**Classification models for predicting a patient clinical status: discussion on the case of the chemotherapy-induced neutropenia duration**

S. San Matías¹, M. Clemente¹ and V. Giner-Bosch¹

¹Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Valencia, Spain.

*Correspondence to: Dr. S. San Matías, Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain. Tel: +34 963877490; Fax: +34 963877499; E-mail: ssanmat@eio.upv.es

**Abstract.**

In this paper, we aim to discuss some particulars about the process of building classification models in oncology applications. In particular, we discuss the results of a recent article by San Matías et al. [1], where the goal was to develop a methodology to predict the duration of a pathology related to the administration of anti-cancer drugs, making use of advanced statistical methods.

1. **Introduction.**

In any medical context, it is very frequent to have situations where anticipating the future outcome of a clinical situation of a patient is essential: dying or not after a heart attack, recovering or not from a disease, etc. In such cases, we certainly know several data about the patient (symptoms, history of previous diseases...), but we would like to predict the unknown future status of the patient. These problems, where the unknown to be predicted for each individual is a dichotomous characteristic, are known as classification problems in the statistical data mining field.

When a classification problem comes up, a tool is required to establishing the relationship between the known patient information (the inputs) and his/her unknown status (the response). Statistics help us to build such predictive tools, the classification models, which are a fundamental part of the knowledge acquisition process. Classification models provide us with accurate decision rules for medical expert systems, improving the efficiency of assessment systems.

Of course, there exist different types of classification tools, which are adequate depending on the characteristics of the problem. In the last years, many different classification problems arising in oncology have been solved using mainly a well-known statistical technique, logistic regression, showing good results and stability of the solutions; as an example, see Klastersky et al. [2]. Nevertheless, this is a constantly
developing field of interest, and of course other methods are also applied more and more on a variety of real oncological cases, with satisfactory results.

In this paper, we revise some recent contributions relating classification models in medicine, and particularly we make a review of our own work in [1]. There, we worked with a real dataset about febrile neutropenic patients in a hospital, and proposed a methodology with the aim of overcoming some drawbacks appearing in many real cases.

2. Improving classification models in medicine: resampling methods.

It is not infrequent to find applications where statistical tools do not reach the desired accuracy in their predictions. Sometimes the problem lies in using an inappropriate model, or even in the lack of good predictors, but not always. When the dataset has a small sample size or when the status to be predicted is highly unbalanced, probably our classification tool will perform poorly. In such cases, how could we improve the performance of our statistic predictive tools? The use of some kind of resampling technique can be a great help.

Resampling methods are becoming very popular because their simplicity and robustness. They involve sampling the original data numerous times, where the samples can be generated either with or without replacement. A randomized sample is generated by scrambling the existing data (sampling without replacement), while a bootstrap sample is generated by sampling with replacement from the original sample. Thus, in any particular bootstrap sample, we expect some data points of the original sample to be present two or more times, while others could be absent.

Among resampling methods, bootstrap techniques are one of the most well-known. The main idea of bootstrap is to consider the sample data as the whole population data, and to resample from this original sample in order to obtain new replicated datasets (bootstrap samples) with the same sample size. Then, the statistical analysis will be repeated with each of those bootstrapped samples. For more information on bootstrap techniques, see Efron [3], Efron and Tibshirani [4], Diaconis and Efron [5] and Davison and Hinkley [6].

In the remainder of this section, we include two references as an example of how resampling has been recently applied to improve predictive models in clinical applications.

In Austin and Tu [7], the authors proposed a model selection method based upon drawing repeated bootstrap samples from the original dataset together with automated variable selection methods. The authors apply this methodology to a dataset of patients admitted to hospital with a diagnosis of acute myocardial infarction, with the objective of predicting mortality within 30-day of admission. The
The final sample size was 3882 patients, with about 11% of mortality, and the authors show that selecting those variables that were identified as independent predictors in at least 60% of the bootstrap samples resulted in an accurate parsimonious predictive model.

A recent, different and successful way of using resampling to improving classifiers is given by the ensembles. Ensembles are models build from base learners, and it has been shown that they can substantially improve prediction performance of models. In this line, Ishwaran et al. [8] proposed for the first time the so-called random survival forests (RSF), an ensemble tree method for the analysis of right-censored survival data. RSF consists of random survival trees: using independent bootstrap samples, each tree is grown by randomly selecting a subset of variables at each node and then splitting the node using a survival criterion involving survival time and censoring status information. The authors report the results of the experiments using 11 medical datasets, and they illustrate how RSF produce accurate classifiers at least as good as conventional methods and much more automatically, and moreover, RSF are able to uncover highly complex interrelationships between variables.

Both of these papers present numerical experiments where the dataset sizes were big enough to guarantee the robustness and significance of the developed classifiers. However, as we have already commented, the difficulties arise when dealing with small sample sizes.

3. Review on a classification model for the duration of the chemotherapy-induced neutropenia (CIN).

A good illustration of the problems that we face to when trying to build a good classifier with a small dataset can be found in our recent work [1]. In this paper, our objective was to develop predictive models to classify patients with chemotherapy-induced neutropenia (CIN) into two groups, according to the duration of the episode. Febrile neutropenia (FN) is the most common side effect associated to the administration of anticancer drugs. It is very important to determine the CIN duration at the onset of a febrile neutropenic episode, because of the possibility of suffering serious complications, including death.

3.1. The problem of predicting the duration of CIN.

There is a relationship between the aggressiveness of a chemotherapy regimen and the neutropenia duration (see for example Blay et al. [9]). Consequently, several authors have focused their efforts in developing predictive models for the CIN duration according to the aggressiveness of the cytotoxic regimen. Apparently, in such models the dependent response should be the numerical duration of CIN; however, it is more
interesting for clinical purposes to classify patients according to their duration status (ds), which is defined as a dichotomous variable whose feasible values are either low expected duration or high expected duration. In this way, the originally continuous problem becomes a classification problem.

With this idea, Lalami et al. [10] developed a scoring model, which assigns a numeric value to each chemotherapy drug and defines a global score as the addition of each drug’s individual scores. Therefore, a threshold of 8 points is used to classify a patient under this regimen either into a group of low expected duration of FN (level 1, <8) or high expected duration (level 2, ≥8). The authors used real data from 203 patients that were classified following their system; the results showed a good performance, as the medians duration of CIN in these two groups of patients were significantly different.

In San Matías et al. [1], we proposed an alternative methodology for solving the same problem as Lalami et al., trying to outperform their results and at the same time, to overcome the inherent instability accompanying models developed with such a small sample size. Our proposal uses one input (the haematological toxicity of each chemotherapy regimen) to predict a dichotomous characteristic, the duration status of a FN patient.

3.2. Data and methodology.

In our study, 106 patients with solid tumours and an episode of febrile neutropenia (FN) were eligible. Haematological recovery to ANC≥2.0×10^9/l (grade 1 CIN) was observed in 88 patients.

Basically, our methodology follows the next three steps:

- Phase 1: defining the response variable
- Phase 2: enlarging the training dataset
- Phase 3: building the classification model
- Phase 4: validating the results

3.3. Defining the response variable

Let us call d the variable accounting for the time (in days) to overcome grade 1 CIN from the beginning of the episode. This variable has to be coded somehow into the new dichotomized variable ds, which is going to be our response variable for the classification model.

Our proposal was to carry out the dichotomization by a k-means cluster analysis. Cluster is a multivariate statistical technique whose aim is to classify a sample into a small number of groups or clusters, in such a way that the individual profiles in the
same cluster are very similar and those of the different clusters are as different as possible. For technical details on clustering techniques, see Mardia et al. [11] and Hastie et al. [12].

The process of the definition of the response variable $d$s can be summarized as follows:

**Initialization.** Select the noncensored cases in the sample. Set $k=2$ (number of clusters to be formed), input variable=$d$.

**Step 1.** Perform a k-means cluster analysis.

**Step 2.** Let *cluster 1* be the group formed with the patients with lowest durations, *cluster 2* the group formed with the patients with highest durations.

**Step 3.** Let $M$ be the maximum value of $d$ for patients in cluster 1.

**Step 4.** Define $ds = 1$ for patients with $d \leq M$, $ds = 2$ otherwise.

As a result, this procedure will generate two groups of patients as separated as possible with respect to their neutropenia durations. Analysis of variance test can be performed to confirm the statistical difference between the two clusters.

In our sample, constant $M$ was equal to 3, and then patients were divided into cluster 1 (72.6%, maximum duration equal to 3 days) and cluster 2 (27.4%, minimum duration equal to 4 days).

### 3.4. Enlarging the training dataset.

As our main inconvenient in order to get a robust predictive model was the very limited sample size, our proposal in our paper was to artificially enlarge the training dataset. More specifically, we proposed the use of an extended database obtained by resampling from the original sample (based on some ideas in Gross [13] and Babu and Singh [14]).

Our suggested steps are as follows:

**Initialization.** Let $s$ be the sample size. Let $T$ be the enlarging coefficient, $T>1$.

**Step 1.** Resample the original dataset (with replacement) in order to get an enlarged sample of size $T\cdot s$.

**Step 2.** Extract a new analysis sample with size $s$ from the enlarged sample, using without replacement sampling.

**Step 3.** Randomly divide the new analysis sample into a training dataset (70%) and a test dataset for validation (30%).

We used a value of $T=3.36$, in order to get a test dataset having the original sample size (106 cases).
3.5. **Building the classification model.**

A Classification and Regression Tree (CART) was carried out on the training dataset for classifying patients according to the dichotomous response variable \( ds \). The predictor variables were \( G_i \), being:

\[
G_i = \text{number of different drugs with score } i \text{ supplied to a patient into the chemotherapy, for } i=0 \text{ to } 4.
\]

CART produces a decision tree which is structured as a sequence of simple (if/then) questions based on the predictor variables and it identifies subgroups of patients with a higher likelihood of testing positive for the high-duration status. See Breiman et al. [15] for more information on CART.

When the tree is grown, each of the final nodes represents a decision rule to classifying patients into the majority class of this node.

In our case, the decision algorithm identified six different rules to assign a patient into the high duration group.

3.6. **Validating the results.**

First of all, the accuracy of the classifier was estimated by using the area under the curve (AUC) from the ROC curve, the sensitivity, the specificity and certain other measures.

Anyway, the ability of the classifier for separating patients with different FN durations was tested using a Kaplan–Meier analysis and log-rank tests. At this point, we used again bootstrap to validate the results of the survival analysis. We obtained 500 replicates of our sample, comparing in each one the survival curves. The more distinguishing the classifier, the higher would be the number of replicates in which the survival curves are significantly different.

In our test dataset, the AUC was 0.672, with a sensitivity of 57.6% and a specificity of 69.9%. The classifier distinguished two groups of patients with statistically different neutropenia duration, with median durations until haematological recovery of ANC \( \text{ANC} \geq 2.0 \times 10^9/\text{l} \) of 4 versus 2 days (\( P < 0.001 \)). Figure 1 shows the survival curves corresponding to predicted low-level/high-level groups defined by the classification method, for the test dataset. Considering a significance level \( \alpha = 0.1 \), the results showed that the survival curves were significantly different (P value less than \( \alpha \)) in 99.4% of bootstrap replicates. All these results enhanced the performance of proposals in previous works.
Figure 1. Time to hematological recovery for an ANC≥2×10⁹/l, for predicted low and high duration groups by the classification method in San Matías et al.

4. Conclusions

In this review, we highlight some statistical issues than can be useful to enhance predictive models in oncology applications. At the moment, there is an increasing interest about resampling and bootstrap strategies, as a way to improve the accuracy and robustness of classical predictive methods. The main difficulties appear when developing such a predictive methodology with a training dataset with a small sample size or when we have a rare event to predict. However, the use of the resampling and bootstrap techniques helps to overcome these difficulties and makes statistically significant the validation of the results. In short, models predictive ability can be outperformed using an appropriate combination of statistical techniques, and particularly making an adequate use of resampling techniques.

As a case study, we have summarized the most important aspects of our proposal in San Matías et al. [1]. In this paper, we dealt with a very small sample size and we proposed a methodology that can be easily followed to get new classification models adapted to different samples.

5. References


