Kaplan-Meier curves, Cox model and p-values are not enough for the prognostic evaluation of tumor markers: statistical suggestions for a more comprehensive approach.

Patrizia Boracchi[1], Paola Roccabianca[2], Giancarlo Avallone[3], Giuseppe Marano[1]

1: Department of Clinical Sciences and Community Health, Laboratory of medical Statistics, Biometry and Epidemiology “G.A. Maccacaro”, Università degli Studi di Milano, Milan, Italy

2: Dipartimento di Medicina Veterinaria, Università degli Studi di Milano, Milan, Italy

3: Department of Veterinary Medical Sciences, University of Bologna, Ozzano dell'Emilia, Italy

Corresponding Author:

Giuseppe Marano, Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Biometry and Epidemiology “G.A. Maccacaro”

Campus Cascina Rosa. Via Vanzetti 5, 20133 Milan

Email: giuseppe.marano@unimi.it

Abstract

The assessment of prognostic markers is key to the improvement of therapeutic strategies for cancer patients. Some promising markers may fail to be applied in clinical practice because of misleading results ensuing from inadequate planning of the study and/or from an over-simplified statistical analysis. The main issues involved in an efficient clinical study planning and the subsequent statistical analysis aimed to the prognostic evaluation of a cancer marker will be illustrated and discussed. The aim will be also to extend the offset of most applied statistical models, i.e. Kaplan-Meyer and Cox, to enable the choice of the methods most suited for the study endpoints Specifically, for tumor centered endpoints, like tumor recurrence, the issue of competing risks will be highlighted . For markers measured on a continuous numerical scale, a loss of relevant prognostic information may occur by setting cut-offs, thus the methods to analyze the original scale will be explained. Furthermore, p-value is not a sufficient criterion to assess the usefulness of a marker in clinical practice; to such end, measures for evaluating the ability of the marker to discriminate between “good” and “bad” prognoses are illustrated.

For illustrative purposes, an application of useful methods of analysis to a public dataset from human breast cancer patients, is shown. Tumor size, Tumor grading, number of axillary lymph nodes were considered as known prognostic factors, and the amount of Estrogen receptor content, recorded as quantitative continuous scale, was selected as the prognostic marker

Keywords:

Tumor markers, Prognostic Factors, Survival Analysis, Competing Risks, Cut-Offs. Discriminant Ability

INTRODUCTION AND DEFINITIONS

Oncology research on patient's clinical characteristics and pathological tumor features is aimed to a better understanding of tumor biology, to advance diagnosis and open to new treatment protocols with the final aim to improve prognosis. Although results seem promising, patient's response to anticancer therapies as well as patient's life expectancy are still heterogeneous for the same tumor types. Clinical and pathological characteristics are frequently combined to identify patient groups with different risk of disease progression or treatment response (to date 1240 with oncology and risk groups in title/abstract, see for example: Liu et al[23], Bell et al[4], and Winick et al[38]). This is not surprising as risks groups could be useful to plan first line treatments (e.g. to avoid the potential over-treatment of low-risk patients and/or under-treatment of high risk patients) or to select patients for clinical trials including novel therapeutic principles and protocols according to their health status and probability to treatment response.

The more information available on each specific tumor entity, the greater the possibility of building more effective patient stratification. The addition of tumor markers to other clinical and pathological variables has become a frequent approach because it improves the identification and stratification of cancer patients in different risk groups. According to the definition given by the National Cancer Institute, a tumor marker should be intended as "*anything present in or produced by neoplastic cells or other cells of the body in response to cancer or certain benign (noncancerous) conditions that provide information about a cancer, such as how aggressive it is, whether it can be treated with a targeted therapy, or whether it is responding to treatment. Tumor markers have traditionally been proteins or other substances that are made by both normal and cancer cells but at higher amounts by cancer cells. These can be found in the blood, urine, stool, tumors, or other tissues or bodily fluids of some patients with cancer*".

The oncological research is on-going, and the contribution of tumor markers on body fluids and tumor tissues is investigated in order to explain patient's overall cancer outcome, regardless of therapy (prognostic), or to give information on their effects on a therapeutic intervention (predictive).[26] To decide whether new tumor markers should be included in a strategy for risk group identification their prognostic/predictive role has to be evaluated. For this aim the application of longitudinal studies is common. In this studies, for each patient the time to occurrence of tumor related events (e.g. local recurrence and distant metastases) and death (together with cause of death) are recorded. The prognostic or predictive role of tumor markers is investigated by statistical methods specific for time dependent events (survival analysis).

Since the high number of papers about prognostic tumor markers, it is likely that a lot of clinicians and pathologists are involved in longitudinal studies for markers evaluation. Most of the study investigators are aware of the importance of the application of adequate statistical approaches to obtain reliable results however, this requires a specific statistical "know how" which is not widespread among researchers with clinical and/or biological training.

The aim of this work is to highlight the major issues involved in the planning of statistical analysis to fulfill the study aims, by choosing adequate modeling strategies, and to correctly interpret the results obtained. The aim will be reached by providing clinical examples and by avoiding the use of formulas which could hamper the understanding without adding useful information. This aim will be attained by providing data from a human breast cancer clinical trial that will be used to show statistical analysis methodology and to interpret and discuss model results. Although dataset will refer to woman breast cancer, the analysis exemplifies statistical topics and problems faced also in veterinary longitudinal studies. Since, several tumor markers are considered both in "human" and veterinary breast cancer studies (see for example Kaszak et al[21]).

First, the statistical analysis of tumor markers needs to be planned in advance to achieve study goals. To better explain the statistical methodological approach to longitudinal studies, the four main steps are summarized as follows.

1) Correct specification of the end-point.

In longitudinal studies the end-point is the time elapsed from the study entry of the patient (e.g. date of starting treatment) to the time of the occurrence of the event of interest. Overall survival is considered the most relevant end-point for the evaluation of "patient centered" treatment efficacy.[37,21] The definition is simple and unambiguous. In addition, "tumor centered" end-points are frequently used; such end-points usually include disease-free survival, progression-free survival, relapse-free survival. They combine groups of events selected among tumor progression, local and distant tumor recurrence, metachronous cancer, severe toxicity, death. Thus, they are called "composite" end-points. The respective time to event is the time elapsed from the beginning of follow-up to the time of the first event occurred. It is worth noting the lack of general agreement among authors about the definition of the events that are included in the studied end-point: this attitude may impede the cross comparison of results of different studies. Thus, the end points should be always accurately specified/defined in longitudinal studies to allow standardization of inclusion criteria among studies.

It must be stressed that it is unlikely that the end-point of interest is observed in every patient included in the study. The times of patients who are alive without any event recorded at the end of the study or who are lost to follow-up without events are "censored" at the date of last clinical information. Therefore, in such cases the main end-point may occur after the date of last collected clinical information, but the time to occurrence is unknown (right censoring).

2) Choice of survival analysis models which are appropriate for the chosen end-points.

Survival analysis assessments are commonly based on Kaplan-Meier curves (compared by log-rank test) and Cox regression models. It must be stressed that these methods are based on strict assumptions and are not correct for every end-point, thus biased results may derive from a wrong modeling strategy. Careful evaluation and assistance by experienced statisticians may aid in the consideration and choice of alternative, better suited methods for specific end points. For example, Kaplan Meyer curves and Cox models are correct for comprehensive end-points, such as death for overall causes, or occurrence of any tumor-related event plus death for any cause. For more restrictive end-points, as for example a main end-point including only tumor relapse (thus, not including death not related to cancer), the occurrence of death without tumor relapse prevents the observation of the main end-point. This is the case of competing risks analysis. More in general, competing risks are said to be present when a patient is at risk of more than one mutually exclusive event, and the occurrence of one of these prevents any other event from happening.[10] A typical situation of competing risk is the analysis for causes of death (classified usually as “related to the disease” or not “related to the disease”). The occurrence of death classified as not related to the disease is a competing risks for death classified as related to the disease and vice versa. An adopted solution is to consider the time to occurrence of competing risk (death without recurrence) as a censored time, and to use Kaplan-Meyer and Cox methods. This is adequate only under some assumption, that is, that the probability of death is independent from the probability of relapse, which could be not always tenable according to clinical experience. In fact, death without tumor recurrence is a “peculiar” kind of censoring because we know the patient will never develop a tumor related event, such as tumor relapse, after death. Therefore, statistical methods specific for competing risks need to be adopted.

3) Inclusion of the marker in the statistical analysis.

158 Several markers are measured by qualitative (ordered) scale or quantitative numerical
159 scale. A widely used approach is to subdivide the scale so that define “high” and “low”, or
160 “high”, “medium” and, “low” risk groups. Subdivision of the measurement scale is
161 performed according to criteria defined in previous studies on similar diseases or, else,
162 when such criteria are lacking, by specific and precise definition clearly provided by the
163 authors. The advantages of grouping are the straightforward interpretation of the results,
164 and the possibility of make more straightforward recommendations on the use of the
165 marker for prognostic/predictive aims. The major disadvantage is the potential lack of
166 prognostic/predictive efficacy. As an example, the ER receptor status can be measured in
167 fmol/mg of cytosolic proteins. The complete prognostic information is the ER content in the
168 original measurement scale. If a cut-off is used, e.g. 10 fmol/mg, to define ER- and ER+
169 classes, this implies the assumption that every ER value within each of the two classes
170 has the same prognostic role. The putative prognostic information of the original
171 measurement scale is no longer considered, and this avoids evaluating whether ER in the
172 original measurement scale could provide more accurate information about the prognosis.
173 Moreover, data driven rules for grouping generate heterogeneous choices which make
174 difficult the cross comparison of study results, due to lack of standardization.
175 *4) Quantification of the prognostic/predictive information provided by the marker in addition*
176 *to variables used routinely in clinical practice*.
177 Assessment and quantification of the value of the new marker should be done by adding
178 the marker in a statistical model in which all the variables having a well-known
179 prognostic/predictive role are included. To such aim the “p-value” corresponding to the
180 marker is not exhaustive. As an example, in the case of marker with a single cut-off, if the
181 Kaplan-Meyer survival curves for the two groups are significantly different, this can be
182 interpreted in the following way: for each follow-up time the proportion of surviving patients
183 in one group is different from that of the other group. But this does not imply that each

subject of one group has different survival probability compared with each subject of the other group. For routine clinical practice it is relevant to evaluate if the marker is able to discriminate subjects with different prognosis. This ability is referred to single subjects rather than groups of subjects, and therefore is not assessed by the p-value, but by specific measures of “discriminant ability”. A p-value lower than 0.05 does not imply a satisfactory discriminant ability of the marker.

METHODS

Selection of the statistical model

For overall survival and relapse-free survival where the composite end-point includes all kinds of tumor relapses and death, Kaplan-Meyer and Cox models are adequate only under the strict assumption that censored times are independent from the time to event. As an example, for overall survival it is assumed that patients alive who are lost to follow-up share the same “future” survival probability with those patients who are still in follow-up. This can be true for patients who are lost to follow-up for reasons not depending on their health status, but it is unlike for lost to follow-up diseased patients.

Also, in existence of competing risks, some patients may be lost to follow-up or be free from any event at the end of the study. The assumption that these patients share the same probability of the event with patients who are still in follow-up is still needed. However, the situation is more complex than this. As an example, lets suppose we are interested in “Tumor relapse-free survival” where only related tumor events are of concern (death not included in the end-point). Some patients may die for unrelated tumor causes before a tumor relapse is observed. The independence between times to death and time to relapse is doubtful: in other terms, if the patient was not deceased, could we suppose he/she will have the same “relapse free probability” as patients with observed relapse?

Given the definition of study end-point, the possible presence of competing risks should be carefully considered and an adequate modeling strategy should be adopted. In presence of competing risks, crude cumulative incidences estimators must be considered instead of Kaplan-Meyer curves. Comparison between cumulative incidences among groups should be performed by the Gray test.[12] Concerning regression models, the Fine and Gray model[8] should be used instead of Cox (see for example: Kim,[22] Satagopan et al,[31] Oyama et al[27]). Competing risks analysis tools are available in statistical softwares such as for example R, STATA and SAS.

Inclusion of a novel marker in statistical analysis

The method and the measurement scale of a marker is decided according to criteria established by the lab responsible according to her/his scientific skill. Let us make some consideration about the compliance between the perspectives of biochemical/pathological laboratory techniques and statistical analysis on the use of a marker. To exploit the maximum potential predictive /prognostic role of the marker it is preferable to maintain its original measure scale. Usually a first exploratory evaluation is performed by “univariate” analysis. This approach is simple when the marker is recorded on a quantitative or qualitative scale with a reduced number of levels, because survival or cumulative incidence curves can be traced for each marker level. By the examination of the curves it is possible to evaluate if some marker levels that show similar prognostic/predictive behavior could be grouped together. This approach is not applicable for markers measured on a quantitative scale with several levels.

To maintain the original measurement scale, a regression model is needed. How does it work? Let us consider the most popular model: the Cox model, which is based on the relationship between log(h(t)), i.e. the logarithm of the hazard of the event at time t (the hazard is the rate of the event per time unit) and marker’s levels. The simplest relationship is “linear” i.e., the increase of log(h(t)) for each increase of x units of the marker is the

same for each marker level. For example, let us suppose a marker M can assume levels from 0 to 10: the linear relationships implies that the ratio between the hazards for levels 4 and 5 is the same than the ratio between the hazards of levels 8 and 9 (Fig. 1A). By taking marker level 0 as reference level, it is possible with the Cox model to calculate the relative hazard of each marker level with respect to the reference one. For example, in Fig. 1B, the hazard between levels 5 of the marker and the hazard of the level 0 of the marker is 2.0, and the hazard between level 10 and the hazard of the level 0 is 4.0.

The linear relationship is simple and "user friendly" but it does not always fit the "real world". For some markers, a saturation effect is expected: the increase of log(h(t)) for x units of the marker decreases with the increasing of the marker level, so it no longer constant. For example, in Fig. 1C the ratio between the hazards for levels 2 and 1 is 1.7, which is not the same as the ratio between hazards of levels, e.g., 6 and 5, which is equal to 1.4. By taking marker level 0 as reference level, the ratio between hazards of marker levels 2 and 0 is equal to 2.0, whereas the ratio between hazards of marker levels 6 and 0 is 2.7 (Fig. 1D). Effects more complex than those discussed above could occur and be difficult to interpret.

Since the shape of the relationship between h(t) and marker's level provides insights about the role of the marker on disease progression dynamics, the convenience to categorize marker levels to create risk groups has to be evaluated with care. To provide an alternative to the use of empirical cut-off rules which are not based on the prognostic propensity of the marker (e.g. median or other percentiles of the distribution), statistical procedures for "best" cut-off selection have been proposed (e.g. Faraggi, and Simon,[7] Hilsenbach and Clark,[18] Mazumdar et al[24]). Nevertheless, methodological papers on cancer journals advised against the best-cut-off use mainly for the risk of missing prognostic information or unreliable results (e.g. Altman et al,[2] Altman,[1] Holländer and Schumacher[19]). In addition,

the adoption of user defined cut-off values was criticized also on methodological statistical papers.[30]

Results of univariate analysis are not sufficient to make conclusions on the usefulness of the marker, multivariate analysis is needed to estimate its adjusted effect when other clinical and pathological variables are taken into account.

Quantifying the added prognostic/predictive information provided by the marker

The "statistical significance" of the marker is not the main criterion to be adopted to assess significance. In fact, a statistically significant result does not imply a clinically relevant result and vice versa. It is easy to obtain a statistically significant result for an irrelevant prognostic impact of the marker if a large dataset is analyzed. Conversely, it is not easy to obtain statistically significant results for a clinically relevant prognostic impact of a marker if the sample has a small size. A statistically significant result depends on the power of the statistical test which, in turn, depends not only on the prognostic impact of the marker but also on sample size.

Part 1: sample size considerations

If statistical significance is retained as a relevant criteria for the initial evaluation of the contribution of a marker, the sample size needs to be considered with care. Sample size depends on the level of significance (usually 5%) but also on the power of the test (i.e. the probability of obtaining a statistical significant result when the marker is effectively prognostic in the population of patients to which the sample refers). Usually the power of the test is fixed equal or above 80%.

A key issue is the minimal amount of prognostic impact considered clinically relevant to be detected by the test. For sake of simplicity, let us consider a marker classified into two classes (low and high), and survival curves compared by the log-rank test. After defining statistical significance and power, the information needed is the hazard ratio to be detected (e.g. the ratio between the hazard of end-point of patients with high marker levels

and the hazard of end-point of patients with low marker levels). For example, if a hazard ratio of death of 2 is to be detected, with a significance level of 5% and power of 80%, the total number of deaths to be observed is 56. The sample size depends on this number of events and on the proportion of deaths expected in each group. If in the low marker level group the 20% of patients is expected to die, and 40% of patients are expected to die in the high marker level group, the sample size for each of the groups is 93.

From this example it may be noted that several key information is needed for sample size calculation. The responsibility of clinicians is to provide reliable information about the hazard ratio to be detected and by the proportion of events of interest expected in the two groups. The responsibility of the statistician is to apply correct methods and formulas for sample size calculation.

When the marker is novel it is very difficult to provide reliable information for sample size, and “rule thumbs” may be adopted to perform regression analysis. These rules are based on a quantity defined as event per variable (EPV) ratio and suggest that the maximum number of variables that can be included in the regression model depends on the number of events observed in the sample. The most frequently used rule is that the EPV ratio is equal to ten.[4] In such case if, for example, 50 events are observed then 5 binary variables can be considered (including the marker).

Even in this case the number of events play a key role, and the number of patients required depends on the probability of events in the study population. For a disease with in general a good prognosis (low event probability) a very high sample size will be required. For example, if a 10% of probability of event is expected, to include five binary variables in the regression model, the minimal sample size will be 500 patients.

Part 2: statistical procedures

Clinical and pathological variables which are recognized to be prognostic/predictive factors usually are collected in routine clinical practice as an aid to clinical decision-making

process. Is the availability of information on marker level useful to improve treatment planning? The answer is to evaluate the additional prognostic/predictive contribution of the marker to that provided by the other variables. For this issue, the results of multivariate analysis must be considered.

The evaluation of the prognostic usefulness of the marker in clinical routine practice should be based on the ability of the marker to discriminate patients with different outcomes. A regression model (e.g. Cox) including marker level is performed and for each patient the model's predicted outcome is compared with the observed one. A measure of discriminant ability is the area under ROC curve (AUROC). It is customary that higher marker values are associated to worst prognosis. The AUROC represents the probability that, for a random pair of patients, the patients who has the shorter time to event (worst outcome) has also the higher marker level. In the case of optimal discriminant ability AUROC is equal to 1. AUROC equal to 0.5 indicates the lack of discriminant ability, in fact prediction is like a coin flip. An AUROC measure appropriate for time to event data has to be used, .e.g. Harrell's c statistic.[35] The Harrell's c statistic provides a unique measure on the whole study duration. When both marker levels and individual outcome status change with follow-up time a useful information to investigate could be the minimum follow-up time useful for outcome prediction. To such aim, time dependent AUROC measures can be adopted.[16,20]

In the case of multivariate analysis the marker is included in the model together with other clinical and pathological variables, thus for each patient, model prediction is based on the joint effect of all variables, and the additional contribution of the markers is not highlighted. To such end a naïve method is to estimate AUROC by the model with all the variables but the marker (reduced model) and to compare this AUROC with that of the model also including the marker (full model). The greater the difference between AUROCs of full and reduced model and the greater will be the added discriminant contribution of the marker. It should be stressed that if the observed difference result "negligible" this does not imply a

negligible discrimination improvement. Because of this limitation a more structured approach is the integrated discrimination improvement index.[28] Integrated discrimination improvement values range between 0 (no discrimination improvement) and 1 (maximum discrimination improvement) and the more is the index near to 1 and more the contribution of the marker to discrimination will be.

STRATEGY OF ANALYSIS AND RESULTS: APPLICATION TO THE DATASET OF NODE POSITIVE BREAST CANCER PATIENTS TREATED WITH CHEMOTHERAPY

We used public data made available by the German Breast Cancer Study group: a description about the dataset structure can be found, among others, in Sauerbrei and Royston.[32] The dataset used for the application of the statistical analysis in this paper is available at the following web site: ftp//ftp.wiley.com/public/sci_tech_med/survival. These data were recorded from a multi-center randomized trial on lymph-node positive breast cancer with the primary aim of evaluating recurrence-free and overall survival between three chemotherapy regimens. The dataset consists of 686 records of patients with complete information about major prognostic variables. To apply the statistical methods two end points were considered: 1) death (for all causes) and 2) tumor recurrence. Tumor recurrence was defined as a composite end-point including the first occurrence of either loco-regional or distant recurrence, contralateral tumor, and secondary tumor.

The analysis described in the following sections has been performed only for illustrative purposes, with no intention to provide clinically reliable results. The Authors are aware that to perform an exhaustive prognostic/predictive analysis on human breast cancer, data need a much more complex modeling strategy and the consideration of a larger number of variables. Several analyses can be found in the literature according to prognostic/predictive aims. But this is not the aim in the present report, so a restricted set of variables and only one marker will be considered to illustrate the methodology. Only a

subset of variables will be considered: a group to represent known prognostic factors (tumor size, tumor grade, number of nodes involved) and one, the amount of Estrogen Receptors representing the prognostic marker to be evaluated. Hormone receptor content was measured macroscopically by a dextran-coated charcoal method.[33] The marker was recorded as quantitative numerical scale defined in femtomoles and specifically in fmol/mg of protein. Because patients have been submitted to chemotherapy and not to hormonal therapy, for methodological purposes, we have considered the analysis as prognostic rather than as predictive.

For sake of simplicity, the number of nodes involved will be classified according to (1-3,>3 and <10, >=10), tumor size as T_1 (<=20 mm), T_2 (>20 mm but < 50 mm), T_3 (>50 mm). Concerning the first end-point the following analysis will be performed: the univariate analysis of the marker, firstly dichotomized according to the cut-off reported in the original trial paper (20 fmol), then by data driven best-cut-off, and finally considered as a continuous variable. The prognostic impact of the marker will be evaluated by “p-value”, Harrell’s c statistics, and AUROC during follow-up. Multivariate analysis will be performed considering a model with all the above-mentioned prognostic factors and the marker. The added contribution to discriminant ability of the prognostic marker will be evaluated.

Concerning the second end-point (tumor recurrence) the issue of the competing risk effect due to death without local recurrence will be considered and crude cumulative incidence estimators and regression models for competing risks will be applied, showing the difference with naïve analysis which ignores competing risks. The evaluation of discriminant ability will no longer be showed because the interpretation in the case of competing risks analysis is like that discussed in the analysis of overall survival.

All analyses have been performed with the software R release 3.6.2,[29] with the packages survival,[34] cmprsk,[11] rms,[15] survivalROC[16] and survIDINRI[13] added.

Analysis of time to death (for all causes)

In this paragraph we illustrate analyses to assess the impact on overall survival of ER content as an example of "novel" marker. Standard Cox models and Kaplan-Meyer methods can be used in this case, because the only possible source of censoring is the loss to follow-up and patients alive at the end of the study follow-up period.

Univariate analysis of ER content and time to death

The Kaplan-Meier survival curves when the 20 fmol/mg cut-off was used (ER- if estrogen content <20 fmol/mg and ER+ if content >=20 fmol/mg) are represented in fig.2. A marked difference between the two groups is shown with a significant better prognosis for ER+ (log-rank test= 27.3 p<0.001).

To find the optimal cut-off by the data driven method, a cut-off sequence starting from 5 to 200 fmol/mg was considered, and the cut-off corresponding to the minimum p-value was chosen. According to this criterion the best cut-off was 10 fmol/mg (log-rank test =33.2 p<0.0001). Results seem to be reproducible since this cut-off has been previously identified (e.g. Courdi,[6] Nicholson et al,[25] Andersen et al[3]).

The survival curves obtained by the old and new "best" cut-off are illustrated in Fig. 3. In the comparison, curves for ER+ patients are superimposable while there is a difference in ER - patients with a slight worse prognosis for ER<10 patients.

To consider ER as a continuous variable, a naïve approach is to include ER in a Cox model according to linear relationship with the following results: Hazard Ratio= 0.9985, 95%, confidence interval: from 0.9972 to 0.9998 p-value 0.0281. This finding indicates that prognosis improves with the increasing of ER fmol concentration, for each increase of 1 fmol/mg of ER. Is this result clinically reasonable? To give an answer, the first step is to examine the distribution of the variable. Range: from 0 to 1144 fmol: median=36, mean=96.25, Q_1=8, Q_3=114. The difference between mean and median suggests an asymmetrical distribution of the variable that is clear in Fig. 4A. The distribution of ER concentration is heavily asymmetrical and 72% of patients have ER <=100 fmol. It is likely

that a difference of 1 fmol is more clinically relevant when ER has low values than when ER has high values.

In this situation, the application of a data scale transformation should be preferred in order to: 1) attribute more weight to small differences in fmol starting from low ER values than to high ER values and 2) reduce the spread of the ER values.

A widely diffuse transformation is performed via logarithmic scale. Since some patients have 0 fmol ER recorded in the dataset, an empirical solution that can be adopted is log(ER+1). This scale transformation satisfies both requirements. For the requirement 1) as an example the difference of 5 fmol from 5 to 10 in logarithmic scale is Log(10)-Log(5)=0.693 and from 100 to 105 is Log(105)-Log(100)=0.049. For the second requirement, see Figure 4B.

The prognostic relationship is now evaluated by including ER in log scale (LER) into the Cox model. First, the simplest analysis: a linear relationship. Hazard ratio= 0.81, 95% confidence interval= 0.75-0.87 (p-value<0.0001). These results mean that for each increase of 1 unit LER the ratio between hazard of death for LER=x and the hazard of death for LER=x+1 is estimated to be 0.87 and does not change for each pair of values LER and LER+1. Thus, for example the ratio between the hazard of death for LER=0.69 (about 1 fmol) and LER=1.69 (about 4 fmol) is 0.87 and the ratio between the hazard of death for LER=2.40 (about 10 fmol) and LER=3.40 (about 29 fmol) is 0.87, and so on. To facilitate the evaluation of model results, it is preferred to represent the estimated hazard ratios in the original measurement scale by considering as reference the lowest ER value. In Fig. 5 is shown the shape of the ratio between hazard of death for each ER fmol value and the hazard of death for 0 fmol. The decrease in hazard ratio is steeper for low ER values than for higher ones. For example, the hazard ratio of death from 0 to 10 fmol is 0.52, from 0 to 20 fmol is 0.45, from 0 to 50 fmol is 0.37 and from 0 to 100 fmol is 0.32.

A relevant issue to be analyzed is how much a researcher is confident with a linear relationship. When the linear relationship seems to be too “restrictive”, to address this question the possible existence of a more flexible relationship needs to be investigated, for example by including into the Cox model power functions of LER, such as polynomials or fractional polynomials[32] or spline functions (Harrell et al,[14] Heinzl and Kaider[17]).

As a matter of fact, after including cubic spline functions, a more complex functional relationship than a linear one was found. The comparison of model prediction with spline and model prediction with linear relationship is shown in fig. 6. The difference is a slight increase of the Hazard Ratio from 0 to 2 fmols, and after 200 fmol for the model with spline function whereas in the model with linear relationship the hazard ratio always decreases with the increasing of fmol.

When the models are compared, the linear relationship model results in a likelihood ratio test= 27.34 $p=2*10^{-7}$ and model with spline function results in a likelihood ratio test= 34.64, $p=1*10^{-7}$. Based on the p value, the second model seems to be better. However a lower p-value for the most complex model cannot be used as a criterion to decide which model better represents the prognostic behavior, thus the more complex model can be preferred over the simplest one only if the shape of the hazard shown in Fig. 6 has a credible clinical/biological explanation.

If the aim is to predict outcome, the discriminant ability of the two models (AUROC) should be considered. First, the measure by Harrell’s c statistic for the AUROC on the whole follow-up is equal to 0.634 and to 0.633, respectively for the model with linear relationship and the model with spline functions. The two models provide the same discriminant ability (i.e. about 63% of patients who have longer survival times have higher ER values than patients who have shorter survival times) thus, according to this perspective it seems useless to complicate the model with a spline function. More detailed information can be obtained by dynamic ROC curve which provides cumulative AUROC for selected follow-up

467 times. This allows to investigate the possible time in which to assess the patients for the
468 better model discriminant ability. Dynamic ROC curve for the linear and spline Cox
469 regression models are reported in Fig, 7. Subdividing follow-up time in 180 days intervals,
470 for the model with spline function the highest AUROC value is 0.72 at 180 days and for the
471 model with linear relationship is 0.70 at 360 days. These results may suggest that better
472 ER discriminant is shown at short follow-up times.
473 After 900 days the discriminant ability of the model with spline function is fairly better than
474 that of the model with linear relationship. The maximum of the discriminant ability of ER
475 when considered as dichotomous (cut-off 20 fmol) is at 360 days with AUROC=0.65 which
476 is lower than that obtained when ER is considered in a continuous scale (LER)
477 *Multivariate analysis of ER content and time to death*
478 The first model includes tumor size, number of axillary lymph nodes, tumor grading and
479 LER (log (ER+1). The LER scale was considered for the same reasons discussed in the
480 previous paragraph. Results of the Cox model are reported in Table 1.
481 When the joint prognostic effect of the variables is considered, ER is a statistically
482 significant prognostic factor, as well as grading and axillary lymph nodes but not tumor
483 size. The three pathological variables are categorical (3 classes) and have to be included
484 into Cox model as “dummy variables”; typically, such variables assume only the values 0
485 and 1. For a categorical variable with 3 classes, two dummy variables are needed. One of
486 the classes is chosen as the reference and Cox model estimates the ratio between the
487 hazard of death of each of the two remaining categories and the hazard of death of the
488 reference one. E.g. for grading, the chosen reference category is grade I thus, the hazard
489 of death for patients with grade II is 2.57 times the hazard of death for patients with grade I
490 and the hazard of death for patients with grade III is 3.63 times the hazard of death of
491 grade I patients. Corresponding P-value tests the null hypothesis of hazard ratio=1 (i.e. no
492 difference between the hazard of death for Grading II (or III) and grading I). Both hazard

ratios are significantly different from 1 thus, grading has a significant prognostic effect. The same was shown for axillary lymph nodes but not for tumor size. Together with p-value, 95% Confidence Interval provides relevant information about the value of the hazard ratio that we would find if the whole population of patients were examined. E.g. if the whole population of node positive breast cancer patients submitted to the same chemotherapy were examined and patients with Grade III tumors were compared with patients with Grade I tumors the hazard ratio of death is expected (with a probability of 0.95) to lie between 1.54 and 8.60.

In the previous analysis LER was included as linear effect. Now the question becomes: Is there evidence also in multivariate analysis for a more complex relationship? The inclusion of splines does not suggest any improvement over the previous model. These results can be interpreted as that the complex relationship in univariate analysis may be attributable to the lack of adjustment for other known prognostic factors. This is one of the reasons to evaluate the marker by multivariate analysis, considering other prognostic factors which are likely associated to the marker itself.

Concerning the discriminant ability, the AUROC on the whole follow-up for the multivariate model was Harrell's c statistic =0.731. For marker evaluation the main question is: does ER improve the discriminant ability when added to the other variables? Harrell's c statistic for the model with the prognostic variables and without ER is 0.71. Thus, it seems that the contribution of ER to the discriminant ability of the three prognostic factors is limited.

Because in univariate analysis (previous paragraph) it emerged that the best discriminant ability of ER was shown at early follow-up times, this evaluation was performed also in the multivariate analysis. Again, the highest discriminant ability was found at early follow-up times: 360 days (Fig. 9) and model with ER slightly outperforms the model without the variable (AUROC=0.85 vs AUROC=0.83).

The improvement in discriminant analysis obtained by including in the model ER (as LER) can be evaluated by “integrated discriminant index (IDI) which, at 360 days is 0.0026. The IDI is near to zero indicating a low discriminating improvement. The 95% Confidence interval (from -0.0020 to 0.0103) includes 0 thus there is not “statistical evidence” of a discriminant improvement in prognosis given by ER (in this type of chemotherapy treated patients) when Grade, tumor size and number of axillary lymph nodes are jointly considered.

When ER is considered as dichotomous (cut-off 20 fmol) the highest discriminant ability is again at 360 days and AUC is 0.84, slightly lower than AUROC for the model with ER in continuous scale (LER).

Analysis of impact of ER on time to tumor recurrence

In this paragraph we show the impact of ER on tumor recurrence: as discussed in the methods section, this end-point requires methods for competing risk analysis, because of the presence of death occurrence without relapse, as the competing event preventing the observation of the end-point of interest.

Univariate analysis of ER and tumor recurrence

In the current example, among 171 patients who are dead, 21 had not experienced tumor recurrence. For this end-point, tumor recurrence-free survival curve interpretation is not straightforward because the probability of being free from tumor recurrence is the sum of two probabilities: the probability of being alive without tumor recurrence, plus the probability of being deceased without recurrence. For this reason, the probability of concern is the cumulative probability of tumor recurrence as first event (crude cumulative incidence). Sometimes, a naïve estimate of this probability is mistakenly obtained as the complement to the Kaplan-Meyer estimate of tumor recurrence-free survival, after considering time to death without recurrence as censored. For the data under examination,

543 the crude cumulative incidences and the naïve incidence obtained by 1-Kaplan-Meier are
544 shown in Fig. 9.
545 First let us consider ER as a categorical variable by cut-off 20 fmol/mg. In Figure10, the
546 patterns of the cumulative incidence obtained by the two methods are similar and slight
547 differences can be evidenced only at follow-up times greater than 1500 days. This result is
548 expected in our case since the low number of patients who died without tumor recurrence.
549 In other situations where a higher number of competing events is observed, more
550 substantial differences between the two estimates are expected. Furthermore, the two
551 estimates are nevertheless interchangeable. The naïve Kaplan-Meier estimate is a biased
552 estimate of the probability of recurrence given the “removal” of death without recurrence,
553 i.e. the cumulative probability of recurrence if this could be observed for all patients. The
554 crude cumulative incidence estimator is the unbiased estimate of the cumulative
555 probability of recurrence observed as first event.
556 Concerning the comparison between crude cumulative incidences of recurrence for
557 patients with ER- and ER+ status, a significant difference was found (Gray test= 12.97 p-
558 value=.0003164462): thus, a higher incidence of recurrence is expected in ER- patients.
559 To find the optimal cut-off. ER was dichotomized by a cut-off sequence starting from 5 to
560 200 fmol and the cut-off corresponding to a minimum p-value was chosen. According to
561 this criterion the best cut-off was 9 fmol/mg (p-value of Gray test = $5.148984*10^{-7}$. Results
562 are near to the cut-off 10 found for overall survival.
563 *Multivariate analysis of ER and tumor recurrence*
564 The analysis was performed by estimation of two models: The Fine and Gray regression
565 model for competing risks, and, for comparison purposes, a Cox model for recurrence
566 times, in which times to death without tumor recurrence are censored. Results are
567 reported in Table 2. Although the slight differences among Hazard ratios (again, this result
568 was expected because of the low number of deaths without recurrence) the interpretation

of the results of the two models are different. Both models account for the presence of competing risks but from a different perspective. For the Fine and Gray regression model, if the (sub-distribution) hazard ratio is significantly different from one, then the crude cumulative incidences (for example, between Grade III versus Grade I, see Tab. 2) are different. This relationship cannot be extended to Cox model results because hazard ratio from Cox model does not have a direct relationship with crude cumulative incidences. From the estimates of hazard ratios (Table 2), a significant impact of Estrogen Receptor levels (included in log scale) emerged, with an estimated hazard ration of 0.90 (95% C.I. 0.84,0.97). This result suggests that the hazard of tumor recurrence decreases with increased levels of ER, and, consequently the crude incidence of recurrence is lower in subjects with higher ER levels.

DISCUSSION

This manuscript illustrates, utilizing a human database, some of the most appropriate statistical approaches to analyze prognostic significance of any novel tumor marker, stressing the necessity to plan in advance a statistical approach tailored to the clinical study and providing insights on study planning to provide statisticians with the most useful and adequately numerous dataset. Oncologists and clinicians in general should take into consideration before starting the study on a new marker several matters including: correct specification of the end-point, choice of the best suited survival model for the end-point, qualitative or quantitative measurement scale of the marker, and statistical methods aimed at quantifying the prognostic/predictive information provided by the marker.

Most of the problem that are spotted by a statistician when she/he is consulted, after the end of a study, in order to improve the paper to submit to a scientific journal, or to clarify some technical issues about the statistical analysis, are:

- Lack of representativeness of the sample with respect to a wider population of subjects sharing the same pathology, due to an inadequate sampling plan.
- Data retrieved from medical records (retrospective studies) with insufficient quality of data in order to satisfy the aims of the study.
- Inadequate (small) sample size to the aim of investigating the prognostic value of the variables under examination.
- Choice of cut-offs for numerical markers based on empirical basis without investigation of the marker on the original measurement scale.
- Inadequate statistical methods of analysis that strongly reduce the reliability of results.
- A blind interpretation of p-values that makes statistical significance prevail over the most important clinical relevance.
- Lack of evaluation of the discriminant ability when the statistical model is used as an aid to clinical decision making.

The evaluation of prognostic/predictive tumor markers is challenging and should aim toward a personalized medicine framework, intended to improve clinical decision making. Because of the potentially relevant role of a marker, the statistical analysis needs to be performed in such a way to obtain reliable information. For this purpose, the end-point has to be clearly defined and its choice depends on the study aim, that is, and end point can be “patient oriented” or “tumor oriented”. The most utilized patient oriented end-point is patient overall survival or patient’s quality of life, representing composite end-points in which many events are included (e.g. tumor recurrences and death). The tumor oriented end-point relates generally to the response of the tumor to therapeutic strategies, and different specific end-points are of concern, as for example local relapse, end/or distant metastases and/or contralateral tumors. Since each one of the above mentioned end-points provides different information on the disease course, it is usual and highly recommended to plan the study by considering both patient oriented and tumor oriented

end-points. The latter are only a subset of the events which can be observed and should be planned to take into account the presence of competing risks.

When the goal is to evaluate the prognostic/predictive role of a marker, a multivariate analysis is the adequate strategy. In the model exemplified in this report, all the previously well-known prognostic predictive clinical and pathological variables should be included in such a way to quantify the added contribution provided by the marker, and to allow clinicians to decide whether to include the marker in their routine practice in costs/benefits terms. As an aid to decision, the criteria based on statistical significance are not sufficient and the discriminant ability should be provided in addition.

Number of events is one of the main critical issues in this type of statistical analysis, because an insufficient EPV could determine lack of reliability of multivariate analysis results. In fact, the more variables needed to be included the larger sample size is needed. Methodological papers showed that at least 10 events for each variable should be considered to obtain reliable model results.[5,36] As a consequence, when a low number of events are expected to occur for the disease of interest (e.g. low incidence of tumor recurrence or deaths) the adequate sample size may need to be very large and thus, difficult to reach by a single research center.

The authors believe that the role of each study is to contribute to the scientific background enhancement, and to this aim a study should be correctly conducted non only regarding the experimental components (clinical, biological, pathological) but also with an appropriate statistical analysis. The role of the variables on the disease course is often complex, but it seems that most researchers fail to realize that unreliable results could be obtained by the application of a too simplified statistical approach, eventually adopted by honest researchers who, however, are not experienced in statistical methods. On the other hand, statisticians who are not experienced in medicine could apply complex statistical methods which are inadequate to study aims. In conclusion, the best strategy is to work in

close collaboration with each group providing the study with its own expertise, and learning how to communicate effectively to explain technical issues by using terms and examples which can be understood by each research staff component. We hope this manuscript will facilitate the cooperation among bio-statisticians, oncologists and pathologists.

REFERENCES

1. Altman DG. Suboptimal analysis using 'optimal' cutpoints. *Br J Cancer.* 1998;78(4):556.

2. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using ‘optimal‘ cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst*. 1994;86(11):829-35.

3. Andersen J, Thorpe SM, King WJ, Rose C, Christensen I, Rasmussen BB, et al. The prognostic value of immunohistochemical estrogen receptor analysis in paraffin-embedded and frozen sections versus that of steroid-binding assays. *Eur J Cancer*. 1990;26(4):442-449.

4. Bell EH, Pugh SL, McElroy JP, Gilbert MR, Mehta M, Klimowicz AC, et al. Mocular-based recursive partitioning analysis model for glioblastoma in the temozolomide era: a correlative analysis based on NRG oncology RTOG 0525. *JAMA Oncol.* 2016;3(6):784-792.

5. Concato J, Peduzzi P, Holford, Feinstein AR. Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy. *J Clin Epidemiol*. 1995;48(12):1495-1501.

6. Courdi A. Prognostic value of continuous variables in breast cancer and head and neck cancer. Dependence on the cut-off level. *Br J Cancer*. 1988;58(1):88.

7. Faraggi D, Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Stat Med.* 1996:15(20);2203-2213.

8. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94(446):496-509.

9. Fiteni F, Westeel V, Pivot X, Borg C, Vernerey D, Bonnetain F. Endpoints in cancer clinical trials. J Visc Surg 2014;151(1):17-22.

10. Gichangi A, Vach W. *The analysis of competing risks data: A guided tour*. Odense, Denmark: University of Southern Denmark; 2005.

11. Gray B. *cmprsk: Subdistribution Analysis of Competing Risks. R package version 2.2-10*. 2020. https://CRAN.R-project.org/package=cmprsk

12. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat*. 1988:1141-1154.

13. Hajime U, Tianxi C. survIDINRI: IDI and NRI for comparing competing risk prediction models with censored survival data. R package version 1.1-1. (2013). https://CRAN.R-project.org/package=survIDINRI

14. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15(4):361-387.

15. Harrell FE. rms: *Regression Modeling Strategies. R package version 6.0-0.* 2020. https://CRAN.R-project.org/package=rms

16. Heagerty PJ, Saha-Chaudhuri P. *survivalROC: Time-dependent ROC curve estimation from censored survival data. R package version 1.0.3*. 2013. https://CRAN.R-project.org/package=survivalROC

17. Heinzl H, Kaider A. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Comput Meth Prog Bio* 1997;54(3):201-208.

18. Hilsenbeck SG, Clark GM. Practical p-value adjustment for optimally selected cutpoints. *Stat Med.* 1996;15(1):103-112.

19. Holländer N, Schumacher M. On the problem of using 'optimal' cutpoints in the assessment of quantitative prognostic factors. *Onc Res Treat*. 2001;24(2);194-199.

20. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med Res Methodol*. 2017;17(1):53.

21. Kaszak I, Ruszczak A, Kanafa S, Kacprzak K, Król M, Jurka P. Current biomarkers of canine mammary tumors. *Acta Vet Scand*. 2018;60(1):1-13.

724 22. Kim HT. Cumulative incidence in competing risks data and competing risks regression
725 analysis. *Clin Cancer Res*. 2007;13(2):559-565.
726 23. Liu CJ, Lin SY, Yang CF, Yeh CM, Kuan AS, Wang HY, et al. A new prognostic score
727 for disease progression and mortality in patients with newly diagnosed primary CNS
728 lymphoma. *Cancer Med*. 2020;9(6):2134-2145.
729 24. Mazumdar M, Smith A, Bacik J. Methods for categorizing a prognostic variable in a
730 multivariable setting. Stat Med. 2003;22(4):559-571.
731 25.Nicholson S, Sainsbury JRC, Needham GK, Chambers P, Farndon JR, Harris AL
732 Quantitative assays of epidermal growth factor receptor in human breast cancer: cut-off
733 points of clinical relevance. *Int J Cancer*. 1988;42(1):36-41.
734 26. Oldenhuis CNAM, Oosting SF, Gietema JA, De Vries EGE. Prognostic versus
735 predictive value of biomarkers in oncology. *Eur J Cancer*. 2008;44(7):946-953.
736 27. Oyama MA, Shaw PA, Ellenberg SS. Considerations for analysis of time-to-event
737 outcomes subject to competing risks in veterinary clinical studies. *J Vet Cardiol*.
738 2018;20(3):143-153.
739 28. Pencina MJ, D'Agostino SRB, D'Agostino JRB, Vasan RS. Evaluating the added
740 predictive ability of a new marker: from area under the ROC curve to reclassification and
741 beyond. *Stat Med.* 2008;27(2):157-172.
742 29. R Core Team. *R: A language and environment for statistical computing. R Foundation*
743 *for Statistical Computing*, Vienna, Austria. 2019. URL: https://www.R-project.org/.
744 30. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple
745 regression: a bad idea. *Stat Med.* 2006;25(1):127-141.
746 31. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, & Auerbach AD. A note
747 on competing risks in survival data analysis. *Br J Cancer*. 2004;91(7):1229-1235.

32. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J Roy Stat Soc A Sta*. 1999;162(1);71-94.

33. Schumacher M, Bastert G, Bojar H, Huebner K, Olschewski M, Sauerbrei W, et al. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *J Clin Oncol.* 1994;12(10):2086-2093.

34. Therneau T. _A Package for Survival Analysis in S_. version 2.38, 2015. <URL:https://CRAN.R-project.org/package=survival>.

35. Vergouwe Y, Steyerberg EW. Validity of Prognostic Models: When Is A Model Clinically Useful? *Seminars Urol Onc*. 2002;20(2):96-107.

36. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. 2007;165(6):710-718.

37. Wilson MK, Karakasis K, Oza AM. Outcomes and endpoints in trials of cancer treatment: the past, present, and future. *Lancet Oncol*. 2015;16(1):e32-e42.

38. Winick N, Devidas M, Chen S, Maloney K, Larsen E, Mattano L, et al. Impact of initial CSF findings on outcome among patients with National Cancer Institute standard-and high-risk B-cell acute lymphoblastic leukemia: A report from the Children's Oncology Group. *J Clin Oncol* 2017;35(22):25-27.

FIGURE LEGENDS

Figure 1: theoretical hazard ratios for a linear and a non-linear relationship between hazard and marker level.

Marker levels are placed on the X axis. Panels A) and B) : linear relationship. A: Ratio between the hazard for a marker level x and the hazard for marker level 0; B: Ratio between the hazard for a marker level x+1 and the hazard for marker level x. Panels C) and D) : non-linear relationship. C: Ratio between the hazard for a marker level x and the hazard for marker level 0; D: Ratio between the hazard for a marker level x+1 and the hazard for marker level x.

Figure 2: Kaplan-Meier survival curve for ER

(ER- if fmol/mg<20 and ER+ if fmol/mg>=20)

Figure 3: Kaplan-Meier survival curve for ER for old cutoff

(ER- if fmol/mg <20 and ER+ if fmol/mg>=20) and new (“better”) cut-off (ER- if fmol/mg <10 and ER+ if fmol/mg >=10)

Figure 4: histogram of the ER distributions

Panel A: original scale (fmol/mg); panel B: log(ER+1) scale

Figure 5. Ratio of the hazard of death for each fmol ER and the hazard of death for 1 fmol ER

Figure 6 The ratio of the hazard of death for each fmol ER and the hazard of death for 1 fmol ER Fi

Gray line: model with regression spline; Black line: model with linear relationship.

Figure 7 Discriminant ability by dynamic ROC curve.

Black line: model with linear relationship. Gray line: model with regression spline.

Figure 8. Discriminant ability by dynamic ROC curve.

Gray line: model with axillary lymph nodes, grading and tumor size. Black line: model with the three pathological variables plus LER.

Figure 9. Cumulative incidence of tumor recurrence for ER (ER- if fmol<20 and ER+ if fmol>=20).

Solid lines crude cumulative estimates, dashed lines: naïve Kaplan-Meier estimates. Black lines ER-, Gray lines ER+.