Journal of
Proteomics & Bioinformatics

Truntzer et al., J Proteomics Bioinform 2013, 6:8
http://dx.doi.org/10.4172/jpb.1000276

Research Article Open Access

Impact of the Selection Mechanism in the Identification and Validation of New "Omic" Biomarkers

Caroline Truntzer[1]*, Delphine Maucort-Boulch[2,3,4] and Pascal Roy[2,3,4]

[1]*Proteomic Platform CLIPP, CHU Dijon, Dijon, France*
[2]*CNRS, UMR 5558-Team Health Biostatistics, Villeurbanne, France*
[3]*Biostatistics Health Laboratory, University Claude Bernard Lyon 1-UMR 5558, Villeurbanne, France*
[4]*Department of Biostatistics, Hospices Civils de Lyon, Lyon, France*

Abstract

Background: High throughput analysis like mass spectrometry dedicated to clinical proteomics offers new insights into clinical research. This promising technology generates high-dimensional datasets with a huge amount of biological input. Working with these high-dimensional datasets has created challenges for statistical methods and there are still weaknesses in current statistical analysis that have to be overcome to get an accurate interpretation of "omics" studies. The central question is that of a reliable identification of new prognostic and diagnostic biomarkers. Although observed in previous studies, these mechanisms of identification and validation of new markers have been inadequately explained and often dissociated.

Results: The aim of our study was therefore to show how candidate markers are sometimes selected in identification studies because of biased estimations of their effect. To achieve this goal, this work was conducted through the simulation of high-dimensional studies concerning survival. We showed how the selection mechanism involved in identification studies influences a mechanism called regression to the mean. This in turn leads to a biased estimation of the effect size and thus to optimism when considering validation studies.

Conclusions: This study demonstrated why the discovery of new robust markers is only possible through well-designed studies relying on consistent sample sizes for the identification step. Due to the above mentioned mechanisms of identification and validation, pertinent candidate biomarkers in high-dimensional clinical studies require non-biased estimation, and this right from the identification step. Only then will it lead to consistent studies and thus reach benefit in terms of health care.

Keywords: High dimension; Biomarker selection; Validation; Regression to the mean

Background

Nowadays, cancer research is making use of new high throughput technologies, like mass spectrometry in the field of clinical proteomics. Mass spectrometry signals show the proteomic profiles of the individuals under study at a given time. These profiles correspond to the recording of a large number of proteins, much larger than the number of individuals. Thus, this leads to the generation of high-dimensional datasets with a huge amount of biological input. Working with high-dimensional datasets has created a number of challenges for statistical methods. In particular, it raises two main statistical questions: the identification of candidate markers and their validation in further studies.

A classical clinical study starts with an a-priori hypothesis made by the clinician about the potential prognostic or diagnostic effect of one particular clinical factor. Usually, only one or very few variables are tested in a single study. Statistical models aim to validate (or not) this hypothesis by estimating the strength of the association between this variable and the clinical outcome of interest, and by testing its significance. In contrast, there is no a-priori hypothesis in "omic" studies, where a huge number of variables are tested simultaneously. Thus, a two-step strategy is needed for these studies.

The first step corresponds to identification studies designed to select a list of candidate biomarkers tested among a high number of biological parameters; this is conducted by analyzing a sample of the population under study. This identification step can in turn be broken down into two sub-steps: estimation and selection. In fact, the selection of relevant markers relies on the estimation of the strength of association between each biological input and the clinical outcome of interest. Only values with a significantly high enough strength of association are selected.

These studies lead to the acquisition of large number of variables. These variables are potential biomarkers and may be of several types. In the context of clinical proteomics, data generated by mass spectrometers correspond to the proteomic profile of each of the individuals under study at a given time. These profiles correspond to the recording of the intensities of a high number of proteins expressed by the genome of the individuals. Besides proteomics datasets, other "omics" datasets representing other biological levels are also concerned by this huge quantity of variables. Genomics aims at learning about genes through the study of SNPs for example (Single Nucleotide Polymorphism, i.e. DNA sequence variation), while transcriptomics aims at learning the expression and regulation of genes through the study of RNA. All these studies are characterized by the same huge amount of variables. Given the cost or the difficulty to get biological samples, this large number of variables often go with relatively low number of observations. By

***Corresponding author:** Caroline Truntzer, Proteomics Platform CLIPP, CHU Dijon, Dijon, France, E-mail: caroline.truntzer@clipproteomic.fr

Received June 12, 2013; **Accepted** August 12, 2013; **Published** August 16, 2013

Citation: Truntzer C, Maucort-Boulch D, Roy P (2013) Impact of the Selection Mechanism in the Identification and Validation of New "Omic" Biomarkers. J Proteomics Bioinform 6: 164-170. doi:10.4172/jpb.1000276

chance alone, many potential markers may be found significantly associated with the outcome, even though most of them may not actually be linked to diagnosis or prognosis. The question of multiple testing was hot debated through the definition of the False Discovery Rate (FDR), for example, that is, the expected proportion of false positives among the genes declared as significant [1]. When looking for differential genes, Pawitan et al. [2] showed that the FDR is mostly influenced by the proportion of truly differentially expressed genes, and by the sample size. Some additional works showed how the type I error rate was related to the power in high-dimensional setting [3,4], the increase in the power being at the cost of the FDR. Some authors illustrated the issues related to the selection process in the "omic" field. For example, Michiels et al. [5] showed through the well-known dataset from Van't Veer et al. [6], that the list of selected candidates was highly unstable and depended on the composition of the identification set. Later, Ein-Dor et al. [7] proposed a tool to evaluate the selection process and showed that thousands of samples are needed to ensure a robust signature.

The second step corresponds to validation studies designed to confirm the previously selected candidate(s) as biomarker(s). This step can also be broken down into two sub-steps: re-estimation and confirmation. These studies aim to re-estimate on independent datasets the strength of association of the previously selected candidates, and thus confirm (or invalidate) the relevance of candidates as markers. Relevant markers may have optimal predictive quality. Some authors illustrated the divergence between the strength of association estimated in identification studies and that estimated in validation studies. In particular, Michiels et al. [5] showed on several well-known datasets how an inadequate validation led to the publication of overoptimistic results compared with those from their own analyses. Later, Truntzer et al. [8] showed how high-dimensional data analysis was subject to greater optimism-that is to say an over-estimation of the strength of association- compared with analysis of classical clinical variables.

Some solutions were proposed to correct the optimism bias linked to the selection process. Using resampling methods [9,10] or penalized regression [11-13] are such examples. The objective of this paper, however, is not to propose new solutions, but rather to explain how it works. Indeed, to our knowledge, the questions of identification and validation have been highlighted, but neither the mechanisms, nor the ways in which these two steps are strongly associated have been thoroughly explained. In this work, we propose to explain the link between estimation and selection. To understand the process involved, we will analyse how candidate markers are selected in identification studies and how their estimated strength of association may be reduced-and thus not confirmed-when re-estimated in validation studies. In other words, to better understand how selection leads to optimism, we propose to show how the estimation bias that may occur in the identification steps leads to selection of inappropriate candidate markers. For recall, regression toward the mean refers to the phenomenon that a variable that is extreme on its first measurement will tend to be closer to the mean of the distribution on a later measurement [14,15]. In fact, let consider a given variable. Its measurement varies around its mean following a given distribution. When sampling a first measurement from this distribution, there is a low probability of observing it extreme. So, if a first measurement is extreme, there is high probability that the second one spontaneously regress towards the mean value [16].

Methods

Simulation of the datasets

Comprehension of the mechanisms involved in the identification and validation steps was achieved through simulations of survival "omic" datasets. Indeed, simulations have main advantage of offering a situation in which the truth is known and can even be controlled. The same processes as those encountered in real-life clinical studies can also be reproduced with the advantage that all the parameters can be controlled.

The same simulation process as described in a previous paper by the same authors was used [8]. Here is a brief description of this process. A classical way to link variables to censored survival data is to use the Cox proportional hazards model. Let us de ne X an *(n, p)* matrix of *p* variables for n individuals. For each of the n individuals, the follow-up times were noted $t_1,\ldots, t_n$ as were the event-indicators $\delta_1,\ldots,\delta_n$ with $\delta_i=1$ if the event occurred and $\delta_i=0$ if it did not occur. At time t, the Cox proportional model is given by

$$\lambda(t|X) = \lambda_0(t)\exp(\beta' X) \quad (1)$$

Where $\lambda_0(t)$ is a baseline hazard function, $\beta=\{\beta_1,\ldots,\beta_n\}$ is the vector of parameters and $X_1,\ldots,X_p$ are the vectors of length *n* describing each of the *p* variables for the *n* patients.

We simulated a virtual population of size *n* in which each individual is described by *p* "omic" variables -with $n<<p$- and survival information. Normal distributions N(0,1) were assumed for the "omic" variables. A Weibull distribution with shape parameter 5 and scale parameter 2 was used for the baseline function. For censoring times, a uniform U(0,8) was used, leading to about 40% censoring. Only p_1 of the *p* variables were considered as related to survival; the remaining p_0 were under the null hypothesis H_0 of no association with survival. p_1 coefficients of the Cox model were thus set at 0.2, β_j, $j=1,\ldots,p_1$, and the remaining p_0 were set at 0, β_j, $j=p_1+1,\ldots,p$. Note that $p=p_1+p_0$. This is represented by the left panel labeled "truth" in Figure 1.

For a fixed set of parameters *p* and p_1, 200 identification sets of *n* patients were simulated according to the above design. For each of these identification sets, 50 corresponding validation sets were drawn up following the same design. This overall process was performed by considering *n* in {100,200,400,1000}. In this study we chose p=1000 and p_1=20.

For each simulated identification set, univariate Cox regression models were used to estimate the strength of the association of each variable through survival model parameters. Based on these estimations, the R most contributive variables were selected in a univariate way.

Selecting variables in the multiple hypothesis setting results in considering the problem of testing simultaneously *p* null hypotheses, leading to different situations, described in Table 1 [17]. Among the *p* corresponding variables, p_1 are under the H_1 hypothesis (H_i^j ; j=1,…., p_1), while p_0 are under the H_0 hypothesis (H_0^j ; $j=p_1+1,\ldots,p_0+p_1$). The test leads to the rejection of R hypothesis.

Among the R rejected hypotheses, V are under the null hypothesis (False Positives or FP), whereas S are actually under the alternative hypothesis (True Positives or TP). In the same way, among the (p-R) variables, T were wrongly not selected. The Type I error concept had to be newly defined to take into account the huge number of hypotheses tested. The basic idea is to adjust p-values of usual test statistics in order

to control the global error rate. For this purpose, the control of the False Discovery Rate, that is the expected proportion of Type I errors among the rejected hypotheses $FDR = E(V/R)$, is commonly used [1]. In general, one would like to minimize the number V of false positives, or Type I errors and the number T of false negatives, or Type II errors, thus maximizing the power, defined as $E(S)/p_1$.

Note that the identity of the R variables depends on the identification set. Indeed, the same set of variables is not systematically selected from one dataset to the other.

- In this work we were interested in the estimation of the strength of association for the R variables, which was estimated through univariate Cox survival models. Once the R variables had been selected on one identification set, the Cox coefficients of these R variables were then re-estimated on the corresponding validation sets. Estimations for the V and the S variables were stored separately. Note that there was no new variable selection on the validation sets. In parallel, Cox coefficients for the p_0 and p_1 variables were estimated on each identification set. The whole process was performed for each of the 200 identification sets, and is illustrated in Figure 1. To sum up, we considered the distributions of the Cox coefficients estimated for:The p, p_0 and p_1 variables over the 200 identification sets. The sets of p, p_0 and p_1 variables are respectively denoted Ω_p, Ω_{p0} and Ω_{p1}, hereafter.
- The V, S and R variables over the 200 identification sets. These sets of variables are respectively denoted Ω_V,Ω_S and Ω_R, hereafter. The Ω_V, Ω_S and Ω_R sets were defined separately on each identification set. Keep in mind that the variables constituting these sets of variables are not the same, depending on the identification dataset. While R depends on the datasets when the FDR control is applied, it is fixed otherwise. The distributions of the estimates for these sets of variables will make it possible to understand what happens in identification studies where p "omic" variables are tested, without a-priori hypotheses about their relationship with survival.
- The V, S and R variables over 50 validation sets for each of the 200 identification sets. We insist on the fact that Ω_V,Ω_S and Ω_R were not newly defined on the validation sets. At this step, the selection process is over, and the corresponding validation studies are conducted. We also remind the reader that the identification and the validation sets are the same size.

Results and Discussion

Results

The comparison of the above described densities were used to illustrate how the selection mechanism involved in identification studies influences regression to the mean, and how it leads to over-estimation of the strength of association and thus to optimism.

Results are shown through histograms that display the density of each of the distributions of interest. Each of the following figures is related to one particular set of variables (Figure 1) from the last, but one column of Table 1. Whatever the figure, each of the four panels was obtained with a specific sample size, with n={100; 200; 400; 1000}. The vertical line with abscissa 0.2 corresponds to the simulated strength of association; in other words, it corresponds to the mean distribution of the Ω_{p1} estimates. Note that the number of estimates contributing

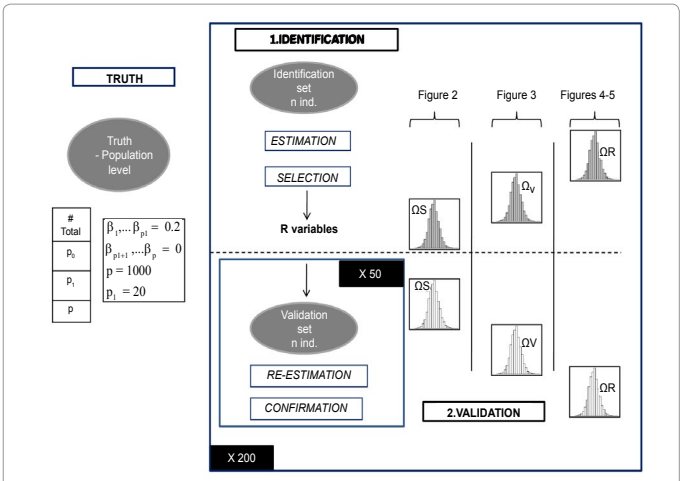

Figure 1: Flowchart of the whole process conducted in this study. Representation of the whole process involved in the discovery of new biomarkers as simulated in this study. The "truth" left part of the figure corresponds to the simulation settings representing the global population. A total of p=1000 variables were under study, among whom p0 were under the null hypothesis H0 of no association with survival, and p1=20 under the alternative hypothesis. The right part of the figure corresponds to the identification and validation steps occurring in "omic" studies. R variables are selected on the identification set, and then validated in 50 distinct validation sets.

	#Not rejected	#Rejected	#Total
# True hypothesis	U	V	p_0
#Not true hypothesis	T	S	P_1
# Total	p-R	R	p

U: True negatives; T: False negatives; V: False positives; S: True positives.

Table 1: Multiple testing setting.

to the distributions density is not the same for the identification and validation sets. In the following, results concern the "top-20" approach.

Figure 2 concerns distributions of the strength of association estimated over 200 identification sets for the Ω_{p1} (grey histograms) and Ω_S (horizontal hatching) sets of variables. Let us restate that Ω_S corresponds to TP. So, Ω_S is a subset of Ω_{p1}, selected because of the high strength of association estimated for the corresponding variables. First, one observes that the more patients included in the study, the narrower the distributions. This is a well-known statement according to which variance decreases with sample size. Second, the first panel shows that with n=100 patients, variables from the Ω_S set are selected in the right extreme of the distribution of the Ω_{p1} estimates. As a consequence, the mean distribution of the Ω_S estimates is far away from the mean distribution of the Ω_{p1} estimates (vertical line). When n increases, $_S$ variables are still selected in the right extreme of the Ω_{p1} estimates distribution, but as the distribution for Ω_{p1} narrows around its mean, estimates are less extreme. As a consequence, the distribution mean of Ω_S estimates decreases and tends toward the distribution mean of Ω_{p1} estimates. With n=1000, distributions for the estimates of Ω_{p1} and Ω_S variables are almost superimposed.

Figure 3 concerns distributions of the strength of association estimated over 200 identification sets for the Ω_{p0} (grey histograms) and Ω_V (horizontal hatching) sets of variables. Let us restate that Ω_V variables correspond to FP, and are also a subset of Ω_{p0}. The distribution of the estimates for the Ω_{p0} variables is around 0 (its true mean) and gets narrower with increasing sample sizes. The distribution of the estimates

for the Ω_V variables is bimodal. Variables from the Ω_V set were selected in both extremes of the Ω_{p0} estimates distribution. With n=100 or 200, the estimates of the variables selected in the right extreme are even higher than the true strength of association simulated for variables related to survival (vertical line). In parallel, one observes that each mode of the distribution of the estimates for Ω_V is far away from the mean distribution of Ω_{p0} estimates. When increasing the sample size, the distribution of Ω_{p0} estimates gets narrower and thus the extreme of the distribution moves away from the vertical line. In parallel, each mode of the distribution of the estimates for Ω_V approaches zero.

To go further, Figure 4 illustrates the mechanism encountered when considering the estimates of all selected variables constituting Ω_R. This figure concerns both identification and validation steps. For this purpose, each panel shows the distribution of the estimated strength of association for Ω_R computed over 200 identification sets parameters (horizontal hatching) and over 200*50 validation datasets (diagonal hatching). The vertical dotted line corresponds to the mean distribution of Ω_R estimates computed on the validation sets. For the identification datasets (horizontal hatching), the distribution started from bimodal with n=100 to unimodal with n=1000 individuals. With small sample sizes, the modes of the distribution are far from 0.2. With n=1000, the left mode vanishes and the mean distribution of the estimates for Ω_R tends toward the mean distribution of the estimates for Ω_{p1}, that is 0.2. For the validation sets, the distribution is first unimodal (n=100), and tends toward 0, indicating that the majority of the Ω_R variables are in reality false positives. With n=200, a shoulder appears on the left of the distribution, and the shoulder moves from the left to the right. In parallel, the mean distribution of the strength of association estimated for Ω_R tends towards 0.2 and matches this value with n=1000. Another noticeable observation is that estimates of $_R$ from the identification and the validation sets come closer with increasing sample sizes. With n=1000, the estimations obtained on identification and validation sets join.

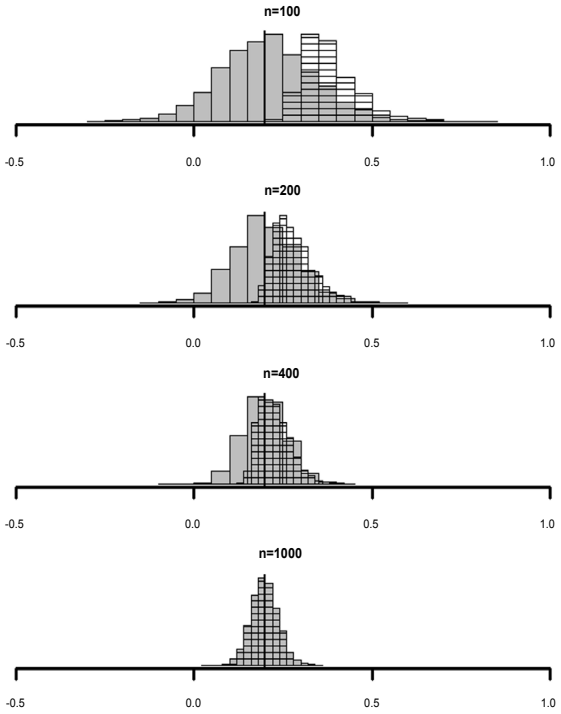

Figure 2: Parameter estimates with varying n for variables under the H_1 hypothesis. This figure only concerns estimates for identification sets. Each of the four panels was obtained with a specific sample size with n={100; 200; 400; 1000}. Whatever the panel, the following distributions were plotted: 1-distribution of the estimates for the Ω_{p1} variables computed over 200 identification sets (grey histogram). 2-distribution of the estimates for the Ω_S set of variables computed over 200 identification sets (histogram with horizontal hatching). The vertical continuous line indicates 0.2.
With n=100, estimates for Ω_{p1} are highly fluctuating, as shown by the wide distribution. Variables are selected in the extreme of the distributions of the estimates for Ω_{p1}, and these mean estimates are thus far from their true means. When increasing the sample sizes, the mean distribution of Ω_S estimates tends toward the mean distribution of Ω_{p1} estimates. This is an illustration of the regression to the mean phenomenon that leads to over-estimation of the strength of association for true positives.

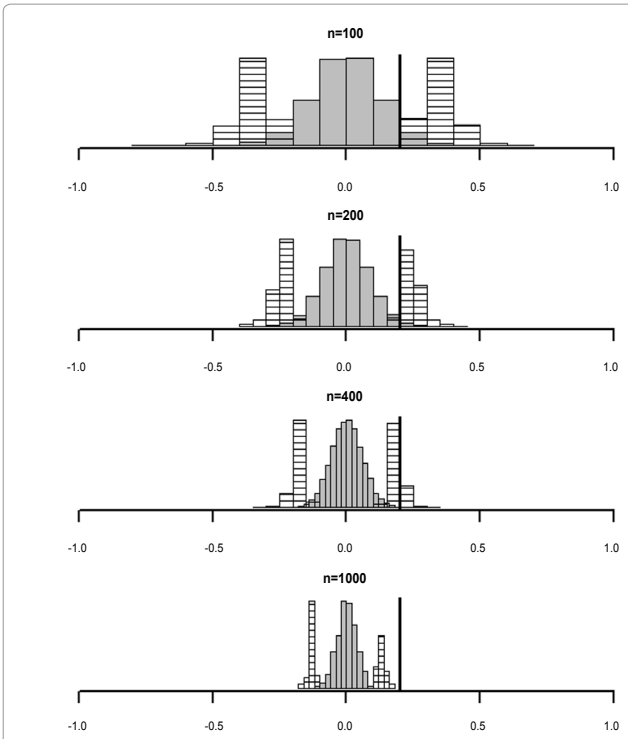

Figure 3: Parameter estimates with varying n for variables under the H_0 hypothesis. This figure only concerns estimates for identification sets. Each of the four panels was obtained with a specific sample size with n={100; 200; 400; 1000}. Whatever the panel, the following distributions were plotted: 1-distribution of the estimates for the Ω_{p0} variables obtained over 200 identification sets (grey histogram). 2-distribution of the estimates for the Ω_V sets obtained over 200 identification sets parameters (histogram with horizontal hatching). The vertical continuous line indicates 0.2.
With n=100, estimates for Ω_{p0} are highly fluctuating, as shown by the wide distribution. Variables are selected in the extreme of the distributions of Ω_{p0} estimates and the mean estimates of Ω_V variables are thus far from their true means. When increasing the sample sizes, the distribution of the estimates for the Ω_{p0} variables gets narrower and the mean distribution of the Ω_V variables estimates decreases. This illustrates the regression to the mean phenomenon that leads to the inappropriate selection of some FP variables that have in fact no effect on survival.

Discussion

With too few individuals under study, estimates for Ω_{p1} (Figure 2) and Ω_{p0} (Figure 3) are highly fluctuating, which is indicated by the wide distributions. Variables are selected because of their high estimates, which lie in the extreme of the distributions of the estimates for the Ω_{p0} or Ω_{p1} variables. As a consequence, the mean estimates obtained for selected variables were far from their true means: the means for Ω_V and Ω_S were higher than the means for Ω_{p0} and Ω_{p1}, respectively.

This demonstrates a selection bias. When increasing the sample sizes, the mean distribution of the estimates for Ω_S tends toward the mean distribution of the estimates for Ω_{p1} due to a decrease in the selection bias. This is due to the regression to the mean phenomenon, which is influenced by the selection process. This phenomenon affects both Ω_{p0} and Ω_{p1} estimates with different consequences. As for Ω_{p1}, regression to the mean leads to over-estimation of the strength of association for true positives. As for Ω_{p0}, the poor distribution of its estimates leads to the inappropriate selection of some FP variables that have in fact no effect on survival.

In the light of these comments, Figure 4 shows how the above cited mechanisms affect the selection of candidate biomarkers, and the further consequences in terms of their confirmation. Because of the regression to the mean phenomenon described above, R is a mixture of S and V. Whatever the dataset (identification or validation), the right mode mostly consists of the estimates of true positives (Ω_S), whereas the left mode mostly consists of estimates of false positives (Ω_V). Increasing the sample size results in an increase in S at the cost of a decrease of V. Modification of this mixture explains the modification of the shape of the distribution of the distribution for the Ω_R estimates.

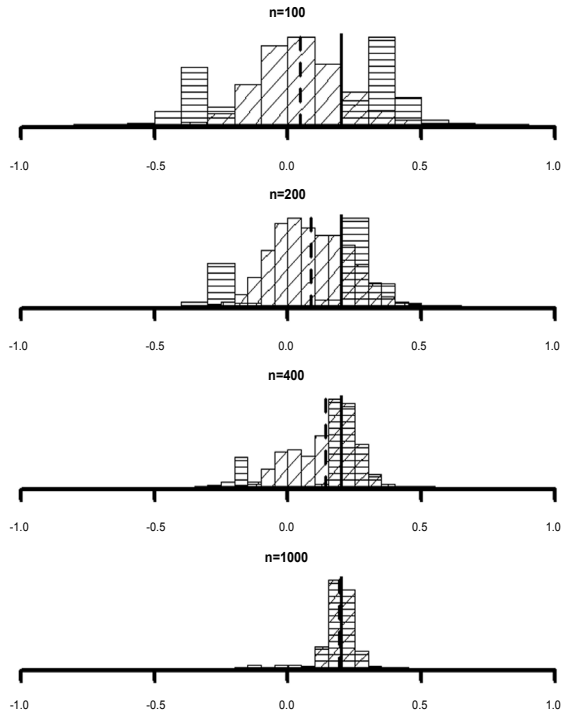

Figure 4: Parameter estimates with varying n for all selected variables. Each of the four panels was obtained with a specific sample size with n={100; 200; 400; 1000}. Whatever the panel, the following distributions were plotted: 1-distribution of the estimates for the Ω_R sets over 200 identification sets parameters (histogram with horizontal hatching). 2-distribution of the estimates of the Ω_R sets over 200*50 validation datasets (histogram with diagonal hatching). The vertical dotted line indicates the mean of the latter distribution. The vertical continuous line indicates 0.2.
The R selected variables are a mixture of False Positives (V) and True Positives (S) that respectively corresponds to the left and the right mode of the distributions of Ω_R estimates. When increasing the sample size, S increases at the cost of a decrease of V, thus modifying the shape of the distribution of the estimates for the Ω_R variables. When n=1000, there are fewer FP and the estimates for TP are no longer over-estimated. This shows the consequence of the regression to the mean phenomenon in terms of confirmation of the selected candidate biomarkers.

In real life studies, variables are selected according to the strength of association estimated during the identification step. The strength of association has then to be re-estimated during the validation step, in order to confirm the effect on survival of the corresponding candidate biomarkers. It appears that the distance between the estimates on the identification and the validation datasets is high with n=100 individuals: when re-estimating the strength of association of Ω_R on independent datasets, it falls and tends toward 0. This divergence between the first estimation and the re-estimation is an illustration of optimism. Thus, it shows how regression to the mean leads to optimism. By quantifying this divergence, it is possible to quantify optimism. With n=1000, distributions of the estimates for Ω_{p0} and Ω_{p1} variables fluctuated to a lesser degree. Thus, selection bias and regression to the mean decreases; Ω_R is almost completely composed of Ω_S and the two distributions superimpose: there are fewer FP and the estimates for TP are no longer over-estimated.

These results demonstrate how a biased estimation of the parameters on the identification sets influences the selection of TP and FP, and illustrates how power increases and optimism decreases with increasing sample size.

These comments demonstrate why large sample sizes in high-dimensional studies are important. Indeed, the estimation of the strength of association from the identification step is critical, because it influences regression to the mean through the selection of variables, and therefore, their validation on new independent datasets. This is important to keep in mind when calibrating new "omic" studies. At present, many current studies are designed to identify new markers on small sample sizes; this choice is justified by claiming that the candidate biomarkers will be validated on larger sample sizes. In complement to Figure 4, Figure 5 shows that this reasoning is incorrect (same legend as for Figure 4), and prevents the identification of relevant markers. The above panel of Figure 5 shows the results obtained with identification and validation datasets of respectively 100 and 1000 individuals. Because of the poor estimates obtained with n=100, many variables are wrongly selected, and have far lower estimates on validation datasets, due to regression to the mean. As demonstrated through the above results, the identification step can only be improved by generating non-biased estimates of the strength of association for Ω_P; this is made possible by using larger sample sizes in the identification step. Increasing the size of validation sets cannot improve the first estimation obtained during the identification step. This is confirmed on the bottom panel of Figure 5, where estimation and selection were conducted on identification sets of 1000 patients and validated on 100 patients. This time, the strength of association is correctly estimated on the identification sets, and thus, confirmed on the validation sets, even though they are of smaller size.

All results discussed above were obtained when the 20 most relevant variables were selected using the log-rank statistics. Similar results were obtained with control of the FDR (results not shown). In fact, controlling the proportion of FP does not correct the estimation and selection bias, thus also leading to regression to the mean and optimism. The main difference lies in the way variables are selected. When n is small, the estimation bias is high, therefore, many FP would be selected; thus type I correction leads to small R. When n increases, both R and S increase so that the proportion of FP stays constant. Thus, variables with lower strength of association than the simulated ones have to be selected to satisfy this constraint. For this, the minimum value of selected variables is moved to the left when n increases (results not shown), leading to a decrease of the selection bias.

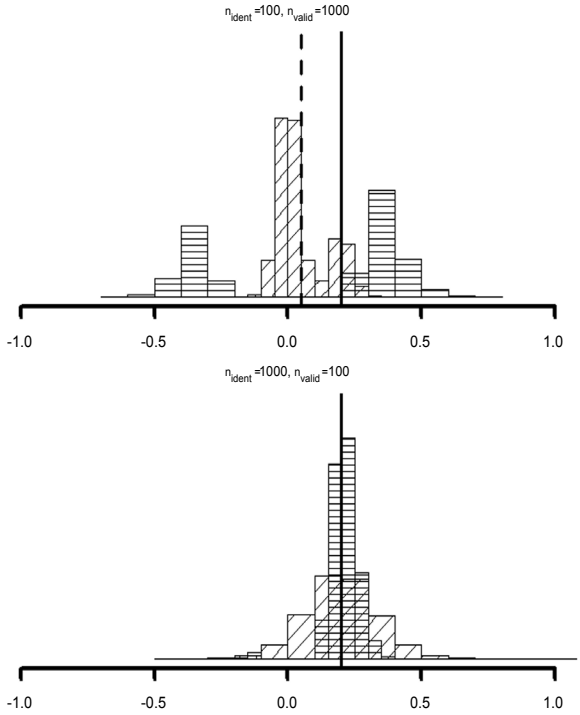

Figure 5: Parameter estimates for all selected variables with n={100; 1000} and n={1000; 100}, respectively, for identification and validation sets. At the top n={100; 1000}; at the bottom n={1000; 100}, respectively, for identification and validation sets. The following distributions were plotted: 1-distribution of the estimates for the Ω_R sets over 200 identification sets parameters (histogram with horizontal hatching). 2-distribution of the estimates for the Ω_R sets over 200*50 validation datasets (histogram with diagonal hatching). The vertical dotted line indicates the mean of the latter distribution. The vertical continuous line indicates 0.2.
With n=100, estimates of the strength of association are poor and have far lower estimates on validation datasets, even large ones like with n=1000. With n=1000 in the identification datasets, however, the strength of association is correctly estimated; as a consequence, it is confirmed on the validation sets, whatever their size.

Nevertheless, these results were observed when using a high value of FDR (FDR=0:2). This high value was chosen to mimic the exploration situation. It would be interesting to evaluate the influence of the choice of the FDR value on these results.

In this paper, we chose to focus on the regression to the mean phenomenon, to explain and illustrate through a simulation process how it takes a major place in the mechanisms of identification and validation of candidate biomarkers. However, it is useful to keep in mind that this is not the only issue that may affect the relevance of "omics" studies.

In particular, we chose not to consider in the simulation process, the role of the variables database defining the set wherein candidate markers will be searched for and selected. In fact, the selected candidates are highly dependent on the choice of the initial set of variables. In the case of genomics or transcriptomics, for example, this will influence the choice of the array. This choice is oriented by the clinical question underlying the study. In the case of clinical proteomics in discovery stage, the initial database reference is partially defined by the type of biological material used as sample (blood, tissue, biopsy, etc...), and by the retained purification method. Selected candidate biomarkers highly depend on these choices, as the reference database will highly vary depending on the type of purification, and/or biological material. It is to note that for clinical proteomics in discovery stage; by contrast with genomics or transcriptomics, the choice is only partial because the exact content of proteins or peptides under study is not known a priori. Moreover, only proteins already identified in known data banks will be used. These choices may then be analyzed in complementary ways to take benefit from the distinct information coming from each of them. In the case of proteomics dataset, another issue may occur. In fact, the reference database may not exactly reflect the content of the processed sample due to technical artifacts like limit of detection, and/or resolution of the measure instrument, statistical preprocessing, and so on. This leads to "technical" missing values that are then missed from the statistical study and this without any biological basis.

To sum up, the database reference is a finite ensemble, and is chosen with an a priori knowledge, and this a priori may lead the investigator to miss some interesting candidate biomarkers.

Conclusions

The objective of this work was to demonstrate how the discovery of important variables directly follows from the estimation of effect sizes, and how it is also influenced by the selection to the mean phenomenon. In fact, the two questions cannot be separated, as this of the statistical power. When searching for new markers, the true strength of association is not known. Sampling of the population concerned is used to obtain an unbiased estimation of it, and thus, to select relevant markers among a large number of "omic" variables. This exploratory stage involves two must-have steps: an identification step and a validation step. During the identification step, potential markers are selected on the basis of their estimated strength of association. In this work, we showed how the selection process influences regression to the mean. Understanding the phenomenon is a first step to overcoming the problems caused by regression to the mean. Only variables with extreme estimated strength of association are selected. This leads to a selection bias that is all the more strong that sample sizes are low. This favors regression to the mean and optimism, the impact of which is then highlighted through validation studies, with bad consequences. In fact, it will finally lead to the selection of variables wrongly considered as candidate biomarkers.

Pertinent "omic" clinical studies are only possible if the strength of association is estimated in a non-biased way in the identification step. Consistent sample sizes will have two effects: 1-improvement in the accuracy of estimates due to regression to the mean. 2-gain in power. Only then will it be possible to identify relevant markers whose effects will be confirmed on independent datasets, and thus be used in the clinical practice.

Acknowledgement

We wish to thank Philip Bastable for editing the manuscript.

References

1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B 57: 289-300.
2. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A (2005) False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 21: 3017-3024.
3. Lee ML, Whitmore GA (2002) Power and sample size for DNA microarray studies. Stat Med 21: 3543-3570.
4. Efron B (2007) Size, power and false discovery rates. Ann Stat 35: 1351-1377.

5. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: A multiple random validation strategy. Lancet 365: 488-492.

6. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536.

7. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103: 5923-5928.

8. Truntzer C, Maucort-Boulch D, Roy P (2008) Comparative optimism in models involving both classical clinical and gene expression information. BMC Bioinformatics 9: 434.

9. Molinaro AM, Simon R, Pfeiffer RM (2005) Prediction error estimation: A comparison of resampling methods. Bioinformatics 21: 3301-3307.

10. Wu LY, Lee SS, Shi HS, Sun L, Bull SB (2005) Resampling methods to reduce the selection bias in genetic effect estimation in genome-wide scans. BMC Genet 6: S24.

11. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58: 26788.

12. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16: 385-395.

13. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32: 407-451.

14. Bland JM, Altman DG (1994) Regression towards the mean. BMJ 308: 1499.

15. Bland JM, Altman DG (1994) Some examples of regression towards the mean. BMJ 309: 780.

16. Davis CE (1976) The effect of regression to the mean in epidemiologic and clinical studies. Am J Epidemiol 104: 493-498.

17. Dudoit S, Shaffer J, Boldrick J (2002) Multiple hypothesis testing in microarray experiments. UC Berkeley Division of Biostatistics Working Paper Series Working Paper 110.

Citation: Truntzer C, Maucort-Boulch D, Roy P (2013) Impact of the Selection Mechanism in the Identification and Validation of New "Omic" Biomarkers. J Proteomics Bioinform 6: 164-170. doi:10.4172/jpb.1000276