, using a normal distribution with mean 0 and SD 0.7 [$N(0, 0.7)$] as noise distribution. The ROR-S score performed best, with 506 patients without any misclassification, respectively. Interestingly, PAM50 with a newly trained model seems to overfit and yields for many patients different predictions when noise is added. Only 219 patients were never or only once misclassified. Similarly, scmgene is very sensitive to noise and yields different predictions for nearly all patients: only 44 patients were never or only once misclassified. The robustness to noise does not seem to depend only on the number of genes used by the classifier, as, e.g. the RS that uses 21 genes, performs slightly worse than EPclin that uses only 7 genes and 2 clinical features. It might rather depend on the way the gene expression measurements are used or which genes are selected by the classifier.

We repeated this simulation using a smaller noise term sampled from an $N(0, 0.3)$. The newly trained PAM50 and scmgene still yielded many misclassifications for most patients. The other risk

scores, however, became comparable to ROR-S, except that all but the RS yielded more patients who were misclassified in many of the noisy datasets (≥5 misclassifications).

Moreover, we calculated for each patient and classifier, how much each individual gene would have to differ to change the prediction. The probability of observing noise at least that high can be calculated if a given noise distribution [e.g. $N(0, 0.7)$] is assumed. These probabilities range from 0 (for measurements that would have to be changed a lot to alter the prediction) to ∼0.6 for our cohort and are available in the interactive report. This way, measurements that are very susceptible to noise can be identified and if possible replicate measurements can reduce the impact of noise for these measurements.

3.5 Interactive report

Figure 4 shows two screenshots of the iReport. The screenshot on the left is part of the overall view that shows a summary of the main results discussed in the paper. It shows an interactive version of the concordance plot of Figure 2. The user can select which features are included in the plot and by which classifiers the patients should be ordered. This allows to compare several features at once. In Figure 4a, the patients are ordered first by the tumor grade and then after the GGI risk score that was developed to determine the grade by gene expression. The corresponding two rows are shown at the top of the plot. As can be seen there is some concordance between the two features, with patients with low grade (purple block on the left in first (grade) row) have predominantly low GGI scores, and patients with high grade (green block on the right) have higher GGI scores. However, the majority of patients have intermediate grade and these patients show a distribution of both, high and low, GGI scores.

The screenshot on the right (Fig. 4b) shows the comparison view for PAM50 and EPclin. It contains the two Kaplan–Meier plots side by side and a contingency table below. The cutoff of EPclin that is used to divide the patients into low and high risk can be modified and the corresponding Kaplan–Meier plot will be updated accordingly. The contingency table shows how many patients are classified by the different combinations of subgroups of the two classifiers. The numbers in this table are linked to the corresponding list of patients, so that by clicking on them a table showing all available features of the patients is shown. This way subsets of patients can be analyzed in more detail. For example, by clicking on the corresponding entry in the contingency table, all information for the 82 patients who were classified as Luminal A by PAM50 and high risk by EPclin is shown. This allows the user to look at the survival status of these patients and see that only 11 of these 82 patients are still alive after five years, which justifies the high-risk prediction of EPclin.

Table 3. Correlation for risk scores (above) and Cramer's V for classifiers (below)

	RORS	RS	EndoPredict	EPclin	GGI
RORS	1.000	0.800	0.811	0.754	0.857
RS		1.000	0.824	0.748	0.770
EndoPredict			1.000	0.889	0.753
EPclin				1.000	0.715
GGI					1.000

	PAM50	PAM50 new	scmgene	histopathological
PAM50	1.000	0.837	0.484	0.478
PAM50 new		1.000	0.486	0.578
scmgene			1.000	0.419
histopathological				1.000

4 Discussion

The Fluidigm IFC platform allows to measure the expression of many genes for many patients at rather low cost and with little effort. In this paper, we showed that it can be used to measure the genes required for several breast cancer classifiers in a large cohort, which enabled us to systematically compare and evaluate these classifiers. For a smaller set of five genes, we measured the expression also on a different qPCR platform and the results showed a good agreement between the different platforms after normalization.

The comparison of the classifiers showed that they all performed well on our independent cohort. This shows that the classifiers do not overfit for the cohort on which they were trained but that they are applicable also using a different methodology (Fluidigm) and this new cohort. They provide good estimates of the risk of recurrence of the individual patients. Also their predictions were concordant, which also

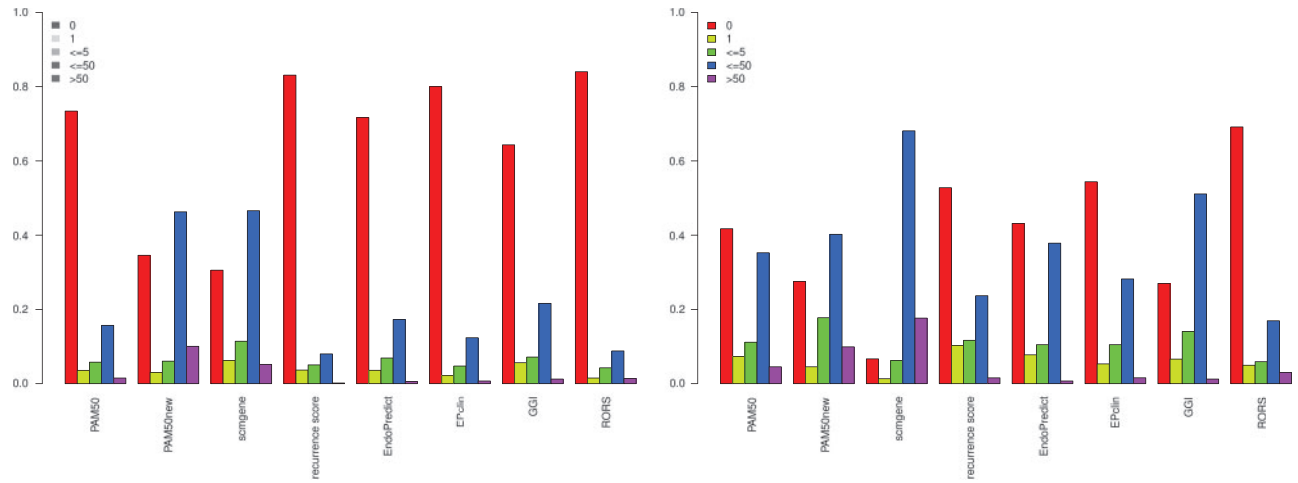

Fig. 3. Number of patients with a given number of misclassifications for each classifier when noise sampled from $N(0, 0.7)$ (left) and $N(0, 0.3)$ (right) is added to the measurements. For each patient random noise is added 100 times and the patients are classified according to the number of misclassifications in the 100 simulations. (Color version of this figure is available at *Bioinformatics* online.)

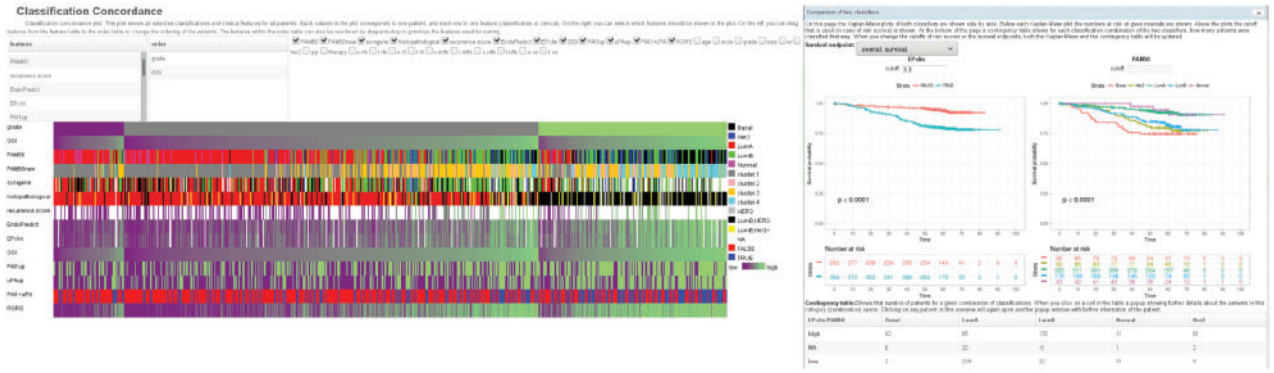

Fig. 4. Screenshots of the iReport. On the left the concordance plot sorted by grade and GGI (top two rows) is shown. The sorting can be modified interactively so that the plot can be used to compare different features. On the right, the comparison of EPclin and PAM50 with both Kaplan–Meier plots and the contingency table is shown. The cutoff used to separate high- and low-risk patients of EPclin can be adapted and the contingency table is linked to a table showing all available features for the patients in a specific cell. (Color version of this figure is available at *Bioinformatics* online.)

explains why a combined risk score integrating several classifiers yielded only a slightly better performance. It should be noted that also the histopathological classification that is easily available in routine practice performed well concerning the discrimination of high- and low-risk patients as well as allocating a similar number of low-risk patients compared with the other classifiers. Most classifiers performed similarly well, only scmgene performed less good and also the concordance to the PAM50 classifier was lower as described in its original paper (Haibe-Kains *et al.*, 2012; 57% identical predictions compared with 70% reported in the paper). As it only uses the expression of three genes it is also less robust to noise. However, we cannot decide whether this difference is due to the different cohort or due to the different experimental platform.

Moreover, we analyzed the robustness of these classifiers with respect to noise by simulating noisy measurements by adding a random noise term. The results showed that especially the classifiers that are newly trained on each cohort, like scmgene or a newly trained model using the PAM50 algorithm, are very sensitive to noise. This also indicates that the cohort that is used to train a new classifier must be of very good quality as noisy measurements can greatly impair the quality of the classifier. Furthermore, also between the classifiers with a fixed model there were large differences in their robustness to noise, as e.g. GGI yielded the same prediction for all 100 noisy measurements only in half as many patients as ROR-S. Furthermore, this kind of noise analysis can also be used to attribute each measurement with a probability that noise changes the prediction for a given patient. This can be used to identify measurements for additional replicates to reduce the impact of noise.

It has to be noted that our unselected cohort was comprised of patients with relatively good clinical prognostic factors. Those HER2-positive or receptor-negative cases who received neoadjuvant chemotherapy were not included since fresh frozen material has not been available. This leads to underrepresentation of clearly high-risk HER2-positive patients and under-representation of clearly low-risk very small tumors. Thus the proportion of certain high- and low-risk patients is reduced and effects probably become smaller. The classifiers perform differently on cohorts with higher proportions of these patients. In this work, we demonstrated feasibility to analyze a large number of genes by qPCR and use the publicly available research versions of the classifiers on that same cohort. Second, because we used the research versions of the classifiers and not the commercial versions, the results may differ slightly. Third, we could not include information on therapy which certainly had an effect on outcome: chemotherapy improving survival of high-risk patients, endocrine treatment improving survival of ER-positive patients and targeted therapy improving survival of HER2-positive patients. Thus the differences between the high- and low-risk groups are diminished.

All the results of this paper are also available as interactive report (iReport) on the accompanying website in order to make all results reproducible and transparent. This website allows to analyze the results and especially the differences between the classifiers in much more detail as is possible in a paper. The online tool allows selection of cases, strata, classifiers, endpoints and visualization of results. Cross-sectional comparison of clinical and histopathological data and classifiers assigned to each patient can be seen. Longitudinal data are shown as Kaplan–Meier curves as by defined groups. Thus on the one hand, the iReport provides an easy to use interface to results that cannot be shown in a paper due to page limitations, as e.g. the Kaplan–Meier plots for all classifiers for all survival endpoints. On the other hand, it also includes much more detail for individual results by linking the raw data to the summarized result, as is, e.g. done by showing the patient lists with all available data for the contingency table of the classifications of two classifiers. We believe that this detailed data can help to generate new hypotheses, e.g. about the patients who are discordantly classified and can thus help the further development of new classifiers.

Acknowledgements

We would like to thank all patients and investigators of the following hospitals and cooperating institutions that recruited patients for this study: Volker Hanf, Department of Gynaecology, Nathanstift, Hospital Fürth, Tilmann Lantzsch, Department of Gynaecology St. Elisabeth & St. Barbara Hospital Halle (Saale), Christoph Uleer, Medical Office Uleer Hildesheim, Susanne Peschel, Department of Gynaecology St. Bernward Hospital Hildesheim, Jutta John, Department of Gynaecology Helios Hospital Hildesheim, Marleen Poehler, Department of Gynaecology, Asklepius Harzkliniken, Goslar, Edith Weigert, Institute of Pathology Hospital Fürth, Jörg Buchmann, Institute of Pathology Hospital Martha-Maria Dölau and Karl-Friedrich Bürrig, Institute of Pathology Hildesheim. We would like to thank Kathrin Stückrath and Sandy Kaufhold for performing experiments. In addition, we like to thank the TATAA Biocenter Core Facility for provision of the gene expression analysis services.

Funding

This study was supported by intramural research funding through the Roux-Program of the Martin-Luther-University Halle-Wittenberg [number 25/36

(2012) and 21/14 (2010)]. R.Z. acknowledges partial funding of this work from the DFG (SFB 1123).

Conflict of Interest: none declared.

References

Bernard,P.S. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Buus,R. *et al.* (2016) Comparison of EndoPredict and EPclin with oncotype DX recurrence score for prediction of risk of distant recurrence after endocrine therapy. *J. Natl. Cancer Inst.*, **108**, djw149.

Cardoso,F. *et al.* (2016) 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.*, **375**, 717–729.

Dai,X. *et al.* (2015) Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.*, **5**, 2929–2943.

Fan,C. *et al.* (2006) Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.*, **355**, 560–569.

Filipits,M. *et al.* (2011) A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. *Clin. Cancer Res.*, **17**, 6012–6020.

Gendoo,D.M. *et al.* (2016) Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*, **32**, 1097–1099.

Goldhirsch,A. *et al.* (2013) Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann. Oncol.*, **24**, 2206–2223.

Haibe-Kains,B. *et al.* (2008) Comparison of prognostic gene expression signatures for breast cancer. *BMC Genomics*, **9**, 394.

Haibe-Kains,B. *et al.* (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst.*, **104**, 311–325.

Hudis,C.A. *et al.* (2007) Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. *J. Clin. Oncol.*, **25**, 2127–2132.

Kelly,C. *et al.* (2012) Agreement in risk prediction between the 21-gene recurrence score assay and the PAM50 breast cancer intrinsic classifier in early-stage estrogen receptor-positive breast cancer. *Oncologist*, **17**, 492–498.

Lundberg,A. *et al.* (2017) Gene expression signatures and immunohistochemical subtypes add prognostic value to each other in breast cancer cohorts. *Clin. Cancer Res.*, **23**, 7512–7520.

Martin,M. *et al.* (2016) Prognostic ability of EndoPredict compared to research-based versions of the PAM50 risk of recurrence (ROR) scores in node-positive, estrogen receptor-positive, and HER2-negative breast cancer. A GEICAM/9906 sub-study. *Breast Cancer Res. Treat.*, **156**, 81–89.

Paik,S. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.*, **351**, 2817–2826.

Pelaez-Garcia,A. *et al.* (2017) Comparison of risk classification between EndoPredict and MammaPrint in ER-positive/HER2-negative primary invasive breast cancer. *PLoS One*, **12**, e0183452.

Prat,A. *et al.* (2012a) Concordance among gene expression-based predictors for ER-positive breast cancer treated with adjuvant tamoxifen. *Ann. Oncol.*, **23**, 2866–2873.

Prat,A. *et al.* (2012b) PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Res. Treat.*, **135**, 301–306.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sestak,I. *et al.* (2018) Comparison of the performance of 6 prognostic signatures for estrogen receptor-positive breast cancer a secondary analysis of a randomized clinical trial. *JAMA Oncol.*, **4**, 545–553.

Sobin,L.H. *et al.* (2011) *TNM Classification of Malignant Tumours*. Hoboken, John Wiley & Sons.

Sparano,J.A. *et al.* (2015) Prospective validation of a 21-gene expression assay in breast cancer. *N. Engl. J. Med.*, **373**, 2005–2014.

Sparano,J.A. *et al.* (2018) Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med.*, **379**, 111–121.

Spurgeon,S.L. *et al.* (2008) High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. *PLoS One*, **3**, e1662.

Toussaint,J. *et al.* (2009) Improvement of the clinical applicability of the Genomic Grade Index through a qRT-PCR test performed on frozen and formalin-fixed paraffin-embedded tissues. *BMC Genomics*, **10**, 424.

Varga,Z. *et al.* (2013) Comparison of EndoPredict and Oncotype DX test results in hormone receptor positive invasive breast cancer. *PLoS One*, **8**, e58483.

Venet,D. *et al.* (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.*, **7**, e1002240.

Von Minckwitz,G. *et al.* (2012) Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *J. Clin. Oncol.*, **30**, 1796–1804.