Immunocytochemical stem cell markers can predict clinical stage of breast cancer

PEDRO J. GUTIÉRREZ DIEZ[1], YANRONG SU[2] and JOSE RUSSO[2]

[1]Department of Economic Theory, University of Valladolid, School of Economics, Valladolid, Spain;
[2]The Irma H. Russo, MD - Breast Cancer Research Laboratory, Fox Chase Cancer Center, Temple University Health System, Philadelphia, PA, USA

Received October 13, 2016; Accepted April 3, 2017

DOI: 10.3892/or.2017.5820

Abstract. We present a computational-statistical algorithm that, from data on the staining degree of immunocytochemical markers: i) evaluates the ability of the considered immunopanel in predicting the breast cancer stage; ii) makes the accurate identification of breast cancer stage possible; iii) provides the best stage prognosis compatible with the considered sample; and iv) does so through the use of the minimum number of markers minimizing time and resource costs. After running the algorithm on two data sets [triple-negative breast cancer, (TNBC), and estrogen receptor-negative breast cancer, (ERNBC)], we conclude that EpCAM and β1 integrin are enough to accurately predict TNBC stage, being ALDH1, CD24, CD61, and CK5 the necessary markers to exactly predict ERNBC stage.

Introduction

In breast cancer, staging is determined by the combination of tumor size (T), lymph node status (N), and metastases (M). The stage of a cancer helps clinician to evaluate the status of the disease and makes the treatment decision. Therefore, any advance that contributes to determine the proper staging will have importance in applying the correct treatment and improving patient outcomes. The use of immunocytochemical markers is used in the diagnosis and prognosis of breast cancer; however, they have not yet been used for staging the disease. This is due to two reasons. On the one hand, the use of immunocytochemical markers is hampered by the lack of clear correlation between the marker staining degree and the disease stage. For each particular marker, its reaction to the progress of the disease is usually non-proportional and irregular because of the heterogeneity of the disease, being an obstacle for an accurate staging when using a unique immunocytochemical marker. On the other hand, the existence of complex and very different uneven behavior for the distinct markers prevents medical researchers from obtaining useful marker combinations to accurately and rapidly predict the disease stage: when the individual marker responses to the disease progress are diverse and erratic, it is not clear at all how to obtain information from the joint behavior of a set of markers in a manageable and operative way. As a consequence, although the use of immunopanels becomes mandatory to gain diagnosis and prognosis capability, only under appropriate mathematical and statistical analyses can the use of additional markers shed further light on the specific stage of the disease without the use of clinical and pathological data (1-30).

At the present time due to the aforementioned reasons, prognosis and diagnosis based on immunocytochemical markers are far from substituting clinical and pathological verification. However, it will be of great advantage to obtain a more exact and reliable use of these markers, since it would allow a rapid and costless monitoring, evaluation and verification of the disease staging, response to treatment, and origin of the cancer without the use of clinical and pathological data. In this respect, the objective of this study is to design an efficient method for assessing the ability of immunocytochemical markers, especially stemness markers, in predicting the specific stage of breast cancer. More specifically, we present a computational-statistical algorithm that, making use of data on the staining degree of immunocytochemical markers: i) evaluates the ability of the considered immunopanel in predicting the breast cancer stage; ii) makes the accurate identification of breast cancer stage possible; iii) provides the best stage prognosis compatible with the considered sample; and iv) does so through the use of the minimum number of markers, thus minimizing time and resource costs.

To illustrate the applicability of this algorithm we have run it for a set of 16 samples identified as TNBC and another set of 32 samples classified as ERNBC. The stages of these cancers ranged from stage I to IV, and were determined following the guidelines of American Cancer Society and the American

Correspondence to: Dr Pedro J. Gutiérrez Diez, Department of Economic Theory, University of Valladolid, School of Economics, Avda. Valle Esgueva 6, 47011 Valladolid, Spain
E-mail: pedrojos@fae.uva.es

Key words: triple-negative breast cancer, estrogen receptor-negative breast cancer, immunocytochemical marker, stemness, efficiency, algorithm, multinomial regression model for ordinal responses, prognosis

Joint Committee on Cancer (AJCC) TNM system (31,32). We have used the algorithm to test the predictive capability in breast cancer of ALDH1; CD24; CD44; CD49f; CD61; cytokeratin 5 (CK5); EpCAM; β1 integrin; and vimentin. These markers, suitable for paraffin embedded tissue *in situ* localization of breast cancer stem cells, are subject of recent and increasing research given the importance of breast cancer stem cells in the development of cancer and the complexity of the responses of these markers to the disease status (33-42). After running the algorithm, we conclude that EpCAM and β1 integrin are enough to accurately predict TNBC stage, being ALDH1, CD24, CD61, and CK5 the necessary markers to exactly predict ERNBC stage. From this perspective, our analyses are relevant not only for allowing total accuracy in identifying the specific breast cancer stage from a reduced number of markers, but also for contributing to the comprehension of the relationships between the distinct breast cancer stages and the considered immunocytochemical stemness markers.

Materials and methods

Immunohistochemistry (IHC). Formalin-fixed paraffin-embedded primary breast tumor samples were sectioned and evaluated after hematoxylin and eosin staining. Representative neoplastic areas were marked and selected to construct tissue microarray blocks. Tissue sections (4 μm) were used to do IHC staining following the standard protocol. Briefly, sections were deparaffinized and placed in Citra Plus antigen retrieval solution (BioGenex, Fremont, CA, USA; #HK081-20K) in microwave oven for 10 min at 100˚C. After cooling for 20 min, slides were placed in PBS for 5 min, and then the staining took place using i6000 Autostainer (BioGenex). The slides were quenched with Peroxide Block (BioGenex, #HK111) for 10 min, followed by blocking with Power Block (BioGenex, #HK085) for 20 min at room temperature. The sections were then incubated with primary antibody (information of antibodies is shown in Table I) for 30 min, super enhancer (BioGenex HK518) for 20 min, and second antibody (BioGenex, HK519) for 30 min. Washing three times using Super Sensitive Wash Buffer (BioGenex, #HK583) was performed after each step of incubation. DAB was used to develop the staining. Slides were scanned using Aperio Digital Pathology Scaner (Leica Biosystems Inc, Buffalo Grove, IL, USA).

Calibration of the antibodies for IHC. To run the statistical-computational algorithm it is necessary to obtain data on the staining intensity of each marker for each particular breast cancer stage, i.e. a calibration of the considered markers is necessary. This calibration was carried out separately for each type of examined breast cancer, namely TNBC and ERNBC.

The calibration process of the antibodies was as follows: For each type of breast cancer, we began by identifying the specific stage of each breast cancer. On the basis of standard criteria, we considered the following stages: I, IIa, IIb, IIIa, IIIb, and IV. For each stage, using the above-mentioned immunocytochemical markers of stemness, we identified the number of stem cells in primary breast cancer tissue embedded in paraffin. In particular, we examined 16 samples identified as triple-negative breast cancer and 32 samples classified as ER-negative. As explained before, the stem cells were quantified on the basis of the presence of the protein as indicated by staining. Each tumor cell was observed and its staining evaluated, and then classified by using the Cell Counter function of ImageJ software to label the cells. When possible, the stain was evaluated by quantifying the staining intensity on each cancer cell, scoring the cell from 0 to 3+, where 0 was for negative staining, 1+ for weak staining, 2+ for moderate staining, and 3+ for strong staining. This was done for ALDH1, CD24, CD61, EpCAM, and β1 integrin. For the remaining markers CD44, CD49f, CK5 and vimentin, each tumor cell was observed and classified and counted as either positive or negative. This process was repeated for all of the samples, and after quantifying the number of cells in each staining degree for each sample, the results were transferred to an Excel® spreadsheet. Through the use of Excel, a percentage breakdown was calculated for both positive and negative cells. This method was chosen in order to clearly show the amount of stem cell activity within each tumor. Additionally, if a tumor had multiple image cores to be quantified then the results from all of the tumor images were added together and the final percentages were calculated.

Statistical and computational issues. The statistical study of the association between the marker staining degrees and the different stages of breast cancer is the most relevant analysis from the clinical and pathological points of view. By capturing how the distinct markers respond to the advance of breast cancer, it is possible to relate the marker staining intensity with a particular stage of the disease and/or a particular type of breast cancer, opening up the possibility to infer the specific stage at which breast cancer is. In this respect and as commented on before, there is abundant empirical evidence showing that: i) the use of isolated markers is questionable in terms of reliability, the use of immunopanels becoming necessary (1); and ii) when using immunopanels, the selection of the appropriate statistical technique is crucial.

Once accepted the use of immunopanels, the problem is to select the most suitable statistical approach. On this point, since the breast cancer stage, the variable to be explained, is an ordinal variable, and the marker staining degrees, the explanatory variables, are of quantitative nature, the appropriate model to consider is a 'multinomial regression model for ordinal responses'. The intuitive idea behind this model is that the information on the specific stage of breast cancer is contained in the joint behavior of a certain number of markers: The relevant behavior to consider is not the response of an isolated marker, whatever this marker is, but the combined and joint reaction of a specific number of markers. The crucial questions are then: first, to identify the markers providing information; and, second, to specify how the selected marker staining degrees enter into the mathematical expression explaining the stage.

These are the tasks that our algorithm efficiently performs. In a first step, our procedure identifies the combinations of marker staining degrees that imply the desired degree of accuracy, in our case the maximum one with a perfect explanation of the sample. Among them, in a second step, the algorithm selects those combinations entailing the minimum number of

Table I. Antibody information.

Antibody name (clone)	Provider	Catalog#	Dilution
ALDH1 (44/ALDH)	BD Biosciences	611194	1:800
CD24	Abbiotec	251181	1:150
CD44 (DF1485)	BioGenex	AM310-5M	Ready-to-use
CD49f	Lifespan Bioscience	LS-A8769	1:200
CD61	BioGenex	AN482-5N	Ready-to-use
CK5 (EPR1600Y)	BioGenex	AN484-5M	Ready-to-use
EpCAM (E144)	BioGenex	AN489-5M	Ready-to-use
β1 integrin (EP1041Y)	Abcam	Ab529741	1:400
Vimentin (V9)	BioGenex	AM074-5M	Ready-to-use

Table II. Possible regressors.

Breast cancer stage	Regressors																		
	ALDH1			CD24			CD44	CD49f	CD61			CK5	EpCAM			β1 Integrin			Vimentin
	N	W	M	N	W	M	N	N	N	W	M	N	N	W	M	N	W	M	N

N, W and M: percentage of negative, weak and moderate stained cells, respectively.

markers. We refer the interested reader in these statistical and mathematical issues to (43-47).

Computational algorithm. The selection of the best set of markers is based on the following criteria: i) to imply the higher predictive capability; and ii) to involve the minimum number of markers. Obviously, these criteria seek to identify the set of antibodies that are the most reliable in predicting the disease stage and that, in addition, entail the lowest costs in terms of time and resources. To identify this set of antibodies, we wrote computational programs with the following steps: 1) From the set of the considered possible explanatory variables, to generate all the possible combinations of regressors (Table II). In our specific case, with 19 explanatory variables, there exist 524286 possible combinations. 2) For each possible combination of regressors, to run the multinomial regression model for ordinal responses. 3) From step 2, to compute the predictive capability for each possible combination of regressors. 4) From step 3, to select those combinations of regressors that imply the maximum predictive capability. 5) From step 4, among the set of regressors with the maximum predictive capability, to select those implying the minimum number of markers.

To gain computational efficiency, we wrote in Eviews the program to run all the statistical regressions, since this statistical package offers the best results and options for the multinomial regression model for ordinal responses. We programmed in Matlab the remaining steps. After applying this computational algorithm, the outcome is the set of markers ensuring the maximum feasible reliability in predicting the breast cancer stage and implying the minimum time and resources costs.

Results

We have run the algorithm for a set of 16 samples identified as triple-negative breast cancer and another set of 32 samples classified as estrogen receptor-negative breast cancer. The data for these samples are those in Tables III and IV, which collect the observed stages and the staining degree of the considered markers for triple-negative breast cancer and estrogen negative breast cancer, respectively.

Triple-negative breast cancer. After running the algorithm for the triple-negative breast cancer data, we obtained that, for the sample, the stages can be predicted with total accuracy with a minimum number of two markers. More specifically, the computational procedure generated a perfect explanation of the data when the explanatory variables were the negative, weak and moderate staining degrees measured for EpCAM and β1 integrin. Obviously, there exist many other combinations of regressors allowing observed data to be perfectly explained, but all of them entail a greater number of markers and/or regressors.

Since the coefficients for the multinomial regression model for ordinal responses have no clear interpretation and the model main purpose is prediction, we present here the algorithm results related to the predictive ability of the final estimation. The efficiently selected marker staining degrees are those in Table V, which also collects the value for the pseudo-R2

Table III. Triple-negative breast cancer data.

	ALDH1			CD24		CD44	CD49F	CD61		CK5	EpCAM			β1 Integrin			Vimentin
Stage	Negative	Weak	Moderate	Negative	Weak	Moderate	Negative	Negative	Negative	Moderate	Negative	Weak	Moderate	Negative	Weak	Moderate	Negative
SI	87.05%	2.91%	4.53%	0.00%	3.21%	95.41%	59.27%	52.43%	3.38%	73.09%	19.53%	38.40%	0.00%	0.83%	75.33%	12.01%	61.97%	25.89%	0.39%
SI	95.78%	4.22%	0.00%	0.00%	0.00%	98.08%	26.44%	11.31%	5.90%	31.20%	57.30%	23.89%	0.12%	0.66%	31.53%	0.00%	15.91%	81.12%	0.00%
SIIa	99.98%	0.02%	0.00%	0.00%	0.00%	100.00%	77.69%	45.25%	3.56%	37.63%	56.16%	88.09%	0.00%	19.52%	63.25%	0.36%	43.62%	54.53%	13.99%
SIIa	84.80%	0.00%	0.23%	0.04%	0.00%	98.97%	77.30%	19.30%	2.88%	45.97%	47.72%	95.37%	0.00%	0.08%	37.00%	14.11%	85.66%	0.23%	0.00%
SIIa	86.48%	0.00%	0.00%	0.00%	59.86%	39.39%	90.60%	53.28%	16.61%	77.69%	3.53%	96.78%	0.00%	0.00%	87.28%	15.79%	52.62%	30.66%	48.46%
SIIa	87.92%	2.08%	6.96%	0.00%	100.00%	0.00%	80.48%	96.52%	82.39%	4.34%	3.22%	99.14%	0.00%	0.36%	68.11%	97.11%	2.89%	0.00%	96.96%
SIIb	80.53%	0.31%	2.26%	5.55%	92.57%	1.88%	77.24%	84.88%	34.42%	44.07%	13.13%	96.92%	0.06%	1.20%	25.36%	85.82%	6.57%	7.40%	60.99%
SIIb	91.28%	0.00%	0.08%	0.00%	56.09%	43.91%	96.49%	68.50%	8.24%	82.19%	6.91%	98.38%	0.93%	1.62%	50.99%	12.38%	87.62%	0.00%	40.98%
SIIb				0.00%	0.00%	100.00%							97.92%				17.54%	82.46%	0.00%
SIIb	81.66%	0.00%	1.03%	7.16%	92.40%	0.44%	100.00%	66.96%	24.22%	52.46%	13.06%	98.33%	0.00%	0.79%	8.00%	48.49%	27.14%	24.36%	84.36%
SIII				4.54%	95.46%	0.00%	99.95%	100.00%	1.93%	89.87%	7.27%	99.43%	0.19%	0.06%	40.06%	100.00%	0.00%	0.00%	79.42%
SIII	95.72%	0.41%	0.79%	7.10%	91.88%	1.02%	82.51%	91.88%	21.74%	45.92%	29.17%	98.74%	0.00%	0.00%	35.19%	37.29%	61.33%	0.00%	25.71%
SIII	82.10%	0.48%	1.67%	2.86%	97.14%	0.00%	81.56%	32.16%	4.91%	86.53%	5.70%	99.95%	0.07%	0.30%	53.75%	99.74%	0.26%	0.00%	83.28%
SIV	95.08%	0.00%	4.92%	0.00%	44.22%	53.76%	84.51%	23.25%	0.68%	44.78%	45.03%	98.88%				4.58%	83.10%	11.27%	93.02%
SIV	83.92%	0.76%	0.00%	2.27%	73.94%	23.79%	100.00%	21.13%	2.68%	68.09%	21.16%	89.41%	4.32%	6.92%	31.12%	92.16%	6.01%	1.32%	91.79%
SIV	99.50%	0.00%	0.00%	0.00%	74.22%	23.40%	42.06%	14.79%				82.75%	1.06%	0.67%	39.40%	2.28%	38.99%	58.73%	14.20%

(goodness of fit), the likelihood-ratio statistic (LR statistic), and the p-value for the LR test. This LR test uses the LR statistic to compare the goodness of fit of two models, one of which is the null model, in this case the constant probability model, the second being our multinomial regression model for ordinal responses (the alternative model). As it appears in Table V, the null hypothesis of constant probabilities is rejected given that the p-value of the test is extremely low, and we can conclude that our proposed multinomial regression model for ordinal response is much more plausible.

Eviews provides an additional comparison between the constant probability specification and the estimated ordered dependent variable model. This comparison, based on the predictive capability of each model, appears in Table VI. The first column collects the (arbitrary) ordered values assigned to the dependent variable, i.e. to the different stages. For each stage, the second to sixth columns display, respectively, the number of observations, the number of correct and incorrect predictions, and the percentages of correct and incorrect predictions. These results are provided for the ordered model and for the best constant probability specification. Obviously, this comparison allows the significance of the markers in explaining the stages to be visualized: if we would remove the selected marker staining degrees as explanatory variables, we would be able to explain only a 28.571% of the observed data, versus the 100% that the inclusion of the considered marker staining degrees makes possible.

According to our results, it seems that the joint response of EpCAM and β1 integrin completely characterizes the disease stage for the triple-negative breast cancer, a characterization impossible to obtain when using these markers separately. Indeed, for the same sample, no correlation was observed between the expression of any of these two markers and the clinical stage.

When we eliminate the distinction between stages SIIa and SIIb and define a unique SII stage, the algorithm output is that in Tables VII-IX. As is logical, the results and conclusions are very similar, the most remarkable consideration being the disappearance of a level of staining for EpCAM as an explanatory variable. In other words, the introduction for EpCAM of the moderate or negative staining allows the separation between stages SIIa and SIIb to be established, but their consideration is not mandatory if this distinction is not necessary.

As appears in Tables VII and VIII, when the considered stages for TNBC were SI, SII, SIII and SIV, total accuracy in predicting the sample was achieved without considering a staining degree for EpCAM marker, respectively moderate (Table VII) or negative (Table VIII) staining. Table IX collects the predictive power results and the comparison with the constant probability model for both cases.

Estrogen receptor-negative breast cancer. For the estrogen receptor-negative breast cancer we count on a sample with 32 data that differentiate between stages SI, SIIa, SIIb, SIIIa, SIIIb and SIV. After running the algorithm for this sample, we obtained that the stages can be predicted with total accuracy with a minimum number of five markers. More specifically, the computational procedure generated a perfect explanation of the data when the explanatory variables were the staining degrees measured for ALDH1 (negative, weak and moderate),

Table IV. Estrogen receptor-negative breast cancer data.

	ALDH1			CD24			CD44	CD49F	CD61			CK5	EpCAM			β1 Integrin			Vimentin
Stage	Negative	Weak	Moderate	Negative	Weak	Moderate	Negative	Negative	Negative	Weak	Moderate	Negative	Negative	Weak	Moderate	Negative	Weak	Moderate	Negative
SI	87.05%	2.91%	4.53%	0.00%	3.21%	95.41%	59.27%	52.43%	3.38%	73.09%	19.53%	38.40%	0.00%	0.83%	75.33%	12.01%	61.97%	25.89%	0.39%
SI	95.78%	4.22%	0.00%	0.00%	0.00%	98.08%	26.44%	11.31%	5.90%	31.20%	57.30%	23.89%	0.12%	0.66%	31.53%	0.00%	15.91%	81.12%	0.00%
SIIa	99.98%	0.02%	0.00%	0.00%	0.00%	100.00%	77.69%	45.25%	3.56%	37.63%	56.16%	88.09%	0.00%	19.52%	63.25%	0.36%	43.62%	54.53%	13.99%
SIIa	76.01%	0.00%	5.38%	0.28%	98.51%	1.20%	97.61%	68.54%	15.77%	76.34%	5.48%	99.09%	0.20%	1.72%	55.06%	96.79%	0.00%	0.00%	90.45%
SIIa	84.80%	0.00%	0.23%	0.04%	0.00%	98.97%	77.30%	19.30%	2.88%	45.97%	47.72%	95.37%	0.00%	0.08%	37.00%	14.11%	85.66%	0.23%	0.00%
SIIa	86.48%	0.00%	0.00%	0.00%	59.86%	39.39%	90.60%	53.28%	16.61%	77.69%	3.53%	96.78%	0.00%	0.00%	87.28%	15.79%	52.62%	30.66%	48.46%
SIIa	71.16%	0.00%	0.00%	0.00%	1.83%	72.31%	78.51%	86.67%	30.14%	58.91%	6.91%	92.04%	0.00%	0.00%	11.98%	44.49%	24.96%	22.02%	30.48%
SIIa	87.92%	2.08%	6.96%	0.00%	100.00%	0.00%	80.48%	96.52%	82.39%	4.34%	3.22%	99.14%	0.00%	0.36%	68.11%	97.11%	2.89%	0.00%	96.96%
SIIb	80.53%	0.31%	2.26%	5.55%	92.57%	1.88%	77.24%	84.88%	34.42%	44.07%	13.13%	96.92%	0.06%	1.20%	25.36%	85.82%	6.57%	7.40%	60.99%
SIIb	83.97%	0.00%	0.00%	0.00%	0.00%	0.00%	54.81%	55.52%	22.18%	63.06%	9.77%	100.00%	0.00%	0.00%	0.72%	15.56%	7.14%	33.04%	3.92%
SIIb	91.28%	0.00%	0.08%	0.00%	56.09%	43.91%	96.49%	68.50%	8.24%	82.19%	6.91%	98.38%	0.93%	1.62%	50.99%	12.38%	87.62%	0.00%	40.98%
SIIb				0.00%	0.00%	100.00%						97.92%				17.54%	82.46%	0.00%	
SIIb	81.66%	0.00%	1.03%	7.16%	92.40%	0.44%	100.00%	66.96%	24.22%	52.46%	13.06%	98.33%	0.00%	0.79%	8.00%	48.49%	27.14%	24.36%	84.36%
SIIIa	5.87%	86.71%	4.88%	0.00%	43.62%	50.34%	99.90%	32.33%	0.54%	69.29%	26.94%	98.71%	0.00%	3.81%	96.19%	0.10%	42.52%	50.24%	93.60%
SIIIa	91.53%	0.00%	8.47%	0.00%	100.00%	0.00%	100.00%	25.17%				98.04%				0.00%	96.00%	1.00%	52.46%
SIIIa	89.80%	8.37%	1.63%	100.00%	0.00%	0.00%	100.00%	85.69%				99.77%	0.00%	1.09%	69.49%	92.96%	7.04%	0.00%	
SIIIa				4.54%	95.46%	0.00%	99.95%	100.00%	1.93%	89.87%	7.27%	99.43%	0.19%	0.06%	40.06%	100.00%	0.00%	0.00%	79.42%
SIIIa				100.00%	0.00%	0.00%	99.59%	0.00%	6.18%	78.85%	10.96%	99.31%	0.00%	0.00%	12.88%	11.61%	65.45%	9.58%	94.10%
SIIIa	79.07%	0.00%	11.63%				100.00%		15.19%	59.96%	14.99%	88.39%							35.62%
SIIIa	87.18%	0.00%	2.56%					8.28%				98.03%	0.00%	0.00%	96.58%	90.27%	7.73%	1.47%	
SIIIa	85.04%	1.47%	1.61%				56.60%		1.27%	91.61%	5.30%	98.64%							33.10%
SIIIb	26.13%	13.52%	21.05%	0.00%	1.20%	98.80%	81.16%	55.46%	5.95%	75.07%	13.71%	88.62%	0.00%	2.33%	62.07%	74.65%	21.69%	3.27%	21.38%
SIIIb	95.54%	0.28%	0.63%				99.25%		0.53%	46.39%	39.79%	100.00%							97.04%
SIIIb	95.72%	0.41%	0.79%	7.10%	91.88%	1.02%	82.51%	91.88%	21.74%	45.92%	29.17%	98.74%	0.00%	0.00%	35.19%	37.29%	61.33%	0.00%	25.71%
SIIIb	99.26%	0.12%	0.00%	0.00%	0.00%	100.00%	91.10%	13.19%	5.75%	65.38%	16.65%	85.10%	0.00%	0.00%	99.25%	85.25%	14.75%	0.00%	95.96%
SIIIb	82.10%	0.48%	1.67%	2.86%	97.14%	0.00%	81.56%	32.16%	4.91%	86.53%	5.70%	99.95%	0.07%	0.30%	53.75%	99.74%	0.26%	0.00%	83.28%
SIV	88.21%	0.00%	0.38%	0.72%	99.28%	0.00%		8.13%	1.22%	34.71%	50.66%	56.87%	1.08%	0.97%	6.16%	1.38%	0.36%	77.65%	10.45%
SIV	95.08%	0.00%	4.92%	0.00%	44.22%	53.76%	84.51%	23.25%	0.68%	44.78%	45.03%	98.88%				4.58%	83.10%	11.27%	93.02%
SIV	98.61%	0.23%	0.23%	0.00%	0.00%	100.00%	81.95%	13.64%				91.83%	0.00%	0.00%	37.70%				48.74%
SIV	83.92%	0.76%	0.00%	2.27%	73.94%	23.79%	100.00%	21.13%	2.68%	68.09%	21.16%	89.41%	4.32%	6.92%	31.12%	92.16%	6.01%	1.32%	91.79%
SIV	15.56%	14.15%	47.57%	0.14%	0.00%	99.86%	99.07%	1.12%	0.40%	80.41%	13.44%	95.86%	0.00%	0.02%	56.75%	84.63%	0.34%	0.00%	83.86%
SIV	99.50%	0.00%	0.00%	0.00%	74.22%	23.40%	42.06%	14.79%				82.75%	1.06%	0.67%	39.40%	2.28%	38.99%	58.73%	14.20%

Table V. Multinomial regression model for ordinal responses.

Model specification
Dependent variable: Stage (TNBC, stages SI, SIIa, SIIb, SIII and SIV)
Method: Maximum likelihood - Ordered probit
Sample (adjusted): 16 observations
Included observations: 14 after adjustments

Selected regressors
Negative, EpCAM
Weak, EpCAM
Moderate, EpCAM
Negative, β1 integrin
Weak, β1 integrin
Moderate, β1 integrin

Accuracy and model validity	
Pseudo R-squared	1.000000
LR statistic	44.07473
Prob (LR statistic)	0.000000

Table VI. Prediction evaluation.

Prediction evaluation for ordered specification					
Dep. Value	Obs.	Correct	Incorrect	Correct (%)	Incorrect (%)
Estimated equation					
0	2	2	0	100.000	0.000
1	4	4	0	100.000	0.000
2	3	3	0	100.000	0.000
3	3	3	0	100.000	0.000
4	2	2	0	100.000	0.000
Total	14	14	0	100.000	0.000
Constant probability specification					
0	2	0	2	0.000	100.000
1	4	4	0	100.000	0.000
2	3	0	3	0.000	100.000
3	3	0	3	0.000	100.000
4	2	0	2	0.000	100.000
Total	14	4	10	28.571	71.429

TNBC; stages SI, SIIa, SIIb, SIII and SIV.

CD24 (negative, weak and moderate), CD61 (negative, weak and moderate), CK5 (negative) and EpCAM (negative and moderate). As for triple-negative breast cancer, there exist many other combinations of regressors implying total accuracy, but all of them entail a greater number of markers and/or regressors.

Table X collects the selected regressors and the statistics on the model accuracy and validity. These regressors imply a total explanation of the sample and a gain of a 70% of accuracy with respect to the constant probability model, as presented in Table XI.

Table VII. Multinomial regression model for ordinal responses.

Model specification
Dependent variable: Stage (TNBC, stages SI, SII, SIII and SIV)
Method: Maximum likelihood - Ordered probit
Sample (adjusted): 16 observations
Included observations: 14 after adjustments

Selected regressors
Negative, EpCAM
Weak, EpCAM
Negative, β1 integrin
Weak, β1 integrin
Moderate, β1 integrin

Accuracy and model validity	
Pseudo R-squared	1.000000
LR statistic	34.51401
Prob (LR statistic)	0.000002

Table VIII. Multinomial regression model for ordinal responses.

Model specification
Dependent variable: Stage (TNBC, stages SI, SII, SIII and SIV)
Method: Maximum likelihood - Ordered probit
Sample (adjusted): 16 observations
Included observations: 14 after adjustments

Selected regressors
Weak, EpCAM
Moderate, EpCAM
Negative, β1 integrin
Weak, β1 integrin
Moderate, β1 integrin

Accuracy and model validity	
Pseudo R-squared	1.000000
LR statistic	34.51401
Prob (LR statistic)	0.000002

As for triple-negative breast cancer, it can be concluded that, for the considered sample of ERNBC, the five selected markers, namely ALDH1, CD24, CD61, CK5 and EpCAM, are enough to provide all the necessary information to identify the stage at which the estrogen receptor-negative breast cancer is with total accuracy.

If we remove sub-stages 'a' and 'b' for stages SII and SIII, the algorithm totally explain the data with 4 markers and 8 regressors: ALDH1 (negative, weak and moderate), CD24 (weak), CD61 (negative, weak and moderate), and CK5 (negative). The inclusion of EpCAM therefore adds additional

Table IX. Prediction evaluation.

Prediction evaluation for ordered specification					
Dep. Value	Obs.	Correct	Incorrect	Correct (%)	Incorrect (%)
Estimated equation					
0	2	2	0	100.000	0.000
1	7	7	0	100.000	0.000
2	3	3	0	100.000	0.000
3	2	2	0	100.000	0.000
Total	14	14	0	100.000	0.000
Constant probability specification					
0	2	0	2	0.000	100.000
1	7	7	0	100.000	0.000
2	3	0	3	0.000	100.000
3	2	0	2	0.000	100.000
Total	14	7	7	50.000	50.000

TNBC; stages SI, SII, SIII and SIV.

Table X. Multinomial regression model for ordinal responses.

Model specification
Dependent variable: Stage (ERNBC, stages SI, SIIa, SIIb, SIIIa, SIIIb and SIV)
Method: Maximum likelihood - Ordered probit
Sample (adjusted): 32 observations
Included observations: 20 after adjustments

Selected regressors
Negative, ALDH1
Weak, ALDH1
Moderate, ALDH1
Negative, CD24
Weak, CD24
Moderate, CD24
Negative, CD61
Weak, CD61
Moderate, CD61
Negative, CK5
Negative, EpCAM
Moderate, EPCAM

Accuracy and model validity	
Pseudo R-squared	1.000000
LR statistic	66.78321
Prob (LR statistic)	0.000000

information useful to distinguish between sub-stages 'a' and 'b', both for stages SII and SIII. The results for this model are those in Tables XII and XIII, which again support the advantages of considering for predictive purposes the 'multinomial regression model for ordinal responses' that we propose.

As explained above, the proposed algorithm deduces which is the combination of marker staining degrees that, for the considered sample, simultaneously implies the maximum accuracy and the minimum time and resources costs. The algorithm results also allow prediction to be carried out for new patients, this predictive capability being one of its most interesting applications. Once the selected marker staining degrees have been measured for the new patient, we can apply the results of our model to these new observed values and obtain good estimators of the probabilities for each stage. Specialized software running the multinomial regression model for ordinal responses (such as that used here, Eviews®), usually allows these probabilities to be easily calculated simply by introducing the observed staining values. On this point, it is worth noting again that although our sample has a small size, the algorithm can be updated and the new probabilities re-estimated as new evidence appears. As an interesting program of development, the authors open this algorithm to the scientific community to make possible the incorporation of new evidence on the relationship between markers and stages, and therefore, a more exact estimation of the probabilities of presence of the different stages of the disease.

Discussion

In this study, we present an efficient computational-statistical algorithm that assesses the ability of immunocytochemical markers in predicting the specific stage of breast cancer. More specifically, the procedure determines and identifies the minimum number of immunocytochemical markers necessary to determine the stage of the disease at any desired level of reliability without the use of clinical and pathological data, thus reducing time and resource costs. Moreover, by measuring the staining degrees of the selected markers for any new patient, the algorithm also allows the probability of presence of each stage to be determined for the considered patient. Finally, the algorithm can be continuously updated in terms of Bayesian inference by introducing new evidence, in a feedback process that leads to more accurate estimations of the probabilities for each stage as new data on the response of the markers to the disease stage are available.

To illustrate the capability of the proposed algorithm we have run the procedure on two data sets (ERN and TN breast cancer), fixing as desired level of accuracy the maximum one, i.e. requiring total accuracy in prediction. In this respect, when the objective is to perfectly explain the sample, the algorithm has calculated for both types of cancer the minimum number of markers to use, has identified these markers, and has provided the prognosis functions to consider.

Our results confirm a previous finding in the literature: in identifying the stages of breast cancer, the relevant question to consider is not the response of an isolated marker, whatever this marker is, but the combined and joint reaction of a specific number of markers. We also have found that the progression of each type of cancer is signaled by distinct and specific markers. More specifically: i) triple-negative breast cancer stages can be identified using only two markers, namely EpCAM and β1 integrin; ii) Identification of Estrogen

Table XI. Prediction evaluation.

Prediction evaluation for ordered specification					
Dep. Value	Obs.	Correct	Incorrect	Correct (%)	Incorrect (%)
Estimated equation					
0	2	2	0	100.000	0.000
1	6	6	0	100.000	0.000
2	4	4	0	100.000	0.000
3	1	1	0	100.000	0.000
4	4	4	0	100.000	0.000
5	3	3	0	100.000	0.000
Total	20	20	0	100.000	0.000
Constant probability specification					
0	2	0	2	0.000	100.000
1	6	6	0	100.000	0.000
2	4	0	4	0.000	100.000
3	1	0	1	0.000	100.000
4	4	0	4	0.000	100.000
5	3	0	3	0.000	100.000
Total	20	6	14	30.000	70.000

ERNBC; stages SI, SIIa, SIIb, SIIIa, SIIIb and SIV.

Table XII. Multinomial regression model for ordinal responses.

Model specification
Dependent variable: Stage (ERNBC, stages SI, SII, SIII and SIV)
Method: Maximum likelihood - Ordered probit
Sample (adjusted): 32 observations
Included observations: 21 after adjustments
Selected regressors
Negative, ALDH1
Weak, ALDH1
Moderate, ALDH1
Negative, CD24
Weak, CD24
Negative, CD61
Weak, CD61
Moderate, CD61
Negative, CK5
Accuracy and model validity
Pseudo R-squared 1.000000
LR statistic 51.86092
Prob (LR statistic) 0.000000

Receptor-negative Breast Cancer stages requires 4 markers: ALDH1, CD24, CD61, and CK5; and iii) For both types of cancer, EpCAM is a useful marker to identify sub-stages 'a' and 'b'. All our results can be verified by running the multinomial regression model we propose with the data provided in Tables III and IV.

Table XIII. Prediction evaluation.

Prediction evaluation for ordered specification					
Dep. Value	Obs.	Correct	Incorrect	Correct (%)	Incorrect (%)
Estimated equation					
0	2	2	0	100.000	0.000
1	10	10	0	100.000	0.000
2	5	5	0	100.000	0.000
3	4	4	0	100.000	0.000
Total	21	21	0	100.000	0.000
Constant probability specification					
0	2	0	2	0.000	100.000
1	10	10	0	100.000	0.000
2	5	0	5	0.000	100.000
3	4	0	4	0.000	100.000
Total	21	10	11	47.619	52.381

ERNBC; stages SI, SII, SIII and SIV.

From the clinical and pathological points of view, our results are of interest in a double sense. On the one hand, with the estimated coefficients, we can estimate the probability of presence of each stage for a new patient simply by measuring the staining degrees of the selected markers for the new patient, making possible a rapid reliable staging. In this respect, a file with an Excel sheet that calculates the estimated probabilities of each stage for any new patient, i.e., for any value of the identified marker staining degrees, and for the two types of breast cancer, can be provided under request. In this Excel sheet, after introducing the staining degrees of the selected markers for the new patient, we can obtain the probabilities assigned to each stage and the most likely stage according to the sample information, both for ERN and TN breast cancer. On the other hand and as explained in the preceding sections, these probabilities can be updated and re-estimated simply by applying our algorithm to the new evidence: as new verified data on the association between stages and marker staining degrees is available, our algorithm updates the selected markers and re-estimates the β and μ coefficients, allowing new and better estimations of the stage probabilities to be obtained. In this respect, the researchers and practitioners interested in applying this algorithm to alternative and/or complementary data sets and in obtaining updated parameters and probabilities can contact the authors to ask for results, routines and estimations.

Acknowledgements

Dr Pedro J. Gutiérrez gratefully acknowledges financial support from the Spanish Office of Economy and Competitiveness and European FEDER Funds, research project MTM2014-56022-C2-2-P.

References

1. Zaha DC: Significance of immunohistochemistry in breast cancer. World J Clin Oncol 5: 382-392, 2014.
2. Su Y, Gutiérrez Diez PJ, Santucci Pereira J, Russo IH and Russo J: In situ methods for identifying the stem cell of the normal and cancerous breast. In: Techniques and Methodological Approaches in Breast Cancer Research. Russo J and Russo IH (eds). Springer, New-York, pp151-182, 2014. doi: 10.1007/978-1-4939-0718-2_6.
3. Weigelt B, Peterse JL and van't Veer LJ: Breast cancer metastasis: Markers and models. Nat Rev Cancer 5: 591-602, 2005.
4. Allred DC, Harvey JM, Berardo M and Clark GM: Prognostic and predictive factors in breast cancer by immunohistochemical analysis. Mod Pathol 11: 155-168, 1998.
5. Wolff AC, Hammond MEH, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, *et al*; American Society of Clinical Oncology/College of American Pathologists: American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. Arch Pathol Lab Med 131: 18-43, 2007.
6. Kleer CG, Cao Q, Varambally S, Shen R, Ota I, Tomlins SA, Ghosh D, Sewalt RGAB, Otte AP, Hayes DF, *et al*: EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci USA 100: 11606-11611, 2003.
7. Ooka M, Sakita I, Fujiwara Y, Tamaki Y, Yamamoto H, Aihara T, Miyazaki M, Kadota M, Masuda N, Sugita Y, *et al*: Selection of mRNA markers for detection of lymph node micrometastases in breast cancer patients. Oncol Rep 7: 561-566, 2000.
8. Zhu L, Loo WT, Cheng CW and Chow LW: Possible predictive markers related to micro-metastasis in breast cancer patients. Oncol Rep 15: 1217-1223, 2006.
9. Jones C, Mackay A, Grigoriadis A, Cossu A, Reis-Filho JS, Fulford L, Dexter T, Davies S, Bulmer K, Ford E, *et al*: Expression profiling of purified normal human luminal and myoepithelial breast cells: Identification of novel prognostic markers for breast cancer. Cancer Res 64: 3037-3045, 2004.
10. Kristiansen G, Winzer KJ, Mayordomo E, Bellach J, Schlüns K, Denkert C, Dahl E, Pilarsky C, Altevogt P, Guski H, *et al*: CD24 expression is a new prognostic marker in breast cancer. Clin Cancer Res 9: 4906-4913, 2003.
11. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ and Clarke MF: Prospective identification of tumorigenic breast cancer cells. Proc Natl Acad Sci USA 100: 3983-3988, 2003.
12. Zhang M, Behbod F, Atkinson RL, Landis MD, Kittrell F, Edwards D, Medina D, Tsimelzon A, Hilsenbeck S, Green JE, *et al*: Identification of tumor-initiating cells in a p53-null mouse model of breast cancer. Cancer Res 68: 4674-4682, 2008.
13. Sleeman KE, Kendrick H, Ashworth A, Isacke CM and Smalley MJ: CD24 staining of mouse mammary gland cells defines luminal epithelial, myoepithelial/basal and non-epithelial cells. Breast Cancer Res 8: R7, 2006.
14. Perrone G, Gaeta LM, Zagami M, Nasorri F, Coppola R, Borzomati D, Bartolozzi F, Altomare V, Trodella L, Tonini G, *et al*: In situ identification of $CD44^{+}/CD24^{-}$ cancer cells in primary human breast carcinomas. PLoS One 7: e43110, 2012.
15. Lo PK, Kanojia D, Liu X, Singh UP, Berger FG, Wang Q and Chen H: CD49f and CD61 identify Her2/neu-induced mammary tumor-initiating cells that are potentially derived from luminal progenitors and maintained by the integrin-TGFβ signaling. Oncogene 31: 2614-2626, 2012.
16. Vaillant F, Asselin-Labat ML, Shackleton M, Forrest NC, Lindeman GJ and Visvader JE: The mammary progenitor marker CD61/beta3 integrin identifies cancer stem cells in mouse models of mammary tumorigenesis. Cancer Res 68: 7711-7717, 2008.
17. Clarke CL, Sandle J, Parry SC, Reis-Filho JS, O'Hare MJ and Lakhani SR: Cytokeratin 5/6 in normal human breast: Lack of evidence for a stem cell phenotype. J Pathol 204: 147-152, 2004.
18. Böcker W, Hungermann D, Weigel S, Tio J and Decker T: Immunohistochemistry in breast pathology: Differential diagnosis of epithelial breast lesions. Pathologe 30: 13-19, 2009 (In German).
19. Aguiar FN, Mendes HN, Cirqueira CS, Bacchi CE and Carvalho FM: Basal cytokeratin as a potential marker of low risk of invasion in ductal carcinoma in situ. Clinics (São Paulo) 68: 638-643, 2013.
20. Cimino A, Halushka M, Illei P, Wu X, Sukumar S and Argani P: Epithelial cell adhesion molecule (EpCAM) is overexpressed in breast cancer metastases. Breast Cancer Res Treat 123: 701-708, 2010.
21. Naylor MJ, Li N, Cheung J, Lowe ET, Lambert E, Marlow R, Wang P, Schatzmann F, Wintermantel T, Schüetz G, *et al*: Ablation of beta1 integrin in mammary epithelium reveals a key role for integrin in glandular morphogenesis and differentiation. J Cell Biol 171: 717-728, 2005.
22. dos Santos PB, Zanetti JS, Ribeiro-Silva A and Beltrão EI: Beta 1 integrin predicts survival in breast cancer: A clinicopathological and immunohistochemical study. Diagn Pathol 7: 104, 2012.
23. Yao ES, Zhang H, Chen YY, Lee B, Chew K, Moore D and Park C: Increased beta1 integrin is associated with decreased survival in invasive breast cancer. Cancer Res 67: 659-664, 2007.
24. Gonzalez MA, Pinder SE, Wencyk PM, Bell JA, Elston CW, Nicholson RI, Robertson JF, Blamey RW and Ellis IO: An immunohistochemical examination of the expression of E-cadherin, alpha- and beta/gamma-catenins, and alpha2- and beta1-integrins in invasive breast cancer. J Pathol 187: 523-529, 1999.
25. Ivaska J, Pallari HM, Nevo J and Eriksson JE: Novel functions of vimentin in cell adhesion, migration, and signaling. Exp Cell Res 313: 2050-2062, 2007.
26. Satelli A and Li S: Vimentin in cancer and its potential as a molecular target for cancer therapy. Cell Mol Life Sci 68: 3033-3046, 2011.
27. Domagala W, Lasota J, Bartkowiak J, Weber K and Osborn M: Vimentin is preferentially expressed in human breast carcinomas with low estrogen receptor and high Ki-67 growth fraction. Am J Pathol 136: 219-227, 1990.
28. Korsching E, Packeisen J, Liedtke C, Hungermann D, Wülfing P, van Diest PJ, Brandt B, Boecker W and Buerger H: The origin of vimentin expression in invasive breast cancer: Epithelial-mesenchymal transition, myoepithelial histogenesis or histogenesis from progenitor cells with bilinear differentiation potential? J Pathol 206: 451-457, 2005.
29. Hemalatha A, Suresh TN and Kumar ML: Expression of vimentin in breast carcinoma, its correlation with Ki67 and other histopathological parameters. Indian J Cancer 50: 189-194, 2013.
30. Kusinska RU, Kordek R, Pluciennik E, Bednarek AK, Piekarski JH and Potemski P: Does vimentin help to delineate the so-called 'basal type breast cancer'? J Exp Clin Cancer Res 28: 118, 2009.
31. Singletary SE, Allred C, Ashley P, Bassett LW, Berry D, Bland KI, Borgen PI, Clark GM, Edge SB, Hayes DF, *et al*: Staging system for breast cancer: Revisions for the 6th edition of the AJCC Cancer Staging Manual. Surg Clin North Am 83:803-19, 2003.
32. Singletary SE and Connolly JL: Breast cancer staging: working with the sixth edition of the AJCC Cancer Staging Manual. CA Cancer J Clin 56: 37-47, 2006.
33. Park SY, Lee HE, Li H, Shipitsin M, Gelman R and Polyak K: Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer. Clin Cancer Res 16: 876-887, 2010.
34. Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, Dutcher J, Brown M, Jacquemier J, Viens P, Kleer CG, Liu S, *et al*: ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. Cell Stem Cell 1: 555-567, 2007.
35. Hwang-Verslues WW, Kuo WH, Chang PH, Pan CC, Wang HH, Tsai ST, Jeng YM, Shew JY, Kung JT, Chen CH, *et al*: Multiple lineages of human breast cancer stem/progenitor cells identified by profiling with stem cell markers. PLoS One 4: e8377, 2009.
36. Lehmann C, Jobs G, Thomas M, Burtscher H and Kubbies M: Established breast cancer stem cell markers do not correlate with *in vivo* tumorigenicity of tumor-initiating cells. Int J Oncol 41: 1932-1942, 2012.
37. Shackleton M, Vaillant F, Simpson KJ, Stingl J, Smyth GK, Asselin-Labat ML, Wu L, Lindeman GJ and Visvader JE: Generation of a functional mammary gland from a single stem cell. Nature 439: 84-88, 2006.
38. Fillmore CM and Kuperwasser C: Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. Breast Cancer Res 10: R25, 2008.
39. Liao MJ, Zhang CC, Zhou B, Zimonjic DB, Mani SA, Kaba M, Gifford A, Reinhardt F, Popescu NC, Guo W, *et al*: Enrichment of a population of mammary gland cells that form mammospheres and have in vivo repopulating activity. Cancer Res 67: 8131-8138, 2007.

40. Stingl J, Eirew P, Ricketson I, Shackleton M, Vaillant F, Choi D, Li HI and Eaves CJ: Purification and unique properties of mammary epithelial stem cells. Nature 439: 993-997, 2006.
41. Cariati M, Naderi A, Brown JP, Smalley MJ, Pinder SE, Caldas C and Purushotham AD: Alpha-6 integrin is necessary for the tumourigenicity of a stem cell-like subpopulation within the MCF7 breast cancer cell line. Int J Cancer 122: 298-304, 2008.
42. Pommier SJ, Hernandez A, Han E, Massimino K, Muller P, Diggs B, Chamberlain E, Murphy J, Hansen J, Naik A, *et al*: Fresh surgical specimens yield breast stem/progenitor cells and reveal their oncogenic abnormalities. Ann Surg Oncol 19: 527-535, 2012.
43. Dobson AJ and Barnett AG: An Introduction to Generalized Linear Models, Third Edition. Chapman and Hall/CRC. Taylor & Francis Group, 2008.
44. Greene WH: Econometric Analysis. MacMillan, 1993.
45. Aitchison J and Silvey SD: The generalization of probit analysis to the case of multiple responses. Biometrika 44: 131-140, 1957.
46. McCullagh P and Nelder JA: Generalized Linear Models. Chapman & Hall, New York, 1990.
47. Long JS: Regression Models for Categorical and Limited Dependent Variables. Vol. 7. Sage Publications, 1997.