Predicting the duration of chemotherapy-induced neutropenia: new scores and validation

S. San Matías¹*, M. Clemente¹, V. Giner-Bosch¹ & V. Giner²
¹Department of Applied Statistics, Operations Research and Quality, Universidad Politecnica de Valencia, Valencia; ²Oncology Unit, Hospital de Sagunto, Sagunto, Spain

Received 30 April 2009; revised 11 March 2010 & 5 May 2010; accepted 6 May 2010

Background: The objective of this study was to develop predictive models to classify febrile neutropenic patients into two groups, according to a prediction of the duration of the chemotherapy-induced neutropenia episode.

Patients and methods: For this retrospective analysis, 106 patients with solid tumours and an episode of febrile neutropenia (FN) were eligible. A score was attributed to each chemotherapy treatment drug according to its expected toxicity. Three new scores were proposed based only on this classification. Two of them are a combination of the individual drug scores and the third one was built using statistical techniques such as cluster analysis and classification trees.

Results: Statistical techniques produced the best score, distinguishing two groups of patients with statistically different neutropenia durations, with median durations until haematological recovery of absolute neutrophil count $2 \times 10^9/l$ of 4 versus 2 days ($P < 0.001$).

Conclusions: Our methodological approach based on statistical techniques identifies the patients who will need the longest times to recover from FN. The input of this predictive system is only the aggressiveness of the cytotoxic agents in a chemotherapy regimen. Our proposal succeeded in distinguishing two groups of patients and the results show better performance than other scores in previous studies.

Key words: chemotherapy, classification trees, cluster, duration of neutropenia, febrile neutropenia, statistical techniques

Background

Chemotherapy-induced neutropenia (CIN) is the most common side-effect associated with the administration of anticancer drugs. Up to 25% of patients treated with chemotherapy are likely to develop a febrile neutropenia (FN) episode (see for example Crawford et al. [1]), although this percentage could increase up to 96% in some particular types of tumours (Crawford et al. [2]). Previous studies relating to the appearance of FN after the administration of chemotherapy can be found in Blay et al. [3], Silber et al. [4], Ray-Coquard et al. [5], Kondo et al. [6] and Wilson-Royalty et al. [7]. It is very important to determine CIN duration at the onset of a febrile neutropenic episode because a patient with FN will be very susceptible to suffering life-threatening complications, including death, and the higher the duration of FN, the higher the infection risk [8].

It is now accepted that there is a relationship between the aggressiveness of the chemotherapy regimen and FN duration (see reference [3]). Lalami et al. [9] have developed a model that aims to predict CIN duration according to the aggressiveness of the cytotoxic regimen. Among the different proposals of Lalami et al. [9], the best performance was achieved by the so-called S2 score. They assign a score to each chemotherapy drug and define S2 as the addition of each drug’s individual scores. A threshold of 8 points is then 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, $\geq 8$). With their sample, they show that the median durations of CIN in these two groups of patients are significantly different.

However, some data collected at the Oncology Unit in Sagunto Hospital have recently shown contradictory results when applying the proposal by Lalami et al. In fact, Lalami’s proposals did not succeed in distinguishing two groups of different CIN durations with a new sample. In our opinion, the correlation between the chemotherapy regimen and FN duration is clearly proven, but it is not clearly shown that the nature of this relationship must be additive as is suggested in [9]. We think that the use of more specific statistical tools could produce more sensitive models. Moreover, it is possible that different samples show differences in FN duration owing to factors not taken into consideration. If so, it could be useful to
develop a reproducible methodology providing more adapted and accurate models for different sample data, more than a static predictive model.

For this reason, in this work, our goal has been to design a new predictive methodology for FN duration by using advanced statistical tools. In order to get comparable results with those of Lalami et al. [9], in our proposal, the only input is the haematological toxicity of each chemotherapy regimen. We propose different systems to classify patients into two groups of low and high expected duration, and we present and discuss an analytical comparison of their performance. The key of this study is to show how to chain various methods to classify patients accurately according to FN duration and how this conceptual approach can be implemented in practice.

patients and methods

patients

Our research is based on a retrospective observational study. The sample was obtained from patients in the Oncology Unit at Sagunto Hospital (Valencia, Spain). This can be considered as a conventional unit in a general hospital according to its dynamic and patients. Data were collected from June 1997 to October 2006. All the patients selected were outpatients, i.e. they had not been admitted to the hospital at the onset of FN; the eligibility criteria being the following: (i) diagnosis of malignancy treated with chemotherapy. All the patients had solid tumours; (ii) neutropenia caused by this treatment. Absolute neutrophil count (ANC) \( < 0.5 \times 10^{9}/L \); (iii) temperature \( > 38^\circ C \) (FN); (iv) treatment with an appropriate initial empiric antibiotic regimen; (v) age \( \geq 16 \) years; (vi) first episode for each patient.

In order to guarantee accuracy in the computation of duration, we used the analysis data obtained at the Hospital. During the study period, there were no clinical differences in the management of CIN that could distort the results. Moreover, we have studied the time series of all durations and we have observed that there is no statistical trend in the data on the duration of CIN episodes over this period of time.

data

The usual descriptive variables on each patient status were collected. All available blood counts were taken into account to compute the neutropenia duration. For the sake of comparison, we have assumed some of the premises in [9]. Namely:

1. The beginning of FN was set as being the first day with documentation of grade 4 CIN.
2. We will use only the aggressiveness of the chemotherapy regimen as a predictor of CIN duration.
3. Each individual chemotherapy drug is given a score (ranging from 0 to 4) according to its expected haematological toxicity based on data from the literature. This score was assigned as follows: bleomycin, leucovorin, cetuximab and estramusine (score 0); 5-fluorouracil, cisplatin, fludarabine, procarbazine and rituximab (score 1); gemcitabine, melphalan, methotrexate, mitoantronex, mitomycin C, raltitrexed, vinblastine, vinorelbine and vincristine (score 2); carboplatin, cyclophosphamide, epirubicin, ifosfamide, oxaliplatin and Adriamycin (score 3) and irinotecan, docetaxel, paclitaxel, topotecan and etoposide (score 4).

We have focused our study on the time to recovery for an ANC \( \geq 2 \times 10^{9}/L \) (grade 1 neutropenia). We call \( dI \) the variable accounting for the time (in days) to overcome grade 1 CIN from the beginning of the episode. This variable \( dI \) is recoded into a dichotomised variable that we call \( ddI \), which labels each case as either ‘low duration or high duration’. In a subsequent subsection, we explain how to carry out this dichotomisation.

predictive models

Our objective consists of predicting if a patient will belong either to the low- or to the high-duration group using as an input only the information on the cytotoxic regimen. In order to deal with this question, we have followed two conceptually different approaches. Figure 1 summarises these two approaches, including the analytical tools required for them and their outcomes.

We call the classification procedures in the line of Lalami et al. ‘scoring methods’, and we have proposed two new scores belonging to this class. Following the spirit of the scores in Lalami et al. [9], we have defined two new families of scores combining the individual scores of each drug. However, we think that the sum may not be the best way of combining the individual scores if the different drugs supplied to a patient could somehow interact. For this reason, our scores differ in the way we carry out aggregation. The two families that we have built and analysed are ‘magnifying trend scores’ (consisting of increasing weighted sums of the individual drug scores) and ‘reducing trend scores’ (consisting of decreasing weighted sums of the individual drug scores). We have selected the best scores within each trend and we denote them as M1 and R1, respectively. The obtaining of scores M1 and R1 is summarised in Algorithm 1.

Algorithm 1—Computation of scores M1 and R1

For each patient, do the following:

Step 1. Order the scores in an ascending way (0 to 4).

Step 2. Assign ascending weights to each score, starting by 1, 1.1, 1.2, etc.

Add the scores multiplied by the corresponding weights and assign to M1.

Step 3. Assign descending weights to each score, starting by 1, 0.9, 0.8, etc.

Add the scores multiplied by the corresponding weights and assign to R1.

Our second approach can be defined as a ‘statistical classification method’, and in this way, we have developed a new methodology whose application provides the new score T1 as an outcome. Let us define for each patient the following set of variables: \( Gi = \) number of different drugs with score \( i \) supplied to a patient in the chemotherapy regimen, for \( i = 0 \) to 4. Using \( Gi \) as predictor variables, our methodology first builds the score T1 and then predicts variable \( ddI \) for each patient based on the value of T1. The classification procedure is summarised in Algorithm 2.

Algorithm 2—Computation of score T1

Step 1. Definition of Low/High duration.

Dichotomize duration into a new variable \( ddI \) so that \( ddI=1 \) (low duration), \( ddI=2 \) (high duration).

Step 2. Computation of T1.

Step 2.1. Use \( Gi \) to divide patients into small groups according to \( ddI \).

Step 2.2. Compute the percentage of high duration patients within each group.

Step 2.3. For each patient, assign T1 as the percentage of high duration of his/her group.

All the new scores were redefined into two categories. In all cases, a cut-off point \( t \) was determined to carry out this classification.

statistical methodology

The dichotomisation of variable \( dI \) was carried out by a ‘statistical 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. It is an unsupervised learning method because there is not a dependent variable to predict,
only input variables (in this case, d1). There are several procedures for carrying out a cluster analysis. We used K-means cluster, with K = 2 clusters, the maximum number of iterations was equal to 10, the convergence criterion was $10^{-6}$ and the similarity measure was Euclidean distance. For more details on clustering techniques, see Mardia et al. [10] and Hastie et al. [11].

For the development of score T1, we worked with an extended database obtained by resampling from the original sample (see Gross [12] and Babu and Singh [13]). According to this method, we first replicated the observed sample to a bigger size and then used without replacement sampling to select a new analysis sample from that enlarged sample. The analysis sample was then randomly divided into a training dataset (70%) and a test dataset for validation (30%).

A Classification and Regression Tree (CART) was carried out on the training dataset. The purpose of CART analysis is to predict or classify cases according to a response variable. CART uses a binary recursive partitioning method that produces a decision tree. The tree is structured as a sequence of simple (split) questions based on several predictor variables and identifies subgroups of patients with a higher likelihood of testing positive for the high-duration state. In our case, d1 was the response categorical variable, while Gi were the predictor variables used by the decision tree. See Breiman et al. [14] for more information on CART.

The value of the classification cut-off point t was determined for each score by exploring the corresponding receiver operating characteristic (ROC) curve and selecting t as the value for which sensitivity is as high as possible, while sensitivity and specificity are similar.

The validation of the predictive models was carried out using the test dataset. First, we compared the accuracy of the categorised scores T1, M1, R1 and S2 as predictors of d1 using the area under the curve (AUC) from the ROC curves. Moreover, sensitivity, specificity and certain other measures were also evaluated for all the models. Nevertheless, the real validation consists of analysing the impact of all the scores as FN duration classifiers. For this purpose, we used a Kaplan–Meier analysis and log-rank tests to compare the duration estimates between predicted groups. However, on account of our small sample size, we used the resampling technique called ‘bootstrap’ to validate the results of the survival analysis. Bootstrap is a well-known statistical method used to assess the variability of test statistics. Its main idea 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 with the same sample size. Thus, the statistical analysis will be repeated with each of those datasets. For more information on bootstrap techniques or its clinical applications, see Efron [15, 16], Diaconis and Efron [17], Davison and Hinkley [18] and Austin [19].

We obtained 500 replicates of our sample, comparing in each one the survival curves (that is, the graph depicting the number of patients remaining neutropenic as time goes on) for both predicted levels (level 1 = low duration and level 2 = high duration). The more distinguishing the score, the higher would be the number of replicates in which the survival curves are significantly different.

results

patient characteristics

We now present a descriptive analysis of our sample. One hundred and forty-eight patients were eligible for the survey but 42 were excluded because either their commencement of or recovery from neutropenia was not documented. In the following, we will consider only the 106 nonexcluded cases, 21 of whom had a subsequent episode of FN that was not considered for the study. There were 49 women (46.2%) and 57 men (53.8%), and their ages ranged from 16 to 84 years, the average age being around 62 years old and the median age 65 years.

In our sample, we found 19 different types of solid tumours.
(31 cases, 29.2%), lung (28 cases, 26.4%) and head and neck (17 cases, 16%). Other tumours represent 28.4% of the total number of cases. With respect to the number of different drugs supplied to the patients in chemotherapy, 12 patients (11.3%) were treated with monotherapy, 54 (50.9%) with two drugs, 32 (30.2%) with three drugs and 6 (5.7%) with a chemotherapy scheme including four drugs. Only two patients received more than four different drugs. The median number of different drugs was 2.

We observed haematological recovery to ANC ≥ 2.0 × 10^9/l in 88 patients (83%) with a median duration of FN equal to 2.5 days, while 25% remained neutropenic after 4 days. Cluster analysis was carried out using the 88 censored cases. The first cluster was formed by 69 patients whose maximum duration was equal to 3 days. The other 19 patients were grouped into the second cluster, with a minimum duration equal to 4 days. The variable dd1 was then assigned a value of 1 for patients with FN durations of ≤ 3, and a value of 2 otherwise. The analysis of variance test confirms the statistical difference between the two clusters (P < 0.001). In the whole sample, the proportion of patients in the low cluster was 72.6% (77 patients) versus 27.4% (29) in the high cluster.

On the other hand, the replication of S2 in our 106 patients ranged from 1 to 11 with a median of 6.

development of M1 and R1

Once scores M1 and R1 were obtained, the cut-off point was fixed at a value of 6 in both cases. The predicted classification was then as follows: if M1 < 6, classify the patient into the low-duration group; if M1 ≥ 6, classify the patient into the high-duration group (the same for R1). In our 106 patients, the R1 score ranged from 0.9 to 8.9 with a median of 5.6, and M1 ranged from 1 to 14.80 with a median of 6.50.

development of T1

For analysis purposes, we obtained an enlarged sample with n = 356 by applying the resampling procedure described in the ‘Statistical methodology’ section. This sample was then divided into the training dataset (n = 250) and the test dataset (n = 106). The CART procedure was carried out on the training dataset and the resulting decision tree is shown in Figure 2. This tree provided an AUC of 0.814.

A classification cut-off point of 30% was selected for the CART to label a group as high duration. Cases with T < 30 were classified as a low predicted duration (level 1) and cases with T ≥ 30 as a high predicted duration (level 2). This decision algorithm identified the following six groups for the high duration state: (i) G1 = 2 and G4 = 0; (ii) G1 = 0, G2 ≥ 1 and G3 ≥ 2; (iii) G1 = 1, G2 = 0 and G3 ≥ 2; (iv) G1 = 0, G2 = 0, G3 ≥ 2 and G4 ≥ 1; (v) G1 = 0, G2 ≥ 1 and G3 = 1 and (vi) G1 = 0, G3 = 0 and G4 = 0.

Figure 2. Decision tree to classify into low-duration/high-duration groups.
The CART revealed that the most influential variable in the tree classification was $G_1$, while the least relevant were $G_0$ and $G_2$. The first step on the tree shows that 56.5% of the patients with $\geq 2$ score one drugs belong to the high duration level, while this percentage is only 24.7% for patients whose regimen had $\leq 1$ score one drugs. Score $T_1$ ranged from 0 to 100 with a median of 23.5. We should point out that for patients where the underlying tumour was breast, the percentage with a value of $T_1 \geq 30$ was 25.8%, while in patients with other tumours, this value was 44% ($P = 0.08$).

**performance of the predictive scores**

With respect to $T_1$, the training model was found to have an overall sensitivity of 73.9% (51 of 69 high-duration cases) and a specificity of 72.4% (131 of 181 low-duration cases). The misclassification rate was 27.2%, with a positive predictive value equal to 30.5% and a negative predictive value of 87.9%. The likelihood ratio was 2.68 for a positive test and 0.36 for a negative test. The performance results using the test dataset can be found in Table 1, together with those of the scores $M_1$, $R_1$ and $S_2$ considering them as predictors of the variable $dd_1$.

Table 1 summarises also the results concerning the impact of our scores $M_1$, $R_1$, $T_1$ and Lalami’s $S_2$ on the duration classification, for our test dataset. For each predictive score, this table includes the Kaplan–Meier estimates of the median of duration, for our test dataset. For each predictive score, this table includes the Kaplan–Meier estimates of the median of duration, for our test dataset. With our data, it seems that score $R_1$ distinguishes slightly better between low-duration groups and high-duration levels than score $S_2$, as we can see from the results in Table 1. This confirms the results of Lalami et al. [9], namely the influence that the aggressiveness of chemotherapy has over the duration of FN. On the other hand, it shows that this influence is not purely additive but can be improved using different weights over the terms of the addition. Additional work on the improvement of the score weights could be conducted on this issue.

**conclusions**

In our sample (from Sagunto Hospital), CIN duration is usually low. For this reason, it is more difficult to distinguish between two groups. Probably, our durations are so low because all our patients were outpatients at the onset of FN and all of them had solid tumours. Therefore, it is possible that they were admitted to the hospital when the neutropenia had already developed. From the results that we have obtained, we can draw the following conclusions:

1. Our analysis showed that $M_1$ was not able to distinguish between low-duration/high-duration groups, with poorer performance than the rest of the scores, including $R_1$. The underlying hypothesis for this behaviour could be that some noxious effects may be common when administering drugs with similar components and that they could produce lower toxicity than was expected.

2. With our data, it seems that score $R_1$ distinguishes slightly better between low and high duration levels than score $S_2$, as can be seen from the results in Table 1. This confirms the conclusion of Lalami et al. [9], namely the influence that the aggressiveness of chemotherapy has over the duration of FN. On the other hand, it shows that this influence is not purely additive but can be improved using different weights over the terms of the addition. Additional work on the improvement of the score weights could be conducted on this issue.

3. Our proposal $T_1$ improves on scores $M_1$, $R_1$ and $S_2$, as can be seen from the results in Table 1 and Figures 3 and 4 (this available online only). This shows that the use of more sophisticated statistical techniques such as classification trees leads to obtaining more robust predictive models. CART analysis is a powerful technique that allows the construction of easily interpretable decision rules that can be applied in clinical practice. Some advantages of tree methods are they do not require that the nature of the relationship between the predictor variables and the outcome be parametrically specified; they are nonparametric and nonlinear; the final results can be summarised in a series of logical ‘if–then’ conditions (tree nodes); they are very useful when there are many possible predictor variables, which make the task of variable selection difficult and they are able to model complex interactions or patterns in the data.

4. Score $T_1$ is more intuitive than the rest of the scores as it ranges between 0 and 100.

5. The relation between the duration of neutropenia and the aggressiveness of chemotherapy is nonlinear and $T_1$ has the ability to model this intrinsic nonlinearity.

The results of this study highlight that the use of advanced statistical techniques for predicting neutropenia duration can lead to better results than existing models even when using the same information as input. In contrast to scoring systems based on subjectively determined values, our methodology...
provides an objective, statistically derived predictive system. Although we have just a small sample, the use of the resampling and bootstrap techniques makes the validation of our results statistically significant and it allows the results to be extended to other potential cases from the same area.

We are aware that our study has the limitation of the sample size. Patients with FN who were not included or who had missing data may have been different from the patients described in this study. However, there is no reason to suspect that they would have been more or less likely to experience a longer FN duration than the included patients.

In conclusion, we propose more than a predictive system, which can be applied directly to other data, as we provide a methodology that can be easily followed to get new predictive models adapted to different samples. This study aims to highlight the importance of future research in this area.

acknowledgements

The authors wish to acknowledge the Oncology Unit in Sagunto Hospital for the facilities that they provided us with for carrying out this study, and the reviewers for their valuable comments.

disclosure

None of the authors declare conflicts of interest.

references