Generation of a HER2 Breast Cancer Gold-Standard Using Supervised Learning from Multiple Experts

Violeta Chang(✉)

Laboratory for Scientific Image Analysis SCIANLab, Anatomy and Developmental Biology Department, Faculty of Medicine, University of Chile, Av Independencia 1027, Block A, 2nd Floor, Independencia, Santiago, Chile
vchang@dcc.uchile.cl

Abstract. Breast cancer is one of the most common cancer in women around the world. For diagnosis, pathologists evaluate the expression of biomarkers such as HER2 protein using immunohistochemistry over tissue extracted by a biopsy. This assessment is performed through microscopic inspection, estimating intensity and integrity of the membrane cells's staining and scoring the sample as 0 (negative), 1+, 2+, or 3+ (positive): a subjective decision that depends on the interpretation of the pahologist.

This work is aimed to achieve consensus among opinions of pathologists in cases of HER2 breast cancer biopsies, using supervised learning methods based on multiple experts. The main goal is to generate a reliable public breast cancer gold-standard, to be used as training/testing dataset in future developments of machine learning methods for automatic HER2 overexpression assessment.

There were collected 30 breast cancer biopsies, with positive and negative diagnosis, where tumor regions were marked as regions-of-interest (ROIs). Magnification of 20× was used to crop non-overlapping rectangular sections according to a grid over the ROIs, leading a dataset with 1.250 images.

In order to collect the pathologists' opinions, an Android application was developed. The biopsy sections are presented in a random way, and for each image, the expert must assign a score (0, 1+, 2+, 3+). Currently, six referent Chilean breast cancer pathologists are working on the same set of samples.

Getting the pathologists' acceptance was a hard and time consuming task. Even more, obtaining the scoring of pathologists is a task that requires subtlety communication and time to manage their progress in the use of the application.

Keywords: Breast cancer · Intra-variability · Inter-variability
Expert opinion · Biopsy score consensus

Supported by FONDECYT 3160559.

D. Stoyanov et al. (Eds.): CVII-STENT 2018/LABELS 2018, LNCS 11043, pp. 45–54, 2018.
https://doi.org/10.1007/978-3-030-01364-6_6

1 Introduction

Breast cancer is one of the most common cancer in women around the world [19]. In Chilean women, 17% of cancer cases corresponds to breast cancer that constitutes the deadliest cancer for women in the country [31].

For cancer diagnosis purposes, the pathologists evaluate the expression of relevant biomarkers (e.g. HER2 protein) using immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) over cancer tissue extracted by a biopsy. IHC provides a measure of protein expression while FISH provides a measure of gene copy amplification [26]. Usually, HER2 overexpression assessment has been manually performed by means of a microscopic examination, estimating the intensity and integrity of the membrane cells' staining and scoring the sample as one of the four labels: 0, 1+, 2+, and 3+; where 0 and 1+ are negative, 2+ is equivocal, and 3+ is positive [34]. The label 2+ refers to a borderline case, which means that a confirmation analysis is required for the complete diagnosis. A common HER2 confirmation test is performed by means of FISH, that analyses gene amplification status and counts the HER2 gene copy number within the nuclei of tumor cells. Many studies have focused on the correlation of IHC and/or FISH for HER2 evaluation [2,29]. It is recommended to perform HER2 evaluation using IHC analysis to determine negative, equivocal, and positive specimens, and further evaluation of equivocal cases with FISH, according to the latest guidelines from the College of American Pathologists (CAP) and the American Society of Clinical Oncology (ASCO) [34].

In this sense, HER2 overexpression assessment is based on a subjective decision that depends on the experience and interpretation of the pathologist [1,11,21]. This non-objective decision could lead to different diagnosis reached by different pathologists (pathologist's inter-variability). Even more, there is evidence that the same HER2 sample evaluated by the same pathologists in different periods of time could lead to dissimilar diagnosis (pathologist's intra-variability). The variability among pathologists for cancer tissue samples is significantly high [15,17,18,22,30], which directly impacts therapeutic decisions, making the reproducibility of the HER2 overexpression assessment a difficult task. There is clearly a need for quantitative methods to improve the accuracy and reproducibility in the assessment of HER2 using IHC.

Additionally, there is a lack of pathologists that could conjugate their experience for homogeneus cancer diagnosis. Just as an example, one of the largest pathology anatomy laboratories in Chile, located in Santiago, performs more than 30,000 biopsies per year. However, there are very few specialists in the country: 1 per every 100,000 inhabitants. The vast majority of pathologists are concentrated in the capital of the country. However, as it was aforementioned, pathologists have an important role in cancer care, because their diagnoses usually serve to establish the oncological treatment plan. Obviously, the lack of specialists, also leads to a lack of standards in less specialized laboratories and a notorious difference of experience among the pathologists of different laboratories.

One way to have a reproducible and objective procedure for HER2 assessment is by means of an automatic classification method that discriminates among four scores given a digital biopsy [5,6,8,13,32]. However, despite decades of research on computer-assisted HER2 assessment [7,13,14,23], there are still no standard ways of comparing the results achieved with different methods. Published algorithms for classification of breast cancer biopsies are usually evaluated according to how well they correlate with expert-generated classifications, though it seems that each research group has its own dataset of images, whose scores are based on the subjective opinion of only one or two experts. The fact that there are non-public datasets makes direct comparison between competing algorithms a very difficult task.

Even more, knowing that a ground-truth represents the absolute truth for a certain application, one would like to have one for HER2 assessment. Unfortunately, for HER2 overexpression assessment, it is very complicated to count with a ground-truth because of the subjectivity of the task. The absence of a gold standard for HER2 assessment makes evaluation of new algorithms a challenging task. In this way, correlation of IHC with FISH was used compare experts versus automatic assessment of HER2 [12]. Using agreement analysis is a different approach to performance evaluation in the absence of ground truth. A valid alternative consists of asking many experts in the field for their opinion about specific cases to generate a gold-standard [17].

Motivated by this challenge, this research work is aimed to achieve consensus opinion of expert pathologists in cases of HER2 breast cancer biopsies, using supervised learning methods based on multiple experts and considering different levels of expertise of experts. The main goal of this research is to generate a realiable public breast cancer gold-standard, combining the pathologists' opinions and FISH results, to be used as training/testing dataset in future developments of machine learning methods for automatic HER2 overexpression assessment. Also, it is expected to evaluate intra- and inter- variability of the experts, using the same data generated by the manual score assignment process. To guarantee a reliable gold-standard, there is available the FISH result for all the biopsy samples, that must be used to evaluate the performance of the machine learning method for getting pathologists' consensus. This would be a very significant contribution to the scientific community, because at present there is no public gold-standard for HER2 overexpression assessment, so the existing methods cannot be properly evaluated and compared.

This paper is organized as follows. In Sect. 2 we review the research work in the area, justifying the need for a gold-standard for HER2 overexpression assessment. Section 3 is devoted to describing in detail the process for collecting the biopsy sections and opinions from experts, as well as to give an overview of the methods for combining opinion from experts. The final remarks and conclusions can be found in Sect. 4.

Table 1. Summary of previous non-public datasets for HER2 overexpression assessment.

Publication	Cases	Experts	Source
Lehr et al. [25]	40	1	Beth Israel Deaconess Medical Center, Boston, MA, USA
Camp et al. [9]	300	1	Department of Pathology School of Medicine, Yale University New Haven, CT, USA
Dobson et al. [13]	425	1	Beaumont Hospital Adelaide and Meath Hospital Dublin, Ireland
Laurinaviciene et al. [23]	195	1	Oncology Institute of Vilnius University, Lithuania
Brugmann et al. [7]	72	5	Institute of Pathology, Aalborg Hospital, Aarhus University Denmark

2 Related Work

The importance of having an image database containing ground-truth labelings has been well-demonstrated in many applications of computer vision: handwriting recognition [24], face recognition [33], indoor/outdoor scene classification [28] and mammal classification [16]. As said before, a ground-truth represents the absolute truth for a certain application that is not always available or costly. Unfortunately, for many applications, especially in biomedicine, it is impossible to have a ground-truth and a valid alternative consists of asking experts in the field for their opinion about specific cases, in order to generate a gold-standard [17]. The need for a gold-standard in biomedical applications has been demostrated in PAP-smear classification [20], human sperm segmentation [10], and sub-celullar structures classification [3,4], among others.

No gold-standards are publicly available for HER2 overexpression assessment. Instead, several research groups have independently gathered cancer breast biopsy images and run different sets of tests, with different performance measures. In Table 1, it is shown a list with several breast cancer biopsy datasets currently used in publications on automatic HER2 overexpression assessment. None of them is a public dataset.

3 Materials and Methods

3.1 Collection of Biomedical Samples

The dataset entailed 30 whole-slide-images (WSI) extracted from cases of invasive breast carcinomas. The Biobank of Tissues and Fluids of the University of

Chile managed the collection of HER2 stained slides obtained from the two main Chilean pathology laboratories: (1) Service of Pathological Anatomy from Clinical Hospital of the University of Chile, and (2) Service of Pathological Anatomy from Clinical Hospital of the Catholic University of Chile.

All the biopsies have known positive and negative histopathological diagnosis (equally distributed in categories: 0, 1+, 2+, and 3+). Each one of these samples was digitalized at SCIANLab, using a whole-slide imaging tissue scanner (Hamamatsu NanoZoomer). Over each digitalized biopsy sample, the tumor regions were marked by an expert pathologist as regions-of-interest (ROIs), see Fig. 1. There were considered between 3–4 ROIs in each sample.

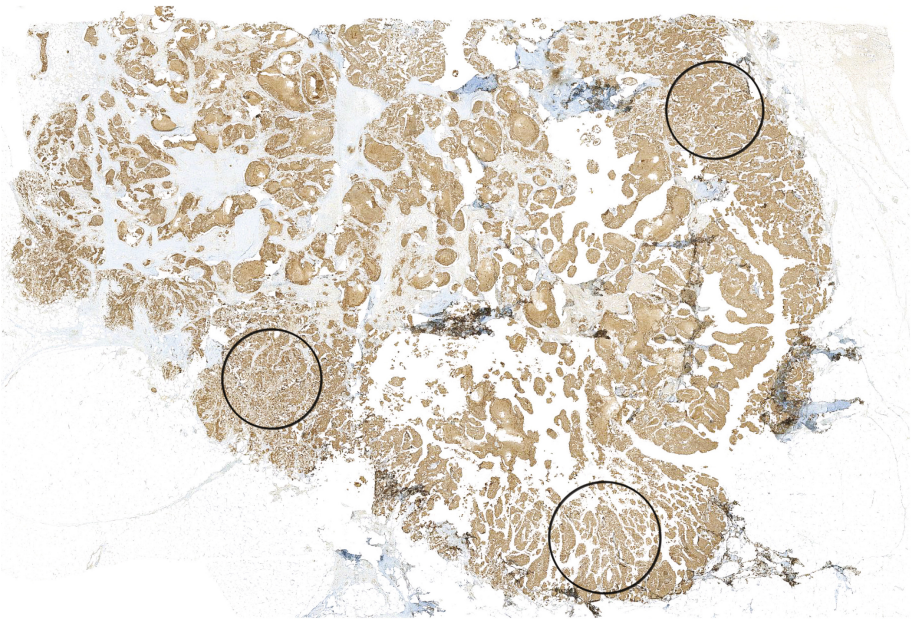

Fig. 1. Whole-slide-image, scanned using Hamamatsu NanoZoomer at SCIANLab, with the regions-of-interest (ROIs) marked on by an expert pathologist.

Then, to simulate real microscopic examination performed by pathologists and according to their opinion, magnification of 20× was used to crop non-overlapping rectangular sections according to a grid over the ROIs. A total of 1,250 biopsy sections were obtained. Aimed to evaluate intra-variability, each biopsy section was geometrically transformed (rotation, vertical flip, and horizontal flip). With all biopsy sections transformed two times, the complete dataset has 3,750 images.

All cases were subjected to supplemental FISH analysis, which is regarded as the gold-standard method by the ASCO/CAP guidelines [34]. This was done with the objective of guaranteeing a reliable gold-standard. Thus, available FISH

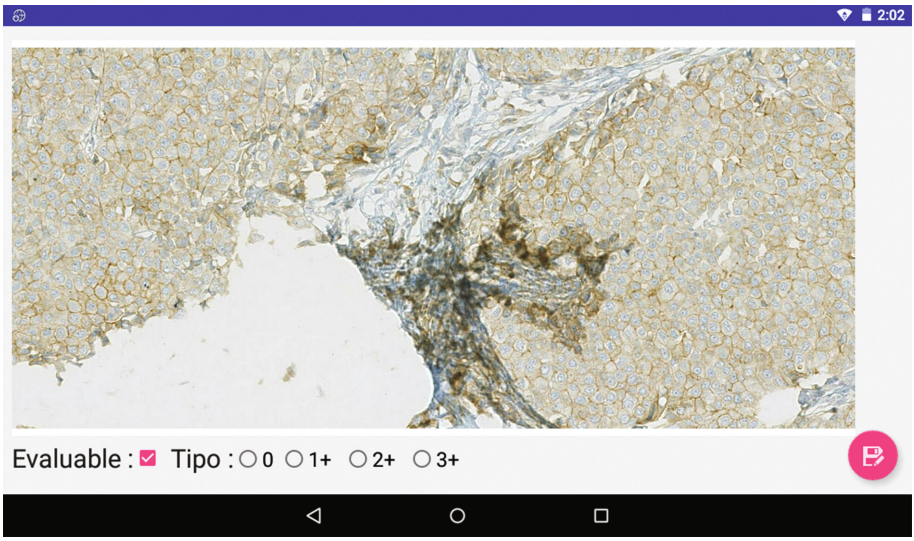

Fig. 2. Screen-shot of the dedicated Android application interface. This application will register the expert's opinion over the same image dataset, under the same conditions of visualization, allowing intra- and inter- variability analysis.

results must be used in two ways: (1) to generate a model along with expert's opinions, training the machine learning method to get results as good as FISH ones. In this way, a model to get consensus opinion could be generated without requiring FISH results, just expert's opinions, and (2) to evaluate the performance of the machine learning method for getting pathologists' consensus.

3.2 Collection of Expert's Opinions

In order to collect the expert pathologists' opinions, an Android application was specially designed and developed. It runs in a dedicated device (Tablet Acer Iconia One, 7-in IPS screen with 800×1280 pixels resolution, dual-core processor, 1GB of RAM). It is expected that each pathologist has the same device under the same conditions, to have a controlled scenario to evaluate inter-observer variability. The underlying idea is that the interface between the application and the pathologist is friendly, easy and intuitive to use and that the remote registration of the opinions of pathologists is carried out in an imperceptible way.

The biopsy sections are presented in a random way, and for each image, the expert must indicate whether the image is evaluable or not (according to his/her opinion) and must assign a score among 0, 1+, 2+, and 3+ (see Fig. 2). All the scores are registered locally in the device and remotely in a dedicated server, if an internet connection is available.

The ongoing work includes the compromise of six referent Chilean breast cancer pathologists, willing to participate in the study. Currently, all of them

have the same device with the same Android application installed on. So far, one pathologist have assigned score to 100% of the samples and two of them have assigned score to 40% of the samples.

3.3 Combination of Expert's Opinions

It is expected to count on the expert's opinion process finished to continue with the stage of combining those opinions. The idea beyond this consensus process is to use a supervised learning based on multiples experts that allows obtaining: (1) an estimated gold-standard that consensus labels assigned by experts, (2) classifier that considers multiple labels for each biopsy section, and (3) mathematical model of the experience of each expert based on the analyzed data and FISH results.

To evaluate the quality of the estimated gold-standard, area-under-curve (AUC) will be calculated using the estimated gold-standard versus labels according to the FISH results of each biopsy. To measure the reliability of the estimated gold-standard, AUC will be evaluated versus individual labels of each expert. In addition, different performance metrics will be measured for each expert regarding the estimated gold-standard: sensitivity, specificity, NPV (predictive value negative) and PPV (positive predictive value).

As an additional impact of this tudy, it is expected to assess the intra-expert variability. In this sense, it was considered during the Android application development to presenting the same biopsy sections to each pathologist in random order. In addition, presentation of the same sections contemplates a previous transformation of flipping and rotation of 90 degrees to increase the recognition complexity. The Kappa statistic will be used [27] to measure the degree of inter-expert and intra-expert variability, considering for this last case, each repetition of the manual classification process as a distinct entity.

4 Final Remarks

Getting the pathologists' acceptance was a hard and time consuming task. Even more, obtaining the scoring of pathologists is a task that requires a lot of subtlety and kind communication and time to manage their progress in the use of the application.

Considering the lack of specialists, it is very understandable how little free time they could have to participate in the study. However, there is a very good disposition and interest in collaborating in a study that will allow to standardize a very common practice in a pathological anatomy laboratory.

The methodology presented in this work is being applied to breast cancer biopsies. However, it would be easy extended/modified to be applied to different cancer tissues. Also, the developed Android application is extendable for other similar tasks and it showed robustness to work with many experts at the same time.

When this breast cancer gold-standar be publicly available, it would be a very significant contribution to the scientific community, because at present there is no public gold-standard for HER2 overexpression assessment, so the existing automatic methods cannot be properly evaluated and compared.

Finally, it is worth to remark that the techniques developed for automatic HER2 assessment will contribute to the valuable efforts in interpretation of biomarkers with IHC, increasing its reproducibility. However, the first step for generating confidence in their clinical utility is by means of a reliable gold-standard to evaluate their performance. The way of getting the confidence of pathologists to widespread the use of machine learning methods for clinical decisions in this field is to generate ways to use the opinion of a diversity of experts as the base of knowledge for automatic methods, tackling with all kinds of bias and known subjectivity.

Acknowledgements. Violeta Chang thanks pathologists M.D. Fernando Gabler, M.D. Valeria Cornejo, M.D. Leonor Moyano, M.D. Ivan Gallegos, M.D. Gonzalo De Toro and M.D. Claudia Ramis for their willing collaboration in the manual scoring of breast cancer biopsy sections. The author thanks Jimena Lopez for support with cancer tissue digitalization and the Biobank of Tissues and Fluids of the University of Chile for support with the collection of cancer biopsies. This research is funded by FONDECYT 3160559.

References

1. Akbar, S., Jordan, L., Purdie, C., Thompson, A., McKenna, S.: Comparing computer-generated and pathologist-generated tumour segmentations for immunohistochemical scoring of breast tissue microarrays. Br. J. Cancer **113**(7), 1075–1080 (2015)
2. Barlett, J., Mallon, E., Cooke, T.: The clinical evaluation of her-2 status: which test to use. J. Pathol. **199**(4), 411–417 (2003)
3. Boland, M., Markey, M., Murphy, R.: Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry **33**(3), 366–375 (1998)
4. Boland, M., Murphy, R.: A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. Bioinformatics **17**(12), 1213–1223 (2001)
5. Braunschweig, T., Chung, J.-Y., Hewitt, S.: Perspectives in tissue microarrays. Comb. Chem. High Throughput Screen. **7**(6), 575–585 (2004)
6. Braunschweig, T., Chung, J.-Y., Hewitt, S.: Tissue microarrays: Bridging the gap between research and the clinic. Expert. Rev. Proteomics **2**(3), 325–336 (2005)
7. Brugmann, A., et al.: Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. Breast Cancer Res. Treat. **132**(1), 41–49 (2012)
8. Camp, R., Chung, G., Rimm, D.: Automated subcellural localization and quantification of protein expression in tissue microarrays. Nat. Med. **8**(11), 1323–1327 (2002)
9. Camp, R., Dolled-Filhart, M., King, B., Rimm, D.: Quantitative analysis of breast cancer tissue microarrays shows that both high and normal levels of HER2 expression are associated with poor outcome. Cancer Res. **63**(7), 1445–1448 (2003)

10. Chang, V., et al.: Gold-standard and improved framework for sperm head segmentation. Comput. Methods Programs Biomed. **117**(2), 225–237 (2014)
11. Chen, R., Jing, Y., Jackson, H.: Identifying Metastases in Sentinel Lymph Nodes with Deep Convolutional Neural Networks arXiv:1608.01658 (2016)
12. Ciampa, A., et al.: HER-2 status in breast cancer correlation of gene amplification by fish with immunohistochemistry expression using advanced cellular imaging system. Appl. Immunohistochem. Mol. Morphol. **14**(2), 132–137 (2006)
13. Dobson, L., et al.: Image analysis as an adjunct to manual HER-2 immunohistochemical review: a diagnostic tool to standardize interpretation. Histopathology **57**(1), 27–38 (2010)
14. Ellis, C., Dyson, M., Stephenson, T., Maltby, E.: HER2 amplification status in breast cancer: a comparison between immunohistochemical staining and fluorescence in situ hybridisation using manual and automated quantitative image analysis scoring techniques. J. Clin. Pathol. **58**(7), 710–714 (2005)
15. Feng, S., et al.: A framework for evaluating diagnostic discordance in pathology discovered during research studies. Arch. Pathol. Lab. Med. **138**(7), 955–961 (2014)
16. Fink, M., Ullman, S.: From aardvark to zorro: a benchmark for mammal image classification. Int. J. Comput. Vis. **77**(1–3), 143–156 (2008)
17. Fuchs, T., Buhmann, J.: Computational pathology: challenges and promises for tissue analysis. Comput. Med. Imaging Graph. **35**(7–8), 515–530 (2011)
18. Gomes, D., Porto, S., Balabram, D., Gobbi, H.: Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast. Diagn. Pathol. **9**, 121 (2014)
19. Gurcan, M., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. **2**, 147–171 (2009)
20. Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B.: PAP-smear benchmark data for pattern classification. In: Proceedings of Nature inspired Smart Information Systems (NiSIS 2005), pp. 1–9 (2005)
21. Khan, A., et al.: A novel system for scoring of hormone receptors in breast cancer histopathology slides. In: 2nd IEEE Middle East Conference on Biomedical Engineering, pp. 155–158 (2014)
22. Lacroix-Triki, M., et al.: High inter-observer agreement in immunohistochemical evaluation of HER-2/neu expression in breast cancer: a multicentre GEFPICS study. Eur. J. Cancer **42**(17), 2946–2953 (2006)
23. Laurinaviciene, A., Dasevicius, D., Ostapenko, V., Jarmalaite, S., Lazutka, J., Laurinavicius, A.: Membrane connectivity estimated by digital image analysis of HER2 immunohistochemistry is concordant with visual scoring and fluorescence in situ hybridization results: algorithm evaluation on breast cancer tissue microarrays. Diagn. Pathol. **6**(1), 87–96 (2011)
24. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
25. Lehr, H., Jacobs, T., Yaziji, H., Schnitt, S., Gown, A.: Quantitative evaluation of HER-2/NEU status in breast cancer by fluorescence in situ hybridization and by immunohistochemistry with image analysis. Am. J. Clin. Pathol. **115**(6), 814–822 (2001)
26. Masmoudi, H., Hewitt, S., Petrick, N., Myers, K., Gavrielides, M.: Automated quantitative assessment of HER-2/NEU immunohistochemical expression in breast cancer. IEEE Trans. Med. Imaging **28**(6), 916–925 (2009)

27. McHugh, M.: Interrater reliability: the kappa statistic. Biochem. Med. **22**(3), 276–282 (2012)
28. Payne, A., Singh, S.: A benchmark for indoor/outdoor scene classification. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005, Part II. LNCS, vol. 3687, pp. 711–718. Springer, Heidelberg (2005). https://doi.org/10.1007/11552499_78
29. Prati, R., Apple, S., He, J., Gornbein, J., Chang, H.: Histopathologic characteristics predicting HER-2/NEU amplification in breast cancer. Breast J. **11**(1), 433–439 (2005)
30. Press, M., et al.: Diagnostic evaluation of HER-2 as a molecular target: an assessment of accuracy and reproducibility of laboratory testing in large, prospective, randomized clinical trials. Clin. Cancer Res. **11**(18), 6598–6607 (2005)
31. Prieto M.: Epidemiología del cáncer de mama en Chile. Revista Médica Clínica Las Condes (2011)
32. Seidal, T., Balaton, A., Battifora, H.: Interpretation and quantification of immunostains. Am. J. Surg. Pathol. **25**(1), 1204–1207 (2001)
33. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. IEEE Trans. Pattern Anal. Mach. Intell. **25**(12), 1615–1618 (2003)
34. Wolff, A., et al.: American society of clinical oncology, and college of american pathologists: recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J. Clin. Oncol. **31**(31), 3997–4013 (2013)