MEDICAL SCIENCE TECHNOLOGY

RESEARCH PAPER

ISSN 1895-2089
© Med Sci Tech, 2017; 58: 111-118
DOI: 10.12659/MST.905935

Received: 2017.06.26
Accepted: 2017.08.16
Published: 2017.10.03

Appropriate Sample Size for Standardization Parameters Estimation Reduces Misdiagnoses of Molecular-Based Risk Predictors in Breast Cancer

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

CDEF 1 **Aldana González-Montoro**
DEF 2 **Laura Prato**
CDE 3 **Federico Casares**
C 4 **Mónica Balzarini**
ABCDEFG 5 **Elmer Andrés Fernández**

1 Faculty of Mathematics, Astronomy, Physics and Computation (FAMAF), National University of Córdoba, Córdoba, Argentina
2 National University of Villa María, Córdoba, Argentina
3 LISRA Institute, East Rockaway, NY, U.S.A.
4 Faculty of Agronomy, Department of Biometry, The National Council of Scientific and Technical Research (CONICET), Catholic University of Córdoba, Córdoba, Argentina
5 Center for Research and Development of Emotional Intelligence (CIDIE), The National Council of Scientific and Technical Research (CONICET), Catholic University of Córdoba, Córdoba, Argentina

Corresponding Author: Elmer Andrés Fernández, e-mail: efernandez@bdmg.com.ar
Source of support: This work was supported by grants from the following Argentine institutions: Universidad Católica de Córdoba (BOD/2016 to EAF), Ministerio de Ciencia, Tecnología e Innovación Productiva (PPL 6/2011 to EAF), Secretaría de Ciencia y Tecnología – Universidad Nacional de Córdoba (30720150101719CB to EAF), and the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Background: Accurate risk/outcome prediction, in which molecular signatures (MS) are playing an increasingly important role, is crucial for personalized therapy. Patients require an accurate diagnosis and an appropriate therapy assignment as soon as they arrive at the clinic. However, most MS require gene-based standardization through parameters estimated from an available population sample. Thus, the estimation of gene standardization parameters (SP) turns out to be crucial to avoid misdiagnoses. Although dependency on SP has been recognized, the effect of different sample sizes on estimation of and impact on therapy management has not been reported. Because this is key for clinical application, in the present study we evaluated the impact of SP on outcome prediction error due to sample size. For this, 2 well-known breast cancer (BC) subtype/risk predictors were used on real data under different recruitment scenarios.

Material/Methods: The PAM50 and Gene70 MS were fed with standardized gene expression profiles using SP estimated from different sample sizes to predict BC intrinsic subtypes and progression, respectively. Error sensitivity analysis was based on estimation of outcome prediction error rates against those obtained using SP estimated with all the patients in the cohort (our criterion standard). Seven BC cohorts including TCGA data (2014 subjects in total) were used.

Results: We found that BC outcome prediction is very sensitive to the sample size used to estimate the MS standardization parameters. More than 20% of predicted classes can change when using small sample sizes to compute SP, and more than 20% of subjects can have their predicted outcome changed.

Conclusions: Patients might receive inappropriate therapy if the SP are not carefully dealt with. A pilot study to provide SP that yield a stable prediction is necessary. A method to evaluate the sufficiency of the size of the available sample for parameter estimation is proposed to guide prior pilot study development.

MeSH Keywords: **Decision Support Techniques • Early Detection of Cancer • Gene Expression Profiling • Transcriptome**

Full-text PDF: https://medscitechnol.com/abstract/index/idArt/905935

3625 1 5 25

Indexed in: [PBL] [Chemical Abstracts/CAS] [SCOPUS/EMBASE]

Background

In the last 15 years, several gene-expression-based classifiers, known as molecular signatures, have been developed for cancer stratification with the aim to pursue personalized therapy. These are multi-gene predictors of some outcome that were reported to be potentially useful in predicting prognosis and guiding therapy and in the evaluation of preoperative chemotherapy response [1]. In breast cancer, since the definition of the intrinsic molecular BC subtypes by Perou et al. [2], increased attention has been paid to the design and evaluation of these multi-gene assays of class predictors for several different targets [3–7]. Moreover, several studies have been conducted to validate the intrinsic BC subtypes (Basal-like, HER2-Enriched, Luminal A, Luminal B, and Normal-like) in many existing datasets [7–11], as well as to provide statistical sense to their prediction [12]. Nevertheless, there is still much controversy on their clinical utility, which could be based on the recognized heterogeneity of the BC disease with both clinical and molecular representation, as well as the methodology used to derive these MS. In this context, both similar performance and a strong disagreement among the MS have been reported [8–17].

In a daily clinical setting, a single sample predictor (SSP) should be used, which is an algorithm capable of immediately classifying a subject using SP previously obtained from a training dataset. New cases should be standardized and immediately diagnosed, and this standardization should not depend on other cases [8,15]. Therefore, as mentioned by Sorlie et al. [16], gene-centering procedures are essential before classification since new data should be adjusted to resemble the characteristics of the original dataset. Nevertheless, a broad spectrum of different standardization methodologies are applied, ranging from no gene correction at all [10] to full gene homogenization through the distance-weighted discrimination method [7]. However, the latter is not clinically useful because either not all of the subjects are simultaneously recruited or the method is applied over different populations. On the other hand, performing no standardization at all is also inappropriate in this kind of MS for both research and clinical applications since any classifier is built upon the basic assumption that the training set is a representative sample of the working population that holds all distributional characteristics.

A robust classification rule remains elusive at present and none of the available prediction models provide the standardization parameters. Thus, SP need to be estimated from the actual "working population" in which the prediction models will be used to guide therapy. This is a very important issue since, as reported by Lusa et al. [8], one condition to be met by these diagnostic tools is that they must unambiguously classify a new sample into a specific subtype or risk score, independent of any other samples being considered for classification at that time. To meet this requirement, some issues remain unresolved, such as the number of patients that must be recruited in order to obtain an SP estimate that would lead to an unambiguous classification. This step is critical since, in a clinical environment, subjects arrive one at the time, and they should be diagnosed as soon as possible through on-demand gene expression measures (e.g., through microarray technology or qt-PCR).

Most of the evaluations and comparisons reported in the literature for MS address concordance, prediction strength, and accuracy, with samples being normalized by means of whole-population-based estimates, an approach unlikely to be used in a clinical setting. In this paper, we evaluate the effect of sample size on SP estimation and its impact on outcome prediction. The robustness of the class membership prediction is evaluated using different sample sizes to estimate MS SP. We show that the MS classifiers that use standardizations can provide ambiguous classification results when SP are estimated with an inappropriate sample size. Thus, a patient could be, for example, mistakenly assigned to a good outcome class or to a subtype with evidence of no chemotherapy response, with important consequences for therapeutic behavior, as implied by the recommendations proposed by international committees [18]. Using whole-dataset SP estimates as the criterion standard, we calculated the outcome prediction error as a function of the sample size used to estimate the SP. Finally, we suggest a method to evaluate whether the current number of recruited subjects provides a stable and consistent prediction.

Material and Methods

To evaluate the effect of sample size on MS SP estimation and its impact on outcome prediction, 2 well-known prediction models were used, the PAM50 algorithm [15] and the Gene70 MS [3]; these algorithms are commercially available at Prosigna(r) (Nanostring Technologies, Seattle, USA) and MammaPrint(r) (Agendia, Amsterdam, The Netherlands), respectively. Different sample sizes were simulated by random sampling from these datasets. SP were estimated from each artificial dataset and MS was used to predict outcome over 6 freely available BC cohorts. The algorithms presented in the *genefu* R library [4] from *Bioconductor* website (*www.bioconductor.org*) were used.

The PAM50 single subject predictor

The PAM50 classification algorithm is fed with the subject expression level of 50 genes. Spearman correlation coefficients are calculated against the centroids of each defined subtype: Basal-like, HER2-enriched, Luminal A, Luminal B, and normal groups. Then, the subject is assigned to one of the subtypes according to the maximum achieved Spearman correlation. Since PAM50 algorithm was designed based on deviation from centered

genes, the gene set should be preprocessed before calculating correlation by centering. We use the *genefu* [4] implementation of the PAM50 algorithm (*intrinsic.cluster.predict* function).

The Gene70 prognosis signature (70 GS)

The Gene70 prognosis signature [3] is implemented in *genefu* through the *gene70* function. It provides low- and high-risk scores related to metastasis development. The risk score is also based on Spearman correlation. In *genefu* implementation, before the risk score calculation, data may be standardized to have zero mean and unit standard deviation (SD), known as z-score scaling. Then, the subject is classified as low- or high-risk if (risk score) <–0.3 or >–0.3, respectively.

Prediction and standardization parameters

As already stated, to predict the subtype of a subject, its expression vector has to be standardized. Let us denote the expression matrix of K genes across n subjects of a sample by $G=(g_{kj})$ with $k=1,...,K$ and $j=1,...,n$. The K rows of the matrix G contain the expression of each of the K genes across the sample, and we will denote the vectors for the k-th gene by g_k. The columns correspond to the expression vector for each subject and for simplicity we will denote the vector for subject j by x_j.

The first step of the procedure is to estimate the distribution parameters that need to be used for the standardization. For this, we compute the corresponding sample parameters. We will denote the vector of parameters for gene g_k by θ_k^n, where θ_k^n, is a two-dimensional vector containing a position and a dispersion parameter, by $\theta_k^n=(\bar{g}_k^n, s_k^n)$. In the case of PAM50, $\bar{g}_k^n$ corresponds to the median of g_k over the considered sample of n subjects and $s_k^n=1$, whereas for Gene70, $\bar{g}_k^n$ is the mean value of g_k over the considered sample and s_k^n is its standard deviation. The standardization is performed by each model, using the usual standardization function, obtaining the standardized expression value of gene k for subject j by, $g'_{kj}=\frac{g_{kj}-\bar{g}_k^n}{s_k^n}$.

To predict the subtype of a certain sample, the models need the parameters for all the genes, so let Θ^n denote the matrix for which k-th's row corresponds to θ_k^n.

Once the SP are computed, the model is used to predict the subjects subtype. We will denote this by $PM_j^n=PM(x_j, \Theta^n)$ the predicted class of the j-th subject, using SP Θ^n. Whether we are considering the PAM50 or GENE70 model will be clear from the text.

Datasets

The datasets (Table 1) used in this study are freely available from the Bioconductor website (*www.bioconductor.org*). In total, they involve 2014 BC patients.

Table 1. Datasets used.

Bioconductor bame	N	Technology	Ref
BreastCancerNKI	337	Agilent	3
BreastCancerVDX	334	Affymetrix	5
BreastCancerUPP	251	Affymetrix	6
BreastCancerTRANSBIG	200	Affymetrix	17
BreastCancerUNT	137	Affymetrix	19
BreastCancerMAINZ	198	Affymetrix	20
TCGA	466	Agilent	23

None of the datasets were modified (no further array-based preprocessing step was applied). Genes with multiple probes were averaged to avoid an extra source of variability in the results.

The criterion standard to validate classification

As the real distributional parameters are unknown, we will compare the SP estimated with different sample sizes with the SP estimated using all the data. This is, if N is the size of the cohort (dataset) being used, the criterion standard is defined as the classification achieved using Θ^N.

Simulating several sample size scenarios

We now describe the simulation procedure for each dataset. Consider $n=10,20,...,\left[\frac{0.9}{10}N\right]\cdot 10$. Where $[x]$ denotes the integer part of x. For each n, the procedure is the following. We

1. randomly chose 10% of the total N samples to be a test set. Let this number be N_{test}.
2. Randomly chose n subjects from the remaining 90% to use as training set.
3. Computed Θ^n with the training set.
4. Predicted the subtype for each subject in the test set obtaining PM_j^n, for $j=1,..., N_{test}$.
5. Repeated 1–4 M times to obtain $PM_{m,j}^n$ with $m=1,..., M$ and $j=1,..., N_{test}$.

These predictions were compared with the criterion standard ones made using Θ^N, as described in the following subsection.

Statistical analysis

Prediction models are usually evaluated in terms of ROC curves and or confusion matrices [19]. Here, for each prediction model, we calculate Average Percentage Prediction Errors (APPE) for each Θ^n as

$$APPE_n = 100 \cdot \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{N_{test}} I\left(PM_j^N \neq PM_{m,j}^n\right),$$

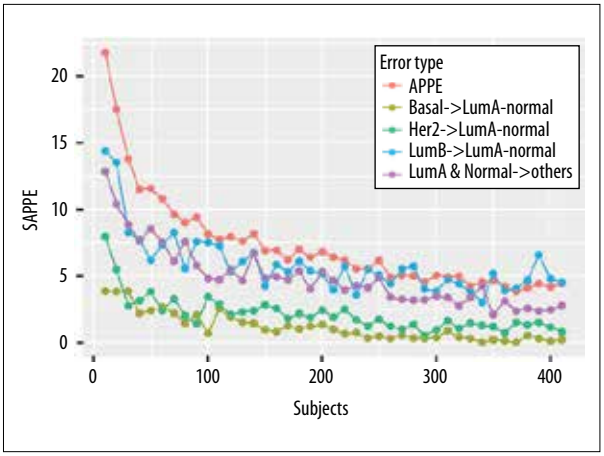


Figure 1. PAM50 Average Percentage Predictive Errors (APPE) regarding all the prediction errors independent of the subtype, and Severity Average Prediction Errors (SAPPE) related to subtype prediction errors between good (bad) prognosis prediction with Θ^N versus bad (good) prognosis prediction achieved with Θ^n.

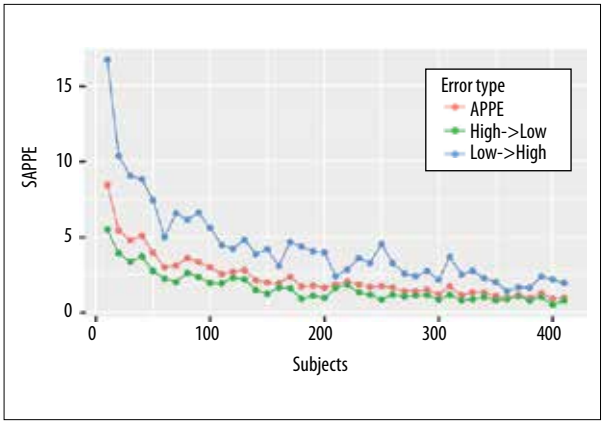


Figure 2. Gene70 Average Percentage Predictive Errors (APPE) regarding all the prediction errors independent of the subtype, and Severity Average Prediction Errors (SAPPE) related to subtype prediction errors between good (bad) prognosis prediction with Θ^{Nn} versus bad (good) prognosis prediction achieved with Θ^n.

where $I(PM_j^N \neq PM_j^n)$ is an indicator function yielding 1, if $PM_j^N \neq PM_j^n$ and 0 otherwise. In addition, taking into account that predicted subtypes by PAM50 and risk prognosis by Gene70 have very different implications for patient outcome, we also define a particular APPE called Severity Average Percentage Predictive Errors (SAPPE) as follows. In the case of PAM50, it is well-known that the intrinsic subtype Luminal A presents a better outcome compared to the other subtypes (Basal, Her2, or Luminal B) [12]. Therefore, we evaluate the APPE when predicted as a good (bad) subtype (i.e., Luminal A or normal) with Θ^N, and to a bad (good) subtype (i.e., Basal, Her2, or Luminal B) when using Θ^n. Thus

$$SAPPE_n^{X1,X2} = 100 \cdot \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N_{test}} I\left(PM_j^N = X1, PM_{m.j}^n = X2\right),$$

where $I(PM_j^N = X1, PM_j^n = X2)$ yields 1 only if $PM_j^N = X1$ and $PM_j^n = X2$.

When the PAM50 model is used, the following pairs were evaluated:

i) Basal to Luminal A or Normal-like: {X1="Basal", X2="Luminal A" or "Normal-like"}
ii) Her2 to Luminal A or Normal-like: {X1="Her2", X2="Luminal A" or "Normal-like"}
iii) Luminal B to Luminal A: {X1="Luminal B", X2="Luminal A" or "Normal-like"}
iv) Luminal A or Normal-like to Basal or Her2 or Luminal B: {X1="Luminal A" or "Normal-like", X2="Basal or Her2" or "Luminal B"}.

We can define SAPPE for the GENE70 model in an analogue manner, using the following pairs:

v) High to Low risk: {X1="High", X2="Low"}
vi) Low to High risk: {X1="Low", X2="High"}

Results

Figures 1 and 2 show APPE and SAPPE errors for the TCGA dataset over PAM50 and Gene70, respectively.

In Figure 1, we can observe that APPE for PAM50 can be up to 20% when using Θ^n with n<50. As a whole, APPE tends to stabilize with n>200 and for n≈100 APPE tends to be below 10%. For all the evaluated datasets, the errors depend on the sample size used to compute Θ^n.

For APPE, the Gene70 case (Figure 2), results were similar to PAM50, with APPE up to 8% when n<50. Thus, both MS algorithms showed a similar behavior, with their performances depending on the number of subjects used to estimate the SP. The same behavior can be observed for the rest of the datasets (see Supplementary Figure 1).

To evaluate error severity for PAM50 subtypes prediction, we compared predictions with worse (Basal, Her2, and Luminal B) and better (Luminal A and Normal-like) prognoses [8,11–13] on the same subjects, standardized with the criterion standard Θ^N, against those normalized with Θ^n for different sample sizes n. These values, SAPPE, are also shown in Figure 1. It is possible to see that for n<50, SAPPE is close to 15%. The Basal subtype was the least sensitive to SP estimation with different sample sizes, followed by the Her2 subtype. The Luminal subtypes were very sensitive. In particular, Luminal B cases were the most sensitive ones $(SAPPE_n^{LuminalB,(LuminalA,Normal)})$.

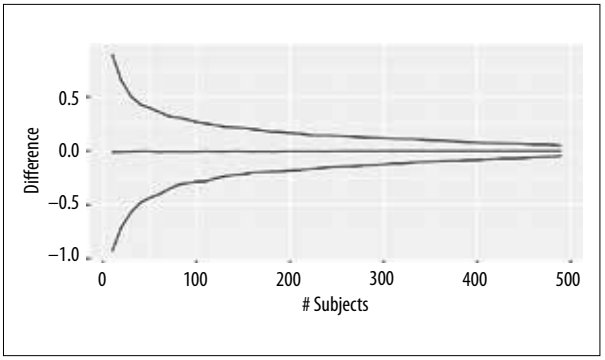


Figure 3. The 10th, 50th (median), and 90th percentiles of the distribution of difference= $\bar{g}_k^N - \bar{g}_k^n$ for the TCGA dataset for all the signature genes in PAM50 and n=10,20,..., $[\frac{0.9}{10}N]\cdot 10$ in the TCGA dataset.

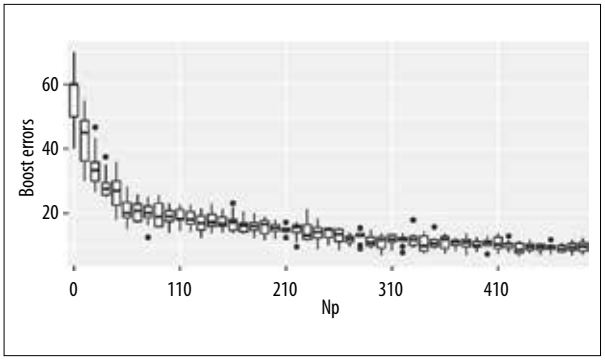


Figure 4. Simulation of the recruitment process for the TCGA BC cohort. BoostErrors=Bootstrapped APPE. N_p – number of recruited subjects.

This means that a subject classified as a Luminal B using a large sample size for SP estimation could be misclassified as a Luminal A (a subtype with good prognosis) when estimating the SP with an inappropriate sample size, which would have a dangerous effect on therapy. On the contrary, Basal and Her2 subtypes tend to be less sensitive to SP estimation.

In the case of the Gene70 MS (Figure 2), the estimation of SP with an inappropriate sample size has a greater effect on low-risk subjects than on high-risk ones. Up to 17% of low-risk subjects can be wrongly assigned to a high-risk group using n<50. To reach a value below 5% for $SAPPE_n^{Low,High}$, n≥100 could be required.

Analysis of population parameters estimates

Figure 3 shows the 10th, 50th (median), and 90th percentiles of the distribution of the difference $\bar{g}_k^N - \bar{g}_k^n$ for the TCGA dataset for all the signature genes in PAM50 and n=10,20,..., $[\frac{0.9}{10}N]\cdot 10$ in the TCGA dataset (R code available by request for the rest of the datasets). It is shown that the median expression value of a particular signature gene can change by up to 1 (±0.5) of a fold change between the whole median population and a smaller sample size for the TCGA dataset, as well as for the NKI dataset, which also uses Agilent microarray technology. For Affymetrix® data, differences greater than 2 (±1) of a fold change were observed when n<80 (data not shown, R code available by request). In general, median differences tended to decrease and stabilize when approximately n<80 subjects were included in the computation of Θ^n. These deviations around the overall median can explain the errors in the subtype and risk predictions, thus emphasizing the need for a pilot study to determine the appropriate number of subjects for a stable application of the MS in breast cancer.

Proposal for a pilot study to achieve consistent standardization parameter estimates

As a consequence of this comprehensive study, a method to determine if the number of recruited subjects is sufficient for an appropriate SP estimation is proposed. In this regard, we propose to recruit subjects in an incremental process and evaluate in each step the prediction error based on a bootstrap sampling process.

To obtain consistent estimation for SP, the following patient recruitment process is proposed:

1. Set N_p=10.
2. Recruit N_p subjects.
3. Generate B=100 bootstrap samples of size N_p.
4. Compute the SP for each sample. This yields $\Theta_1^{N_p}, ..., \Theta_B^{N_p}$.
5. Apply the model for each subject to obtain predictions $PM_{b,j}^{N_p} = PM\left(x_j, \Theta_b^{N_p}\right)$ for b=1,..., B and j=1,..., N_p.
6. Store prediction results for the specific sample size.
7. Evaluate APPE (Bootstrap Errors) for current N_p over the B repetitions.
8. If APPE is not accepted, increase the sample size N_p by 10 and go to step 2; otherwise, use Θ^{N_p} with current N_p and use this estimate later for prediction.

R code and application of this guideline over the evaluated datasets are available in the supplementary material file.

Figure 4 shows the application of the recruitment process algorithm over the TCGA BC dataset. The algorithm was applied 10 times, randomizing all the subjects to simulate different recruitment processes. It is possible to observe that prediction errors (i.e., the number of subjects that change from good/bad to bad/good PAM50 subtype for each n) tend to diminish. In this case, prediction made with the PAM50 model and N=100 yielded a median Bootstrap Error (APPE) below 20%, differing by up to 1.09% from those achieved with n=547 (see supplementary material for more details). Similar results were

obtained with the other datasets. Choosing *n* as the number of subjects satisfying Bootstrap APPE <20% yields errors below 3% when compared with predictions made with SP computed with all the available population of each dataset.

Discussion

Gene expression profiling studies have modified standards of clinical behavior towards BC, with the potential to be used as a molecular-based diagnostic tool. Although evaluation in terms of concordance, efficacy, and prognosis power has been extensively addressed in the literature, clinical use still remains unclear for most of the gene expression signatures. One of the main drawbacks is the lack of standardization of methodological procedures. In this regard, how to preprocess the new subjects in order to feed the classification tool and how to estimate gene-based standardization parameters, among other questions, still require deeper understanding. In addition, their effect on subject outcome prediction and potential misclassification have been vaguely reported.

In the present study, we simulated different sample sizes (resembling a subject recruitment process) to test gene-based diagnostic tools in a real clinical scenario in terms of signature standardization procedures. The aim of this work was to evaluate classification performance and its impact on outcome prediction of well-known MS predictors in BC. For this purpose, several available gene expression datasets were evaluated. Gene-based signature classification was found to be highly sensitive to the sample size used to estimate SP. Our results show that more than 20% of subjects can be misclassified or assigned to the wrong progression risk class if their MS is standardized with SP estimated with a small number of subjects. This suggests that a minimum number of subjects should be recruited to estimate appropriate and stable SP before applying any gene signature for diagnosis that uses population-based gene corrections. This is a necessary requirement to achieve accurate population estimates for the SP in order to make unambiguous predictions for subjects. Hence, once a sufficient number of BC subjects is obtained, the predicted BC subtype will no longer depend on new subjects. Moreover, the minimum number of patients was found not to be universal and should be estimated at each site, probably taking into account different disease type incidences (i.e., different incidence of ER+/–, PR+/–) [8], as well as ensuring quality control for technical microarray issues [20].

Every molecular signature was originally described in a training set and validated with a test set. However, when any molecular signature is intended to be used in clinical application, it must be standardized for the studied population by means of parameters obtained from a representative sample of that population. Our simulations demonstrated that if this standardization is done "on the go" (e.g., taking all the available data on a given day), a subject may be assigned to one class (e.g., good prognosis) if the standardization was done with a limited number of subjects, or to another class (e.g., bad prognosis) if standardization is performed when more subjects become available. This is unacceptable for clinical purposes and demands that a fixed and sufficiently large number of subjects is used to determine the population parameters.

In our study, both PAM50 and Gene70 molecular signature algorithms depended strongly on the number of subjects used to estimate SP. For PAM50, Luminal subtype subjects were the most sensitive, and a wrong estimate of the SP could yield subject assignments to the wrong outcome, probably leading to inappropriate therapy. Among the Luminal cases, the Luminal B seems to be the most unstable one, in concordance with a previous report [21], where it was found that Luminal B subjects cannot be statistically assigned to any subtype. This particular subtype was reported to be highly heterogeneous [22] and has even been suggested to be more than 1 class [12,23]. Thus, Luminal B cases seem to be far from an accurate characterization. On the contrary, Basal-like and HER2-enriched subtypes require smaller sample sizes than the other BC subtypes to reach a consistent classification. This behavior can also be observed when comparing different SSP [10], where the Basal-like intrinsic subtype was consistently identified in the same dataset, even when using different signatures. For Gene70 molecular signature, a similar behavior was found, were High-risk subjects could be wrongly assigned to the Low-risk class, with consequent negative effects on therapy and patient quality of life.

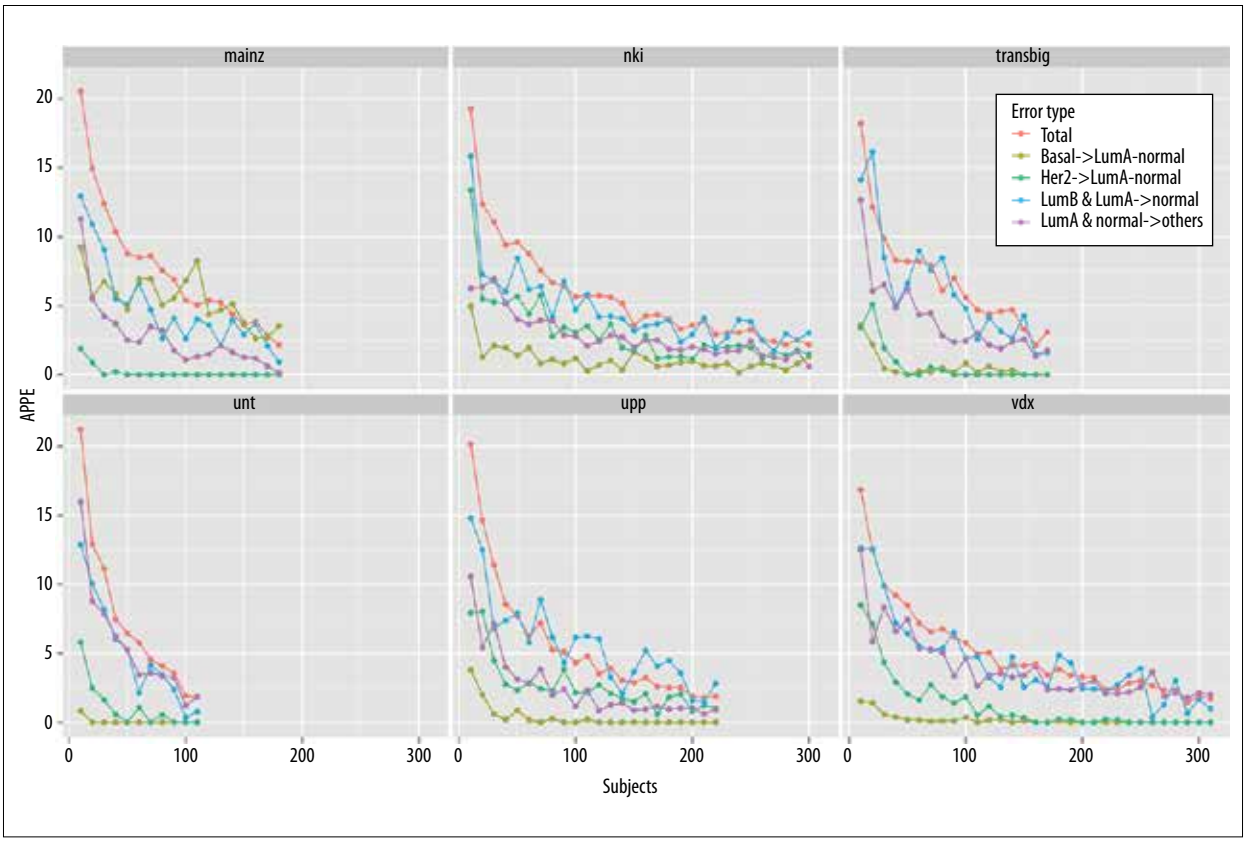
The observed misclassification for a small number of subjects could be due to instability in the estimation of the SP. We found that the estimation of median gene value was variable among different sample sizes compared with the whole sample estimate (criterion standard) (Figure 3). We found up to 2-fold changes of range in the median expression values of a signature gene. This variation might bias the subject-specific gene signature expression, thereby affecting the intrinsic subtype prediction. In general, the median estimates tend to become stable once the number of subjects used for estimation is about n=100. These findings suggest that, when building prognosis methods, the sample size should be large enough to capture gene population parameters, or the prediction method used should not depend on sample-based estimates. In the former case, a pilot study should be conducted, and here we propose a procedure to recognize when the recruited population could yield a stable estimate of the population parameters, which is the sample size required to obtain consistent SP (see R code and evaluation in the supplementary material).

In the present study, we did not address array normalization effects. However, for inter-array normalization, a commonly used approach uses normalization parameters estimated taking into account all of the available arrays. For instance, in one-color chips, robust multi-array average expression measure is typically used, which is based on sample quantiles for chip normalization. For two-color cDNA microarrays, sample quantile inter-array normalization is also available, as well as median absolute deviation [24,25]. In any case, none of these methods are appropriate in a clinical setting, since the arrays should be processed on subject arrival (or at least once a small number of subjects is available). All these inter-array normalization methods, which depend on sample (arrays) statistics, can be affected by the number and quality of arrays included in the normalization step. Accordingly, appropriate quality control procedures and standard operative protocols should be established in the clinical microarray facility for appropriate microarray processing.

Conclusions

In conclusion, our results demonstrate that, if the standardization parameters to be used by the molecular signature algorithm are not appropriately estimated, then subjects might receive inappropriate therapy. This suggests that a pilot study to provide SP that yield a stable prediction should be conducted. In addition, a method to evaluate the sufficiency of the size of the available sample for parameter estimation is proposed to guide prior pilot study development.

Supplementary Figures



Supplementary Figure 1. PAM50 Average Percentage Predictive Errors (APPE) regarding all the prediction errors independent of the subtype, and Severity Average Prediction Errors (related to subtype prediction errors between good (bad) prognosis prediction as in Figure 1 for NKI, VDX, transbig, mainz , unt and upp data sets.

References:

1. Andre F, Pusztai L: Molecular classification of breast cancer: Implications for selection of adjuvant chemotherapy. Nat Clin Pract Oncol, 2006; 3(11): 621–32
2. Perou CM, Sorlie T, Eisen MB et al: Molecular portraits of human breast tumors. Nature, 2000; 406: 747–52
3. van't Veer LJ, Dai H, van de Vijver MJ et al: Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415: 530–36
4. Haibe-Kains B, Desmedt D, Loi S et al: A three-gene model to robustly identify breast cancer molecular subtypes. J Natl Cancer Inst, 2012; 104(4): 311–25
5. Wang Y, Klijn JG, Zhang Y et al: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Lancet, 2005; 365(9460): 671–79
6. Miller L, Smeds J, George J et al: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Nat Acad Sci USA, 2005; 102(38): 13550–55
7. Hu Z, Fan C, Oh DS et al: The molecular portraits of breast tumor are conserved across microarray platforms. BMC Genomics, 2006; 7(1): 96
8. Curtis C, Sohrab P, Suet-Feung Chin S et al: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 2012; 486: 346–52
8. Lusa L, McShane LM, Reid JF et al: Challenges in projecting clustering results across gene expression – profiling datasets. J Natl Cancer Inst, 2007; 99: 1715–23
9. Fan C, Oh DS, Wessels L et al: Concordance among gene-expression-based predictors from breast cancer. N Eng J Med, 2006; 355: 560–69
10. Weigelt B, Mackay A, A'Hern R et al: Breast Cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol, 2010; 11: 339–49
11. Sotiriou C, Wirapati P, Loi S et al: Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst, 2006; 98: 262–72
12. Fresno C, González GA, Merino GA et al: A novel non-parametric method for uncertainty evaluation of correlation-based molecular signatures: Its application on PAM50 algorithm. Bioinformatics, 2017; 33(5): 693–700
13. Parker, JS, Mullins M, Cheang MCU et al: Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol, 2009; 8: 1160–67
14. Paik S, Shak S. Tang G et al: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med, 2004; 351: 2817–26
15. Perou CM, Parker JS, Prat A et al: Clinical implementation of the intrinsic subtypes of breast cancer. Lancet Oncol, 2010; 11(8): 718–19
16. Sørlie T, Borgan E, Myhre S et al: The importance of gene-centering microarray data. Lancet Oncol, 2010; 11(8): 719–20
17. Domany E: Using high-throughput transcriptomic data for prognosis: A critical overview and perspectives. Cancer Res, 2014; 74(17): 4612–21
19. Fernández EA, Valtuille R, Presedo JM, Willshaw P: Comparison of different methods for hemodialysis evaluation by means of ROC curves: From artificial intelligence to current methods. Clin Nephrol, 2005; 64(3): 205–13
20. Geyer Correa F, Reis-Filho S: Microarray-based gene expression profiling as a clinical tool for breast cancer management: Are we there yet? Mol Surg Pathol, 2009; 17(4): 285–302
21. Goldhirsch A, Wood WC, Coates AS et al: Strategies for subtypes-dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the primary therapy of early Breast Cancer 2011. Ann Oncol, 2011; 22: 1736–47
22. Fresno, C, Balzarini MG, Fernández EA: Lmdme: Linear Models on Designed Multivariate Experiments in R. Journal of Statistical Software, 2014, 56(7)
23. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast cancer tumors. Nature, 2012; 490: 61–70
24. Irizarry RA, Hobbs B, Collin F et al: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003; 4(2): 249–64
25. Smyth GK, Speed TP: Normalization of cDNA microarray data. Methods, 2003; 31: 265–73