46th International Society of Oncology and Biomarkers (ISOBM) Congress, 12-17 October 2022, Bled, Slovenia

Logistic Regression Using a Combination Of CA-62, CA15-3, and Age for Detecting Early Stages Of Breast Cancer In Screening

Evgueni Klinski[1] and Ricardo Moro[2]
1 - UCM Technologies Inc., Toronto, Ontario, Canada
2 - Pacific Biosciences Research Centre Inc., Richmond, British Columbia, Canada, ricardoip@yahoo.com

BACKGROUND

Binary Logistic regression estimates the probability of an event occurring - such as having cancer - based on one or ***multiple*** independent variables which can be continuous (e.g., ng/mL of a circulating biomarker), dichotomic (e.g., anorexia/no anorexia) or categorical (e.g., age group). This allows (a) to combine markers and (b) to improve the diagnosis accuracy by adding other relevant clinical or laboratory data. For example, given a certain serum concentration of a biomarker, an older patient with anemia and anorexia is more likely to have a malignancy than a younger patient with no anemia or anorexia. This estimation is usually done in a subjective manner by the physician evaluating the case. Logistic Regression provides a powerful statistical tool to weighing all relevant parameters and estimating mathematically the probability of that individual having cancer.

Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Ln(Odds) , where Odds=p/(1-p):

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Eq. 1

where log is the natural logarithm, β_0 is the intercept and β_1, β_2… β_m are the best fit slopes for each variable, and $x_1…x_m$ are the measured values (be that categorical, binary or continuous).

The beta parameter or coefficient in this model is commonly calculated using the maximum likelihood computation. This method tests different values of β through multiple iterations to obtain the best fit of log odds. These iterations serve to find the maximum log likelihood function to determine the best parameter value. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability.

The outcome is measured as the probability of a binary outcome (e.g., dead or alive, cancer or no cancer) given the variables (or predictors) used. Since the model can compare the predicted results with the actual condition, we can also generate a ROC curve, only that the Specificity and Sensitivity stem from the combination of all the variables used in the model.

Finally, the model can also assess the significance of each variable or predictor used. For example, in combining cancer markers, the value of the markers is highly significant (very low p) whereas age might be not that important ($p > 0.05$), in which case, we can exclude age as a variable.

When we have more than two different unrelated categories, for example the histology is classified into 3 categories; Cancer, Benign lesion and Healthy, rather than in a binary classification (Cancer/No cancer) we need to use Multinomial Logistic Regression. The interesting feature of Multinomial Logistic Regression is that we can now determine at the same time the probability of different outcomes. For example, we can use Cancer, Benign and Healthy and calculate, in one shot, the probability an individual has of bearing each one of those conditions from the set of variables in the model. This is also important for the physician and the patient, but not so much in terms of validation of the model since it does not generate ROC curves .

In a previous communication (Diagnostic efficacy of the new prospective biomarker's combination CA 15-3 and CA-62 for early-stage breast cancer detection: results of the blind prospective-retrospective clinical study. Tcherkassova J. et al. Cancer Biomarkers v.35 (2022) p. 57–69), the cutoff value for each marker was set so that there would be no false positives (100% specificity). Then, the positive cases for each marker were combined to calculate the sensitivity, which was 75% as shown in Figure 1.

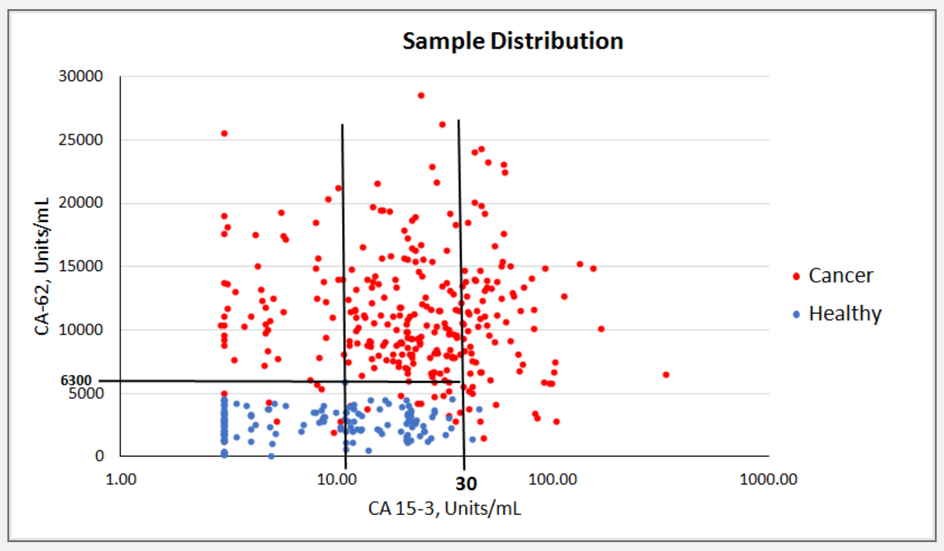

Figure 1: Cutoffs from J. Tcherkassova et al. Cancer Biomarkers v.35 (2022) p. 57–69

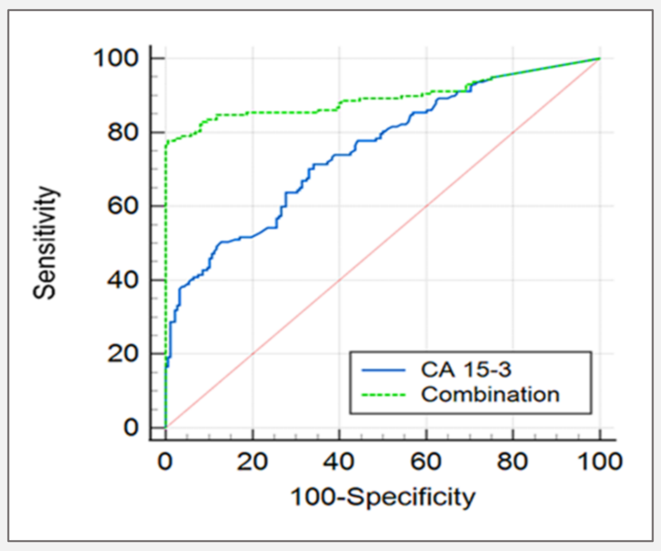

Figure 2: ROC of Stage I with DCIS using CA15-3 alone or in combination with CA-62 from J. Tcherkassova et al.

OBJECTIVES

To determine if a Logistic Regression using the CA 15-3 and CA-62 original values from the abovementioned article plus the age of each individual improved the method's diagnostic efficacy.

METHOD

The serum samples were blinded and had measurements of CA 15-3 (ELISA) and CA-62 (CLIA) as well as their corresponding TNM classification. The study included 300 breast cancer patients (254 at Stages I and II, 20 with ductal carcinoma *in situ* (DCIS), and 26 Stage III and IV patients), and 141 healthy controls.

The Logistic Regression analysis was done with the Real-Statistics Excel® Add-in and included as continuous predictors, CA 15-3, CA-62 and Age.

RESULTS

The logistic regression of Healthy vs. Cancer cases included Age, CA-62 and CA.15-3. The ROC curve is shown in Figure 3.

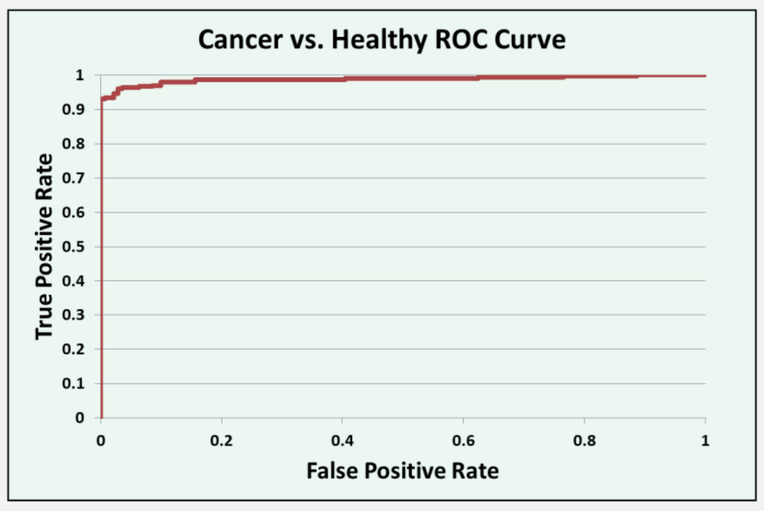

Figure 3: ROC Curve obtained using Binary Logistic Regression on 300 cancer and 141 healthy patients. **Sensitivity = 0.933 with Specificity = 0.993, AUC = 0.988**.

The parameters in Eq 1 (shown again below) are: $\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$

β0 (Intercept) = -8.93, β1 (Age) = 0.0231, β2 (CA-62, U/ml) = 0.00132, β3 (CA15-3, U/ml) = 0.08941

With these values and solving for p we can calculate, for any patient given his/her Age, CA-62 and CA15-3, the probability of having cancer. For example, using the parameters above, for Age = 54, CA-62 = 5,882 U/ml and CA15-3 = 10.8 U/ml, the probability of having cancer = 74%.

DISCUSSION AND CONCLUSIONS

Both statistical methods, Tcherkassova et-al and the one presented here are since there is little or no correlation between the CA15-3 and CA-62 markers. However, Binary Logistic Regression automatically weighs any correlation between variables and provides the corresponding statistical significance.

Another advantage of using Logistic Regression is its ability to combine variables and different types of variables such as binary or categorical predictors in its calculations. Adding significant predictors increases the accuracy of the method as compared to using the customary markers by themselves and even when the physician requests more than one biomarker, the interpretation is subjective rather than mathematical. All these considerations help transform the subjective assessment of the patient by the physician into an objective calculation of the probability of having cancer.

Finally, Logistic Regression provides a probability of having cancer rather than just a value in Units or ng per mL. This is better than interpreting the marker's value since this interpretation requires a good understanding of the shape of the distribution of the marker in both healthy and cancer-bearing patients, which is not common among practitioners. Moreover, a probability can be assessed by the physician as well as the patient. After all, most meteorological reports provide the probability of showers in their daily predictions.

In conclusion, the use of Logistic Regression to assess the risk of a given patient of having cancer appears as advantageous compared to common practice. Furthermore, this statistical method yields better results than providing the marker's value alone since it allows the incorporation of other parameters that are usually related to cancer (e.g., asthenia, anorexia, anemia, etc.) in the estimation of the risk of having a malignancy.

The obtained results show that the application of Binary Logistic Regression on the original dataset considerably improves the prediction of breast cancer diagnosis from 75% with 100% specificity to 95% sensitivity with 99% specificity.

LITERATURE

1. Janneta Tcherkassova, Anna Prostyakova, Sergey Tsurkan, Vladislav Ragoulin, Alexander Boroda and Marina Sekacheva. Diagnostic efficacy of the new prospective biomarker's combination CA 15-3 and CA-62 for early-stage breast cancer detection: results of the blind prospective-retrospective clinical study. Cancer Biomarkers v.35, 2022 p. 57–69
2. Czepiel, S. A. 2002, Maximum likelihood estimation of logistic regression models: theory and implementation. http://czep.net/stat/mlelr.pdf
3. Fagerland M., Hosmer D. and Bofin A., 2008, Multinomial goodness-of-fit tests for logistic regression models, Statistics in Medicine, Wiley Interscience

ACKONLEGMENTS

We thank Dr. J. Tcherkassova for providing the raw dataset values.