New Decision Support Tool for Treatment Intensity Choice in Childhood Acute Lymphoblastic Leukemia

Carlos Eduardo Pedreira, Senior Member, IEEE, Leonardo Macrini, Marcelo G. Land, and Elaine S. Costa

Abstract—Acute lymphoblastic leukemia (ALL), the most common cancer in childhood, has its treatment modulated by the risk of relapse. An appropriate estimation of this risk is the most important factor for the definition of treatment strategy. In this paper, we build up a new decision support tool to improve treatment intensity choice in childhood ALL. Our procedure was applied to a significant cohort of Brazilian children with ALL, the majority of the cases treated in the last decade in the two main University Hospitals of Rio de Janeiro. Some intrinsically difficulties of this dataset introduce an assortment of challenges, among those the need of a proper selection of features, clinical and laboratorial data. We apply a mutual information-based methodology for this purpose and a Neural Network to estimate the risk. Among the relapsed patients, 98.2% would have been identified as high-risk by the proposed methodology. The proposed procedure showed significantly better results when compared to the BFM95, a widely used classification protocol.


I. INTRODUCTION

A CUTE lymphoblastic leukemia (ALL) is the most common cancer in childhood. In children under-15 age group, it corresponds to approximately 80% of the cases of leukemia and to 30% of all cases of cancer. ALL may also occur in adults, but its clinical behavior, prognosis, and treatment are different. For this reason, in this paper, we limit the cohort to children below 12 years old. ALL arises from a single progenitor cell that has undergone genetic damage, leading to uncontrolled proliferation and arrested differentiation. As consequence, the bone marrow is occupied by a malignant cell population and normal hematopoietic production decreases [1].

Nowadays, the probability of event free survival for ALL (an estimate of the rate of cure) is approximately 75% to 80% in 5 years. So, despite the progress of the treatment strategy in the last forty years, 20%–25% of the patients will relapse in a period of five years from diagnosis. An appropriate estimation of the risk of relapse has been considered the most important factor for the definition of treatment strategy since its intensity is modulated by this risk [2], [3]. The consequence of an incorrect decision concerning the intensity of the treatment is twofold: it may increase the risk of relapse; or produce serious therapy-related sequels, including death. Although support care has very much improved in recent years, late therapy-related sequels, that include cardiac toxicity, endocrine toxicity, and secondary cancer, among others, are still a very serious concern, particularly in children [4]–[7]. For these reasons, the classification of the risk of relapse directs the stratification of the intensity of the treatment and is the decisive step of its planning [2].

Investigative efforts, initiated in the 1970s and consolidated in the 1980s, have produced significant contributions concerning patients’ classification in accordance to a set of prognostic factors [8]. This classification gave rise to chemotherapy schemes modulated by patient’s clinical and laboratorial variables measured at the beginning of the treatment. The purpose is to avoid excessive toxicity for the cases with good prognosis and insufficient treatment for the bad ones. The main goal of this paper is to build up a new decision support tool to improve treatment intensity choice in childhood ALL. This decision is founded on an estimation of the relapse risk based on available clinical and laboratorial data at (or just before) the moment when the treatment intensity choice has to be made [1]–[3].

In order to achieve this objective, four central difficulties had to be overcome. The first difficulty is related to the size of the dataset. Since ALL in children is a rare disease (its incidence is approximately 1 case per 100 000 inhabitants per year), the number of cases available for analysis is always restricted [9]. The intrinsically small size of the dataset (if compared to most engineering ones) imposes some computational-engineering tasks, among those an adequate choice of the input variables plays a crucial role in this context. In the present contribution, data are extremely precious and difficult to obtain. Each data point is related to a patient who was observed in a long-term clinical treatment. The second intrinsic arduousness is that although treatment intensity does influence the risk of relapse, it cannot be used as direct input information for modeling since it is what one seeks as outcome. The challenge is to use just the patients’ available information, together with other accessible knowledge, to infer the best treatment to be applied. The third difficulty is that all nonrelapsed patients are right censored data, since one can never guaranty that they will not relapse in the follow-up period, the complete survival time interval has been cut off, i.e., censored at the right side.

1An observation is said to be right-censored (a broadly used term in Survival Analysis) when its exact survival time becomes incomplete at the right side of the follow-up period, the complete survival time interval has been cut off, i.e., censored at the right side.
future. Note that, the larger the follow-up time, the lesser is the chance of future relapse. In fact, due to the course of the disease and treatment, the relationship between the follow-up time and the chances of relapse is not linear, after 36 months the chances of relapse strongly diminish, and it almost vanishes after 60 months, given it had not so far occurred [2], [8], [10]. Actually, very late relapses have been considered in the literature as a new manifestation of the disease [10].

Finally, the fourth hardship concerns the risk of relapse being a hidden output variable in the sense it can never be directly measured.

II. DATASET AND METHODOLOGY

A. Dataset

The dataset consists of 189 possible input features, for each of the 158 children considered in this investigation. Note that these 158 cases constitute approximately the expected number of cases in a population of 5.2 million in three years. The patients were diagnosed in the two main university hospitals of Rio de Janeiro [Federal University of Rio de Janeiro (UFRJ) and University of Rio de Janeiro State (UERJ)] from April 1993 to July 2003. Concerning this investigation, the patients were followed up and the dataset was updated up to April 2004. Patients, whose diagnoses were performed before 1997, were treated with a protocol called modified-BFM95 and after that with the modified-BFM95. These protocols have three therapeutic branches (low, medium, and high-risk groups) to which the patients are stratified following a risk-related rule [2]. These protocols suffered some minor modifications. Effectively, low and intermediary risk branches were identical in both protocols.

The possible input features are clinical and biological data obtained at diagnosis. These input variables may basically be organized in three types: 1) clinical data at the diagnosis: for example, age at diagnosis, sex, race, date of diagnosis, birth date, presence or absence of loss of weight, fever, anemia, arthritis, hemorrhage, increased lymph nodes, increased liver and spleen, infectious focus, abdominal pain, bone pain, mediastinal mass, abdominal mass; 2) laboratorial data at the diagnosis: for example, red blood cell count, hemoglobin concentration, hematocrit, white blood cell (WBC) count, blast count in peripheral blood, platelets count, serum lactate dehydrogenase, percentage of blasts in bone marrow, blast immunophenotype, and cytogenetics abnormalities in blasts; 3) data related to treatment response: for example, total blast cell count in blood at the eight day of treatment, complete remission in the 33rd day of treatment.

B. On the BFM Protocol

The BFM95 is a widely used ALL relapse-risk classification system with an associated treatment protocol. It has been applied in many hospitals around the world, including the two university hospitals where patients considered in this paper were treated. The three levels of risk considered in this protocol are associated to three levels of treatment intensity (see Table I) [2]. Despite its unquestionable importance, approximately 70% of the patients are classified by the BFM95 in the intermediary-risk group, and approximately 35% of these patients will relapse, which suggests that this group contains a heterogeneous cluster of patients [2]. The lack of further discriminating power of this risk classification clearly demonstrates the need to new relapse-risk evaluation methods, which are able to better discriminate the patients in the intermediary group.

C. Methodology: Feature Selection and Neural Network Estimation

The selection of input features plays a central role in the successful solution of classification problems for two main reasons: First, because it is obviously important to use the inputs carrying the maximum amount of information to the output. In noisy environments, redundant or uninformative inputs are not only unhelpful but may also in fact overshadow performance. Second, because the number of parameters is able to properly estimate in a model is directly related to the dataset size. A larger number of input features imply the estimation of a larger number of model parameters. Both these points are particularly relevant in limited-in-size datasets.

Finding the appropriate set of explanatory variables may decisively improve the model performance [11], especially when the input variables feed a nonlinear model. Concerning input variables, there are two major goals: to make the problem manageable regarding the relationship between the number of input variables and the available number of patients and to select the variables that produce effective risk estimation. Since some of the input variables candidates may be redundant or irrelevant, we started by preprocessing the input variables, eliminating obvious redundancies and evaluating the variables discriminating (power and value of positive prediction).

First, we eliminated the variables that contain too many missing values or obvious redundancies, typically rates or factors formed from other variables. For instance, our dataset has a categorical variable named G-age (related to “age” as follows: G_age = 1 for age < 1 or age > 10; G_age = 2 for 1 < age < 10; G_age = 3 for age > 12) that was eliminated in favor of the originally collected variable “age”. The variables “blast cell count in blood at diagnosis” and “white blood cell count at diagnosis” have a strong linear correlation coefficient of 0.99. We opted to use WBC and not “Blast cell count in blood at diagnosis” because WBC may be automatically quantified with

<table>
<thead>
<tr>
<th>Feature</th>
<th>Age year</th>
<th>White cell blood count at diagnosis</th>
<th>T-lymphoma phenotype</th>
<th>Blast cell count in blood at 30th day of treatment</th>
<th>Cytogenetics</th>
<th>Complete remission in 30th day of treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>--------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Standard-risk</td>
<td>Age &gt; 6</td>
<td>&lt;3000</td>
<td>Absent</td>
<td>&lt;1000</td>
<td>Absent</td>
<td>Yes</td>
</tr>
<tr>
<td>Intermediate-risk</td>
<td>Age &lt; 1 or age &gt; 6</td>
<td>&gt;20000</td>
<td>Present</td>
<td>&lt;1000</td>
<td>Absent</td>
<td>Yes</td>
</tr>
<tr>
<td>High-risk</td>
<td></td>
<td></td>
<td></td>
<td>&gt;1000</td>
<td>Present</td>
<td>No</td>
</tr>
</tbody>
</table>

2Refers to “Berlin-Frankfurt-Munich protocol”
more reliability. Nowadays, WBC may be automatically counted by a machine while blast cells, a leucocytes subtype that may abnormally appear in peripheral blood, need to be manually recognized by an expert morphologist with an optical microscopy. The number of attributes was reduced from 189 to 51 (73% reduction) as consequence of this step.

Next, we focused on variables discriminating power and value of positive prediction.

1) **Discriminating power**: The goal is to evaluate the dependence between the input features and the outcome. We build up a two-way contingency table and use a $\chi^2$ statistic to compare the observed and expected counts in each cell of this table [12].

2) **Value of positive prediction**: We evaluate the probability of a patient relapse conditioned to the presence of an input feature candidate. A cut-off at 50% was used.

We verified the coherence, from the clinical point of view, of all eliminated features.

These two steps have produced a reduction to just 17 variables (10.6% of the original 189). The eligible features are at this point: age, bleeding, bone pain, FAB\(^3\) classification, fever, gender, hemoglobin, lymphonodemegaly, liver, malaise, presence of infectious focus, pain, race, spleen, T-immunophenotype, total blast cell count in blood at the eight day of treatment, and WBC. We eliminated 14 patients that have missing in these relevant variables, therefore reducing the number of patients to 144.

Next, we applied a mutual information (MI)-based methodology to order and select the more relevant input features among the preselected 17 variables. Previous paper involving MI in the context of machine learning includes, among others, [13] and [14].

The MI [15], a measure of (in)dependence between random variables, is defined as

\[
I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

Here, $p(x, y)$ is the joint probability density distribution of random variables (RVs) \(X\) and \(Y\). Since two RVs are statistically independent if and only if $p(x, y) = p(x)p(y)$, it follows that independence of \(X\) and \(Y\) implies $I(X, Y) = 0$. Furthermore, note that $I(X, Y) \geq 0$. A large (or small) MI value means that the random variables are very (or little) related. Note that the MI measures possible nonlinear relationships between RVs, in contrast to correlation coefficient that is able to detect linear dependencies only.

The MI between a feature and the outcome is a measure of how much (in a nonlinear way) the outcome is influenced by this particular feature. Following the same reasoning, with the goal of accessing the level of redundancy, one may measure the MI between two or more features. The aim is to search for a set of features that at the same time have a large MI with respect to the outcome and a small MI with the other features in this set. Nevertheless, one would have to calculate the MI in sets embracing all combinations of the feature candidates, what in most real world applications would produce a large (in some cases unfeasible) number of possible associations. A way to overcome this adversity accrues from the MIFS-U [16] algorithm, an improved variant of the MIFS algorithm [17]. It follows a brief schematic description of this algorithm. This procedure was performed considering the 17 preselected features for all patients, except the 28 ones with less than 36 months of follow-up, which were reserved for out-of-sample evaluation only.

Let \(F\) be a set formed by the 17 preselected feature candidates.

**Step 1**: Initialize a set \(S\) as empty (i.e., \(S = \emptyset\)) and compute the MI between each of the features candidates (i.e., each \(f_i \in F; i = 1, \ldots, 17\)) and the output label, call this MI $I(f_i; OUT)$.

Let \(k = 1;\)

**Step 2**: Select, among all \(f_i \in F\), the feature (call it \(f_k\)) that produces the maximum value for $I(f_i; OUT)$. Next, remove \(f_k\) from \(F\) and put it in \(S\), i.e., set $F \leftarrow F - \{f_k\}$ and $S \leftarrow f_k$.

**Step 3**: Choose, among all features in \(F\), the one that, jointly with the one(s) already in \(S\), maximizes the MI with the output, i.e., choose a feature \(f_k\) such that $I(f_k; S; OUT)$ is maximized for \(f_k \in F\). If \(k < 17\), make \(k = k + 1\) and go back to step 2 until, otherwise, stop.

Calculation of $I(f_k; S; OUT)$ in step 3 is quite hard since it involves a conditional probability maximization (find the feature that maximizes the MI with the output, given that a set of features have been already chosen). A simplification that provides a good approximation for $I(f_k; S; OUT)$ was proposed in [16], for technical details we remit the readers to the original paper.

Neural Networks (NN) [18]–[20] have been successfully applied in many fields since the eighties. Its well-known capacity to produce good results for nonlinear problems without the necessity of stressing a physical model \textit{a priori} is among its strong qualities. Different sets of MIFS-U selected features were used, for each patient, as the inputs of a feedforward NN. All the input variables were normalized in the interval \([-1, 1]\). We used the softmax (multiple logistic) activation function [18], [21] as output units such that results can be interpretable as posterior probabilities for the categorical targets.

Because of the limitation in the number of available observations (intrinsic in this type of problem), we opted to use the “leave-one-out” approach [18], [19], a well-known version of cross-validation. In our cohort, the mean of the follow-up time is 79 months (median = 70 months) for nonrelapsed patients and 28 months (median = 16 months) for the relapsed ones. We did not use the 28 nonrelapsed patients with less than 36 months of follow-up in the training set, these being used just for out-of-sample inference. The remaining patients were used for training and also generalization-evaluated in the leave-one-out scheme.

Unfortunately, it is not possible to establish a cohort of patients receiving a uniform chemotherapy protocol. Since the 1970s, it has been well recognized that some clinical and laboratorial characteristics indicate a worse prognosis. So, it is of course antitheics to carry out any treatment protocol that does not consider these well-known prognostic factors to set

\(^3\)“French-American-British” classification for acute lymphoblastic leukemia.
chemotherapy intensity. These variations on treatment intensity may directly impact on a patient outcome, since a high-risk patient may not relapse exactly because he/she has received an appropriate aggressive treatment. In this way, the influence of treatment is an unavoidable bias in all ALL clinical trials. Of course, it is not possible to determine which nonrelapsed patients are in fact high-risk and have survived due to the treatment intensity. Consequently, we have a hidden output target. The positive impact of intensive treatment on patients classified as high risk by the BFM95 has been extensively demonstrated in clinical trials [2], [22]–[24]. Our concern in this paper is directed to the other groups, specially the highly heterogeneous intermediate-risk one. Based on this, the NN target was set as “zero” for nonrelapsed patients that were not classified as BFM high risk, and as “one” for relapsed patients and/or patients classified as BFM high risk.

One central topic in the NN literature is concerned with the model complexity, i.e., how many hidden units, or neurons, should be used. Often, this is done by some “rule of thumb”, by estimating a number of models with different number of hidden units and choosing the one with the best performance. Several alternatives to this, some how arbitrary, way of electing the architecture has appeared in the literature. We employed the Bayesian regularization [25], [26] strategy that turned out to be quite successful in a number of applications. The fundamental idea is to find a balance between the number of parameters and goodness of fit by penalizing large models. The objective function is modified in such a way that the estimation algorithm effectively prunes the network units by driving irrelevant parameter estimates to zero. In this way, the number of hidden units is not arbitrarily chosen by some trial-and-error approach and consequently, the complexity of the NN model is automatically fitted to the complexity of the data. We started the Bayesian regularization algorithm with ten neurons (92 parameters) and ended up with an effective parameters average (after the 144 leave-one-out runs) equal to 63.82, with a standard deviation of 3.44.

D. Performance Evaluation

We chose the following three functions to evaluate the performance of the proposed methodology:

1) Rate of high-risk (RHR) estimated in the relapse group

\[ RHR = 100 \frac{R_{HR}}{R} \]

where \( R_{HR} \) is the number of patients that relapsed among those estimated as high risk, \( R \) is the total number of relapsed patients. The higher this rate, the better is the performance. The HRR reflects the rate of correctness in recognizing the high-risk patients; it measures the model capability of identifying the patients who need more intensive treatment.

2) Rate of overestimated risk (ROR in comparison with the BFM95 protocol).

Since there is no gold standard for this problem, this rate provides a comparison with a widely used protocol. It indicates the proportion of patients whose risk could have been overestimated by the model. Restraining to the group of nonrelapsed patients, the ROR, defined later, is the rate of the ones evaluated as high risk by the model and not as high risk by the BFM95. In those cases, if the BFM95 protocol risk stratification is correct, our model would be indicating a more aggressive than necessary treatment.

\[ ROR = 100 \frac{R_{OE}}{NR} \]

\( R_{OE} \) is the number of nonrelapsed patients who were estimated as high risk by the model and as standard or intermediate risk by the BFM95 protocol. NR is the total number of nonrelapsed patients. The lower this rate, the better is the performance. The ROR may be related to “false alarm rate” concerning the risk estimation.

3) Rate of subestimation risk (RSR in comparison with the BFM95 protocol)

Restraining to the group of relapsed patients, the RSR, defined below, is the rate of the ones evaluated as low risk by the model and not as high risk by the BFM95.

\[ RSR = 100 \frac{R_{SE}}{R} \]

where \( R_{SE} \) is the number of relapsed patients, who were estimated as low risk by the proposed model and as high risk by the BFM criteria. \( R \) is the total number of relapsed patients. The lower this rate, the better is the performance. The RSR may be related to the “probability of miss” concerning the risk estimation.

III. RESULTS AND DISCUSSION

In Table II, one can find the ordering result (after the MIFS-U application) regarding the 17 preselected features. In Table III, we present an overview of the experiments concerning different sets of input variables. In experiment 1, we used all the 17 preselected features, in experiments 2, 3, and 4 just the 7, 6, and 5 more relevant ones, respectively. We have made other simulations with different numbers of variables (respecting Section II.A preselection) and all results can be considered worse than the ones obtained in the referred experiments.

It is interesting to note that although the feature “race” does not appear in most international risk classification systems [2], its relevance was already detected in and it has shown prominence in our results. Its inclusion resulted in an RHR increase from 87% to 98% (compare experiments 4 and 5). Since we used Brazilian data, we conjecture that this variable could be a surrogate of socioeconomic features (income, access to health system, parents’ educational level) [27]. However, biological factors [28], [29] could not be ruled out as possible explanations for prognostic impact of race.

The best result was obtained for experiment 5, which is equal to experiment 4 added feature “race.” This performance can be considered extremely good, 98% RHR with 0% RSR and just 21% of ROR. It is quite relevant that in three of the experiments, we obtained a zero rate for RSR, meaning that the model will not underestimate the risk of any relapsed patient among the
TABLE II

INPUT CANDIDATES AND THEIR MI WITH RESPECT TO THE OUTPUT

<table>
<thead>
<tr>
<th>Ordering</th>
<th>Variable</th>
<th>MI(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>White cell blood count at diagnosis (WBC)</td>
<td>0.197</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>0.086</td>
</tr>
<tr>
<td>3</td>
<td>T-Immunophenotype</td>
<td>0.032</td>
</tr>
<tr>
<td>4</td>
<td>Blast count cell in blood at 8th day of treatment</td>
<td>0.036</td>
</tr>
<tr>
<td>5</td>
<td>Gender</td>
<td>0.024</td>
</tr>
<tr>
<td>6</td>
<td>FAB classification</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>Race</td>
<td>0.012</td>
</tr>
<tr>
<td>8</td>
<td>Malaise</td>
<td>0.004</td>
</tr>
<tr>
<td>9</td>
<td>Fever</td>
<td>0.004</td>
</tr>
<tr>
<td>10</td>
<td>Bone pain</td>
<td>0.002</td>
</tr>
<tr>
<td>11</td>
<td>Bleeding</td>
<td>0.002</td>
</tr>
<tr>
<td>12</td>
<td>Infectious focus</td>
<td>0.001</td>
</tr>
<tr>
<td>13</td>
<td>Pan</td>
<td>0.001</td>
</tr>
<tr>
<td>14</td>
<td>Hemoglobin</td>
<td>0.069</td>
</tr>
<tr>
<td>15</td>
<td>Lymphomendomegaly</td>
<td>0.001</td>
</tr>
<tr>
<td>16</td>
<td>Spleen</td>
<td>0.033</td>
</tr>
<tr>
<td>17</td>
<td>Liver</td>
<td>0.063</td>
</tr>
</tbody>
</table>

TABLE III

EXPERIMENTS WITH DIFFERENT SETS OF FEATURES

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
<th>Experiment 4</th>
<th>Experiment 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE feature</td>
<td>WBC feature</td>
<td>WBC feature</td>
<td>WBC feature</td>
<td>WBC feature</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>count in 2nd</td>
<td>count in 2nd</td>
<td>count in 2nd</td>
<td>count in 2nd</td>
</tr>
<tr>
<td></td>
<td>day of treatment</td>
<td>day of treatment</td>
<td>day of treatment</td>
<td>day of treatment</td>
</tr>
<tr>
<td>1</td>
<td>91</td>
<td>87</td>
<td>81</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>31</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE IV

RELAPSE RISK ESTIMATED BY THE MODEL

<table>
<thead>
<tr>
<th>Risk estimated by the model</th>
<th>Non-relapsed patients</th>
<th>Relapsed patients</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>60 (67.3%)</td>
<td>1 (1.8%)</td>
<td>61 (62%)</td>
</tr>
<tr>
<td>High</td>
<td>30 (33.3%)</td>
<td>53 (58.2%)</td>
<td>83 (58%)</td>
</tr>
<tr>
<td>Total</td>
<td>90 (100%)</td>
<td>54 (100%)</td>
<td>144 (100%)</td>
</tr>
</tbody>
</table>

TABLE V

RELAPSE RISK ESTIMATED BY THE BFM95 VERSUS THE PROPOSED MODEL

<table>
<thead>
<tr>
<th>BFM Risk</th>
<th>Model Risk</th>
<th>Non-relapsed patients</th>
<th>Relapsed patients</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>20 (22.2%)</td>
<td>0 (0%)</td>
<td>20 (13.9%)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>4 (4.4%)</td>
<td>5 (9.3%)</td>
<td>9 (6.3%)</td>
</tr>
<tr>
<td>Intermed.</td>
<td>Low</td>
<td>31 (34.4%)</td>
<td>1 (1.9%)</td>
<td>32 (22.2%)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>15 (16.7%)</td>
<td>23 (42.6%)</td>
<td>38 (26.4%)</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>9 (10.0%)</td>
<td>0 (0%)</td>
<td>9 (6.3%)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>11 (12.2%)</td>
<td>25 (46.3%)</td>
<td>36 (25.0%)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>90 (100%)</td>
<td>54 (100%)</td>
<td>144 (100%)</td>
</tr>
</tbody>
</table>

indicated as high risk by the BFM protocol. Hence, no relapsed patient that would have been treated with intensive drugs, if following the BFM protocol, would be undertreated by applying the proposed model.

The selected features are very coherent from the medical point of view and usually considered to be relevant in the medical literature (apart from race for the reasons discussed above). The BFM95 and BFM2002 protocols [2]–[24] use age, WBC count at diagnosis, and blasts blood count at eight day of the treatment. The consensus of the National Cancer Institute in the USA recommends the use of WBC count at diagnosis and age. It is important to note that although the model has produced a set of features that are in accordance with the “expected from the medical point of view,” we did not use any medical expectancy as a priori information.

We follow by comparing the proposed model results with those referring to the BFM95 protocol. Please see Tables IV and V. For the relapsed patients group, 98.2% of the relapsed patients would be identified as high risk by the proposed methodology, while only 46.3% were set as high risk by the BFM95 criteria. Furthermore, all five patients, the BFM criteria sets as low risk and actually relapsed, are set as high risk by our model. Among the 24 patients that relapse and were classified as intermediate risk by the BFM, 23 would be set as high risk by our method. Focusing on the patients that were classified as low or intermediate risk by the BFM95 criteria and relapsed, 96.6% of those would be set as high risk by the proposed model. This results point in the direction of a considerable decrease in the relapse level if a protocol based on the proposed method is used. Note that 28 of the 29 patients that were set as standard or intermediate risk by the BFM and relapsed would be classified as high risk and consequently received a more intensive treatment.

Concerning the results for nonrelapse, of course, the treatment had a direct impact on the outcome, and so it is not possible, in fact, to have a “gold standard” for low risk of relapse. Patients that did not relapse may have been saved by factors directly linked to treatment. In our model, 66% of the nonrelapsed patients are classified as low risk.

In summary, we have the following:
1) Among the 29 patients evaluated by BFM95 criteria as standard-risk (low), nine were estimated as high risk by the proposed model. Five of these patients effectively relapse up to the Ending Date of this investigation.
2) Among the 70 patients evaluated by BFM95 criteria as intermediate risk, 38 were estimated as high risk by the proposed model. Twenty-one of these patients effectively relapse up to the Ending Date of this investigation.
3) Among the 45 patients classified by BFM95 protocol as high risk, nine were estimated as low risk by the proposed model, none of those have relapsed.

The proposed model would prescribe a more aggressive than necessary treatment (four low risks and 15 intermediate risks by the BFM95, out of 90 nonrelapsed patients) for up to 21% of the nonrelapsed patients. Concerning this numbers, it is important to keep in mind that relapses may still occur in the future for some patients. In this way, this 21% rate is in fact an upper bound for the error.

We end up this section by presenting, for comparison purposes, results generated by linear and nonlinear discriminant...
analysis [30], [31]. The same conditions (applied for the Neural Network model) were imposed: we take only non-relapsed patients with more than 36 months after diagnosis and the same previous selected six variables, WBC; age; T-immunophenotype; blasts blood count at eight day of the treatment; gender and race. We also used the same target-output inversion procedure. We applied generalized discriminant analysis [31] with three types of nonlinear kernels, namely Gaussian, polynomial, and sigmoid. The comparative performances, for experiment 5, can be found in Table VI.

### IV. Final Remarks

The relapse risk estimation in ALL in childhood is one of the most important pediatric oncology problems. ALL is the most frequent cancer in childhood and the prognostic information can decisively modulate the therapy intensity. In this paper, we proposed a new decision support tool that improves the relapse-risk estimation in comparison with a well-established protocol classification system. This method was implemented and tested in a dataset composed by a significant majority of childhood ALL cases in Rio de Janeiro in the last ten years.

The large number of possible input features in the original dataset imposed a careful selection of the appropriated set to feed the NN model. From the medical point of view, the resulting selected features, and their ordering, can be well accepted as explanatory of the relapse risk and in fact, most of them are already used in several relapse-risk estimation systems of international protocols.

Although there is a considerable agreement between the sets of variables selected by the proposed strategy and by the BFM95 (e.g., age, WBC at diagnostic, immunophenotype, blast count cell in blood at eight day of treatment), one of the fundamental differences between these approaches concerns the way these variables are considered in building up the decision criterion. While the BFM95 is based on cutoffs—for age, blast count cell in blood at eight day of treatment and WBC (cf., Table I)—the proposed scheme generates a smooth decision system by combining the selected variables in a nonlinear fashion. Another advantage of presented framework is that as new possible risk factors become available, it is straightforward to rerun the routine including these factors and so establishing updated protocols.

The proposed procedure showed significantly better results when compared to the widely used BFM95 classification protocol, specially, concerning the group classified as intermediate risk. An expressive parcel of this population relapsed and would be classified as high risk by the proposed methodology. So, from the clinical point of view, the proposed methodology allows a better identification of the group of patients that should be treated with more intensive chemotherapy.

### ACKNOWLEDGMENT

The authors are greatly indebted to Prof. J. C. Príncipe for his suggestions, corrections and critiques. His contributions have decisively improved the paper.

### REFERENCES

Dr. Pedreira was the Founding President of the Brazilian Neural Networks Society.

Leonardo Macrini was born in Rio de Janeiro, Brazil, in 1954. He received the M.S. and the Ph.D. degrees in control theory and statistics from the Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, in 2000 and 2004, respectively.

He is currently a Research Assistant at the Institute of Public Health (IESC), Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. His current research interests include pattern recognition, neural networks and multivariate statistics.

Dr. Macrini is currently a Fellow at Brazilian National Research Council (CNPq).

Marcelo G. Land was born in Rio de Janeiro, Brazil, in 1961. He received the B.S. degree in medicine and the M.S. and the Ph.D. degrees from the Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, in 1985, 1993, and 1997, respectively.

He became a specialist in hematology in 1989. He is currently an Associate Professor in the School of Medicine (UFRJ) and director of the UFRJ Pediatrics Institute [Clinical Medicine Graduate Program & Pediatric Institute, (IPPMG)]. His current research interests include diagnostic factors and quality of life in pediatric oncology.

Elaine S. Costa was born in Rio de Janeiro, Brazil, in 1974. She received the B.S. degree in medicine and the M.S. and the Ph.D. degrees from the Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, in 1996, 2003, and 2006, respectively.

She became a specialist in pediatric hematology in 2000. From 2006 to 2008, she was a Research Fellow at the Cancer Investigation Center, University of Salamanca, Spain. She is currently a Practitioner and a Researcher of the Pediatric Oncology Service at the UFRJ Pediatrics Institute (Clinical Medicine Graduate Program & Pediatric Institute, (IPPMG)). Her current research interests include neoplastic diseases in childhood and multiparametric flow cytometry.

Carlos Eduardo Pedreira (SM’03) was born in Rio de Janeiro, Brazil, on April 11, 1956. He received the B.S. and the M.S. degrees in electrical engineering (Systems) from the Pontifical Catholic University (PUC-Rio), Rio de Janeiro, Brazil, in 1979 and 1981, respectively, and the Ph.D. degree from the Imperial College of Science Technology and Medicine, University of London, U.K., in 1987.

He is currently an Associate Professor in the School of Medicine and COPPE-PEE-Engineering Graduate Program, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. His current research interests include pattern classification, cluster analysis, neural networks, and statistical methods for biomedical applications.

Dr. Pedreira was the Founding President of the Brazilian Neural Networks Society.


