Evolutionary computation in the identification of risk factors. Case of TRALI

María Dolores Torres a,*, Aurora Torres a, Felipe Cuellar b, María de la Luz Torres b, Eunice Ponce de León a, Francisco Pinales a

a Universidad Autónoma de Aguascalientes, Av. Universidad # 940, Ciudad Universitaria, C.P. 20131 Aguascalientes, Ags, Mexico
b Centro Hospital Miguel Hidalgo, Galeana Sur # 465, Col. Obraje, C.P. 20230 Aguascalientes, Ags, Mexico

ARTICLE INFO

Keywords:
Genetic algorithm
Evolutionary Hybrid Algorithm
TRALI's risk factors
Testor
Typical testor
Informational weight

ABSTRACT

This paper presents the use of an evolutionary algorithm hybridized with the concepts of testor and typical testor in determining factors associated with transfusion related acute lung injury (TRALI). Although nowadays many cases of this syndrome remain ignored or misdiagnosed, this is the leading cause of morbidity and mortality related to transfusion in the United States.

This research was conducted with data from 174 cases collected in the Centenary Hospital Miguel Hidalgo in the city of Aguascalientes, Mexico, in the period 2007 to 2010.

The proposed algorithm works with information from the model known as “two hits”, in which the first hit is the original disease and the second corresponds to the blood transfusion. This algorithm was strengthened with mechanisms that let it do an efficient search in the whole solution space. In addition to the calculation of the informational weight, the algorithm also establishes the cutoff point that determines the variables that impact the most.

From the results given by the algorithm and the cutoff proposed by the medical staff, a strategy for the treatment of patients that should be transfused was proposed.

This study confirmed some of the risk factors previously reported in the literature, and also made an interesting discovery.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

The influence that a group of variables has on a response variable is a common topic in several areas. When the response variable is numeric, it is feasible to apply tools such as multiple regressions; however, when the response variable is dichotomous or polytomous, the problem must be addressed by other techniques, such as logistic regression. Regardless of the technique used to determine the effect of a group of variables over another, this process is known as risk factors determination.

Identifying risk factors for any health disorder is an area in which researchers have invested many hours. The identification of risk factors is important because it allows physicians to avoid or quickly identify the presence of any adverse medical condition. By having the opportunity to know in advance that a serious condition can occur in a patient, health experts have precious time to establish treatment or actions that could reduce its impact.

This paper describes the application of a hybrid genetic algorithm in the determination of informational weight of variables related to “transfusion related acute lung injury” (TRALI). The used TRALI’s etiology in this work is a model known as “two hits”; consisted of two events: the first of them is related to the clinical picture of the patient (first hit), and the second is the transfusion of blood products (second hit) (Rodríguez-Moyado, 2011a). The determination of informational weight implies the identification of TRALI’s risk factors and the establishment of an assessment to each variable.

The use of conventional statistical techniques (binomial or multinomial logistic regression), may produce unsatisfactory results when analyzed independent variables are highly correlated (McGee, Reed & Yano, 1984). Another aspect that should be considered, in order that logistic regression makes sense, is that monotonous relationship should exist between the explanatory variables and the response variable (Domínguez & Aldana, 2001). Also according to Hsieh (Hsieh, 1989), the number of subjects needed to use logistic regression smoothly, must be greater than $10(k + 1)$ where $k$ is the number of explanatory variables. This apparently innocuous restriction makes impossible the analysis of rare syndromes by these techniques. Due to this, it is desirable...
to develop other techniques to quantify the importance of the relationship between each predictor variable and the dependent variable, without having to meet the mentioned requirements.

Artificial intelligence has provided several tools to make efficient use or interpretation of data, to different areas; without the need to have expert knowledge of the phenomenon under study. With the use of metaheuristics, countless complex problems have been solved efficiently. This paper describes the process and tools used for informational weight determination of TRALI related variables using a metaheuristic hybridized with the concepts of testor and typical testor. A testor is a set of features capable of distinguishing between two classes (may be more); and a typical testor is a minimal testor. These concepts are presented in Section 3.

Because the determination of all typical testors immersed in a basic matrix is a problem of exponential complexity according to Sanchez and Lazo (2008), the use of approximate solution techniques such as genetic algorithm is perfectly justified. Specifically, the approached problem was initially described by 31 variables, so the amount of possible subsets of variables is $2,147,483,647$. The single generation of this many subgroups, even without assessing their ability to describe the syndrome, would take a long time and would be very computational resources consuming. As Torres showed (Torres, 2010), this problem is so hard that some metaheuristics without additional information can be executing for several days and find no one typical testor. In this work, Torres used a basic matrix of 32,768 rows and 29 features.

To better understand the application presented in this paper, Section 2 provides a brief description of the transfusion related acute lung injury and justification of the importance of its early identification. In Section 3, the concepts of typical testor, testor and informational weight are described. Section 4 presents main concepts related with hybrid metaheuristics, whereas Section 5 discusses in detail the used methodology for determining the informational weight of TRALI related variables. Later, in Section 6 it is described the implemented algorithm and operators involved on it. Later, in Section 7 the results and discussion are presented. Finally, in Section 8 the conclusions drawn from this research are discussed.

2. Transfusion related acute lung injury (TRALI)

This paper addresses the identification of risk factors for TRALI with preventive and early identification purposes. This syndrome is characterized by the development of acute respiratory failure and pulmonary edema within 6 h after transfusion, with a wide clinical spectrum of presentation. The term “transfusion-related acute lung injury” was coined in 1985 by Popovsky & Moore (1985).

According to Athon et al., although this is an under-diagnosed and under-reported medical condition, this syndrome is considered the leading cause of transfusion-related death in the United States, and the second one in the United Kingdom (Añón, García, Quintana, González & Bruscas, 2010).

Nowadays many cases of TRALI remain ignored or are misdiagnosed as fluid overload or acute lung injury (ALI) of other etiology according Vlaar et al. (2011), so a strong motivation for the study of this condition is that physicians consider its relationship with transfusion. As stipulated in Rodriguez-Moyado (2011a), doctors of transfusion medicine services must establish a monitoring of patients in surgical and intensive care units, to identify which component is related to the respiratory distress syndrome, identify the donors involved and clarify whether the blood component contains leukocyte antibodies.

The low level of diagnosis of this medical condition, combined with its seriousness, makes it a syndrome that should be avoided, as Cuellar stated (Cuellar, 2012). In the presence of prone patients to this clinical condition, the following actions are recommended: employing blood transfusion components leukocyte-reduced, with less than 72 h of storage and irradiated; for red blood cells, washed units are recommended (Rodriguez-Moyado, 2011b).

Since its initial description in Barnard (1951) non-cardiogenic lung edema related to transfusion has been reported widely using different names, including non-cardiogenic pulmonary edema, pulmonary hypersensitivity and severe allergic pulmonary edema.

Today, this syndrome has gone from being an almost unknown side effect of transfusion, to the leading cause of transfusion related morbidity and mortality.

The formal definition of this medical condition was developed by the National Heart Lung and Blood Institute Working Group on TRALI US in Toy et al. (2005), and is described as new acute lung injury (ALI)/acute respiratory distress syndrome (ARDS) occurring during or within six hours after blood product administration. Although the absence of specific disease markers and diagnostic tests has resulted in a large variation in estimations of incidence, TRALI is generally considered to be a rare event.

Some of the most important elements of the TRALI’s definition according to medical experts are (Kleinmann et al., 2004):

1. Sudden onset of acute lung injury (ALI).
2. Hypoxemia (PaO2/FiO2 (ratio of partial pressure of arterial O2 to the fraction of inspired O2) $\leq 300$ and must be adjusted downward with increasing altitude or $SpO2 \leq 90\%$ on room air or other clinical evidence).
3. Bilateral lung infiltrates on frontal chest radiograph.
4. No evidence of left atrial hypertension (i.e., transfusion-associated circulatory overload).
5. Occurrence during or within 6 h after completion of transfusion.
6. No temporal relationship to an alternative risk factor for ALI.
7. New ALI and no other ALI risk factors present including aspiration, multiple trauma, pneumonia, cardiopulmonary bypass, burn injury, toxic inhalation, lung contusion, acute pancreatitis, drug overdose, near drowning, shock and sepsis.
8. If one or more ALI risk factors are present, possible TRALI should be diagnosed (in patients with an alternative ALI risk factor, TRALI is still possible).

There are different types of blood components, however according to Rodriguez-Moyado (2011a), fresh frozen plasma (FFP) figures prominently as a cause of immediate adverse transfusion effects. Moreover, according to Pita and Ramírez (1999) it has been demonstrated abuse of PFC transfusion in Mexico at a rate of 96.23% of 229 cases in a general hospital. According to the above-mentioned, it is necessary to establish awareness programs and access to information that enables health specialists, to predict the presence of these alterations and even avoid them.

There are two proposed etiologies for this syndrome: the first is an immune mediated episode produced by transfused antibodies directed toward Human Leukocyte Antigens (HLA) or Human Neutrophil Antigens (HNA). The second, is a model consisted of two events: the first of them is related to the clinical picture of the patient (first hit), and the second one is the transfusion of blood products (second hit). This model commonly known as “two hits” was taken as the basis for the classification of the studied variables.

For this reason, we have grouped the variables as first hit variables, second hit variables and other variables.

Because each of the two hits occurs at different time, we propose that health specialists assess first hit variables highly related to TRALI (according with obtained results) in a first moment, and if these variables are presented, then they have to try to avoid transfusion. If transfusion is unavoidable, then it has to be...
administrated carefully and assessing the variables of the second hit associated with this syndrome. Although the literature indicates that TRALI is frequently associated with the transfusion of plasma products, it can also occur in recipients of packed red blood cells due to the residual plasma present in the unit. It was also found that the transfusion of blood products obtained from multiparous women is associated with a significantly higher incidence of hypotension, decreased PaO2/FIO2 relationship and proinflammatory cytokines (Palfi, Berg, & Ernerudh, 2001). Therefore, transfusion of blood components obtained from these donors is a higher risk of inducing immune-mediated TRALI (Muller, 2005). Other factors related to the sensitization of the donors are transplants and previous transfusions. Several studies have explored the prevention of TRALI using three main strategies:

- Establishing a policy of exclusion of donors.
- Establishing and following strict storage criteria of blood components
- Avoiding unnecessary transfusions.

The above strategies have been formulated based on what is currently known or assumed about this syndrome, but the establishment of these policies in Mexican hospitals not seem achievable in the medium term, on the other hand, having an assessment of the variables more related to this condition to diagnose it in advance, is already done.

3. Testors and typical testors

The main advantage of using testors and typical testors to identify risk factors associated with a specific condition (as TRALI) is the fact that expert knowledge about the problem is not needed. Discarding and selecting variables that are truly related to the problem is a consequence of the combinatorial process, and it is not a step in which the experts have to intervene.

Typical testor concept began to be used in the middle of the fifties in Russia (Chegus & Yablonskii, 1955). One of its first applications was related with failure detection in electrical circuits, this application was followed by problems as supervised classification, and variable selection related with the geological area (Alba, Santana, Ochoa & Lazo, 2000).

We can say that one of the most important works in the area of subset selection was the developed by Dmitriev et al., this work is very remarkable because they pioneered the use of typical testors in feature subset selection (Dmitriev, Zhuravlev, & Krendeleiev, 1966).

According to the definition of Shulcoper, Guzmán & Martinez (1999), a testor is a set of features capable of distinguishing between two classes (may be more), because no object from class T0 can be confused with any object from class T1. A testor is called irreducible (typical), if the removal of any feature from it, it becomes a non testor. The term “irreducible”, indicates that no column can be eliminated. The term “typical” conveys the idea of “typicality” for a class of objects, that is, a set of features that typify a class of objects and otherwise distinguish them from other classes (Lazo, 2003).

Thus, according to Torres (Torres, 2010), an easy way to describe testor is: a set of features required to distinguish two (or more) objects that belong to different classes. This set not necessarily has to be minimal; so typical testor is the minimum set of features required to distinguish objects belonging to different classes. In other words: if a feature is eliminated from a typical testor, it will be not more a testor (Santiesteban & Pons, 2003).

In 1965, it was opened an application line of testors’s theory for classic pattern recognition (Shulcoper et al., 1999). Since then, testors have been applied to supervised learning problems fundamentally.

Let us suppose U is a collection of objects, and these objects are described by a set of n features, and these objects belong to k classes.

In order to determine the testors and typical testors of this collection it is needed to build a structure called basic matrix (BM). Basic matrix has information related with the basic differences existing among objects from different classes. So, basic matrix is created based on other structure called difference matrix (DM).

A simple example for obtaining DM and BM is presented. Let us suppose we have a set of objects defined in terms of the features (n = 3) shown in Table 1.

And let us suppose we have the next information about a group of objects. This original matrix of data is known as the learning matrix (LM), see Table 2.

Then, difference matrix (DM) is created by recording the comparison of each feature of the objects from the same class versus objects from the others. Difference matrix is the matrix whose content let us discern objects coming from different classes. DM is a binary coded matrix, where value “1” was used if there are differences, and value 0 was used if there are not differences.

For our example, the difference matrix would be as shown in Table 3.

Once the DM has been constructed, the basic matrix (BM) has to be created. BM is constituted by all the basic rows from DM. Let p and q be two rows from DM. We say that p is a subrow of q if \( v | q_j = 0 \rightarrow p_j = 0 \) and \( |v_j = 0 \rightarrow q_j = 1 \). A row p is called basic if no row in DM is a subrow p. The submatrix of DM containing all its basic rows (without repetitions) is called a Basic Matrix (Díaz-Sánchez et al., 2011).

Based on the previous definition, we can build the basic matrix. The BM for the discussed example is presented in Table 4.

A subset of features T of a certain BM is a testor, if and only if a row of zeros does not appear in the submatrix of BM after the elimination of those columns that belong to T. The set T is typical, if by

<table>
<thead>
<tr>
<th>Feature</th>
<th>Domain</th>
<th>Comparison criterion</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>((-1.0, 1.0))</td>
<td>Acceptable error</td>
<td>( e = 0.5 )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>(Male, female)</td>
<td>Strict equality</td>
<td></td>
</tr>
<tr>
<td>( x_3 )</td>
<td>(Green, blue, black)</td>
<td>Strict equality</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Object</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O_1 )</td>
<td>1.0</td>
<td>Male</td>
<td>Green</td>
<td>1</td>
</tr>
<tr>
<td>( O_2 )</td>
<td>-1.0</td>
<td>Male</td>
<td>Blue</td>
<td>1</td>
</tr>
<tr>
<td>( O_3 )</td>
<td>0</td>
<td>Female</td>
<td>Green</td>
<td>2</td>
</tr>
<tr>
<td>( O_4 )</td>
<td>0.5</td>
<td>Male</td>
<td>Black</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O_1 \times O_1 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( O_1 \times O_2 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( O_2 \times O_1 )</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( O_2 \times O_4 )</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
is essential for each typical testor and therefore, or could be \{0,1\}, the set of real numbers, the set of comparison is performed is described below.

In order to apply the concepts of testor and typical testors, it is important to establish how comparisons are going to be done. One of the most common comparison criterion is known as strict equality, because is applicable to any kind of domain. The way this kind of the most common comparison criterion is known as strict equality is described below.

Let us suppose \( \Omega = \{O_1, O_2, \ldots, O_m\} \) is a set of \( m \) objects and \( \{O_1\}, \{O_2\}, \ldots, \{O_m\} \) their descriptions in terms of their features, \( R = \{x_1, x_2, \ldots, x_n\} \), where each feature \( x_i \) has associated a set of acceptable values \( M_i \). For example \( \{O_1\} = \{x_1(O_1), x_2(O_1), \ldots, x_n(O_1)\} \), and \( x_i(O) \in M_i, j = 1, 2, \ldots, n \) and \( i = 1, 2, \ldots, m \) (Saniestein & Pons, 2003).

The strict comparison criterion “C” between the \( t \)th feature of object \( i \) and \( j \), is described in the next function:

\[
C(x_i(O_1), x_i(O_2)) = \begin{cases} 
0 & \text{if } x_i(O_1) = x_i(O_2); \\
1 & \text{otherwise}
\end{cases}
\]

In this work, this criterion was used with the whole set of variables. In order to facilitate the variable managing, each feature was discretized into classes according to the recommendation of health experts. This process is explained in next section.

4. Methodology

The current research was made with data provided by the “Centenary Miguel Hidalgo” hospital located in Aguascalientes City in Mexico. This is one of the most important health care specialist hospitals in the center of Mexico. The information handled in this work corresponds to the total identified TRALI cases from 2007 to 2010. The cases of the four years were collected in 2011. In order to have a balanced database, for each case a control (case-control) was incorporated. Cases were those patients that presented the syndrome and controls were those that do not. Case-control studies are often used to identify factors that may contribute to a medical condition.

As mentioned above, this work is based on the medical model known as “two hits” that consists of two groups of variables that can be related with TRALI. First hit variables are those that describe the basis disease of the patient, while the second hit variables are the ones related to transfusion. And although TRALI related literature identifies only these two groups of variables, our medical partners incorporated a third group of variables that could not be located into the two mentioned groups. These variables were called “other variables”. These other variables were introduced with the hope to make new discoveries.

At the beginning, the historic archive of patients was reviewed looking for TRALI cases. The period of time considered covers from 2007 to 2010, because this syndrome is rare and the collecting of cases is slow.

In the Table 6 it can be seen the way controls and cases were distributed over the four years of data collection. The total number of cases was 174 of which 87 were cases of TRALI and 87 are controls that were also transfused without presenting any adverse medical condition.

The general approach followed in this work is sketched in Fig. 1. As can be seen, the first phase consisted of a pre-processing with the intention of reducing the size of the problem.

At the beginning, researchers had 31 variables grouped as Table 7 shows.

Once pre-processing has been finished, 9 variables were eliminated and the study continued with 22 of them (the elimination process was applied on patient identification variables that did

<table>
<thead>
<tr>
<th>Table 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic matrix.</td>
</tr>
<tr>
<td>( x_1 )</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informational weight.</td>
</tr>
<tr>
<td>( x_1 )</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>50%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution of cases before pre-processing.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Before pre-processing</td>
</tr>
<tr>
<td>Controls year</td>
</tr>
</tbody>
</table>

Fig. 1. General approach: The whole process begins with data collection and through the steps outlined in the figure, ends with the development of a medical treatment strategy.
This data preparation stage also consisted on the discretization of the whole set of data. The remaining variables were grouped as it is shown in Table 8.

Table 9 shows the group of variables that belong to each group in the previous table. These variables were considered in the genetic process after discretization stage.

Discretization performed on each variable is presented in the following tables. “Base disease” is shown in Table 10. This table presents the reasons a patient had to be hospitalized, and Table 11 shows theirs coding.

For the remaining variables, the criterion for discretization was: presence = 1, absence = 2.

Another important aspect of pre-processing was that a case and a control were eliminated because they contained missing values. Thus, cases and controls kept for research were 172. Its distribution is shown in Table 14.

After the pre-processing stage, the genetic process took place. This process is detailed in the hybrid evolutionary algorithm section (Section 5).
The hybrid evolutionary algorithm (TRALI-GAA) computed the whole set of typical testors associated to the TRALI from the learning matrix. Each testor constitutes the set of variables that best distinguish between patients who will present TRALI from those that will not.

The set of all typical testors associated with a training matrix, enables the calculation of the informational weight of each variable, which is an assessment based on the occurrence frequency of a variable in different typical testors (as it was illustrated in Section 3). This procedure also can be reviewed in Torres, Ponce, Ochoa, Torres & Díaz, 2009).

When the measurement of the influence each variable has for presenting the syndrome is obtained (informational weight), the automatic process computed the cutoff point for remarking the main risk factors.

The cutoff point calculated by the software is made by assigning an index that corresponds to a central value of all weights different to 100.

A second cutoff point was established by medical experts based on their experience to propose a strategy of medical management of patients in two stages: One first attention moment is established to evaluate the variables from the evolutionary cutoff point, and another, to evaluate the remaining variables obtained based on the experience of medical experts from the Centenary Hospital Miguel Hidalgo (second cutoff point). If a patient has risk factors of the first stage, then medical staff has to emphasize care and to track variables of the second stage.

5. Hybrid evolutionary algorithm

The designed algorithm for this study is based on the presented in MICAI 2012 (Torres et al., 2013) and it was also inspired on HEGAFFSSL (Torres, Ponce, Ochoa, Torres & Díaz, 2009). However, this algorithm incorporates new properties.

Previous researches showed that the performance of a hybridized genetic algorithm with the same mechanisms that a hybridized univariate marginal distribution algorithm has better performance (Torres, 2010), so we decide to use the first to address the current problem. Using hybrid mechanisms reported in Torres, Ponce, Torres et al. (2009), TRALI-GA was constructed for determining the risk factors of TRALI. The algorithm reported in this paper is named TRALI-GAA because it produces a cutoff point automatically.

Both TRALI_GAA algorithm as its components, are presented below.

5.1. TRALI-GAA

The Evolutionary Hybrid Genetic Algorithm with Automatic cutoff determination for the identification of risk factors for TRALI is shown below:

As it can be seen on the previous pseudo code, the first stage of the algorithm consists of a pre-processing step; which eliminates some variables that does not add knowledge about the syndrome (TRALI). Once variables have been filtered, the remaining are incorporated to the search of risk factors, these variables are grouped into first hit, second hit, and other variables in accordance with the literature and the suggestion of physicians.

The basic matrix is calculated from the differences matrix; basic matrix (discussed in the Section 3), contains the basic differences among objects that belong to different classes. For this study, two classes were considered: patients with the syndrome and patients without it.

Table 15 presents the parameters used on the evolutionary process.

<table>
<thead>
<tr>
<th>Parameters of the evolutionary process.</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crossover probability</td>
<td>80%</td>
</tr>
<tr>
<td>Generations</td>
<td>100</td>
</tr>
<tr>
<td>Executions number</td>
<td>50</td>
</tr>
<tr>
<td>Mutation</td>
<td>6%</td>
</tr>
<tr>
<td>Population Size</td>
<td>200</td>
</tr>
<tr>
<td>Improvement mechanism</td>
<td>Verified on each individual</td>
</tr>
<tr>
<td>Accelerating Operator</td>
<td>100%</td>
</tr>
<tr>
<td>Elitism</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

Algorithm – TRALI-GAA

begin
  Initial pre-processing
  Generating Difference Matrix phase
  Generating Basic Matrix phase
  Generate initial population (randomly)
  Apply accelerating operator
  Compute fitness
  If necessary
    Apply improvement mechanism
    Compute fitness
  End if
  population — new initial population
  Repeat
    Begin/* New Generation */
    Repeat
      Begin */ reproductive cycle for pairs of individuals */
        Apply selection operator
        Apply crossover operator
        Apply accelerating operator
        Compute fitness
        If necessary
          Apply improvement mechanism
          Compute fitness
        End if
        Count = Count+1
      End
    Until Count = (generation size/ 2)
    Order of population by fitness
    Apply elitism
    Population ← new population
  End
  Until (stopping criterion is reached)
  Final set of typical testors analysis
  Compute informational weight for each feature
  Set automatic cutoff point
  Get final features subset selection
End
the individual is a testor but not a typical one, its fitness is 10; and if the individual does not qualify as a testor, its fitness is 5.

Considering the value of the fitness function (based on the individual’s condition), the algorithm resolve if an improvement is required. Obviously, when an individual is a typical testor, none improvement is required, but if the individual is weighted as a simple testor, a reduction in the number of variables selected is performed; on the other hand, if the individual is weighted with the minimum value, an increment in the number of variables selected as a components of the individual is required for its improvement. So, the improvement mechanism works as a guided mutation (increasing or reducing the number of variables of the individual).

A more detailed explanation of this mechanism is presented in Section 5.3.

When the improvement phase is developed (if it was needed); the new value of the fitness function is computed, and the algorithm begins the creation of the next generation using elitism. Then the iterative process begins. This process consists on the generation of new populations that continually improve the previous ones. When stopping criterion is reached, the final set of typical testors is obtained and the informational weight is computed. This last computation let us know the importance of each variable. The importance of informational weight resides in the fact that it allows medical staff become aware of the effect that each variable has for this medical condition.

On one side, improvement mechanism can be considered as a local search strategy, because it starts from an individual and looks inside it (the individual) for another better. The accelerating operator on the other hand, works as a global search mechanism that allows delimiting those areas of the search space that are really promising.

The algorithm was also strengthening using elitism to accelerate and to guarantee its convergence; because according to Rudolph (1996), the inclusion of an elitist selection rule into genetic algorithm, allows it to converge to the global optimum. Besides this, by means of the local and global search mechanisms, we rescue the better of the interior and exterior scale algorithms for finding typical testors.

The use of a logical combinatory focus, combined with a genetic algorithm and the special group of mechanisms, constitutes an interesting hybridization for supervised learning.

As an immediate result of the proposed system we obtained an efficient exploration and exploitation of the search space.

In addition, an innovative element has been introduced to the algorithm. In the pseudo-code it is marked as “Set automatic cutoff point” and it refers to the establishment of the cutting point for the selection of the most important variables. The algorithm shows all the variables and emphasizes those whose informational weight is over the calculated cutoff point. All the variables are presented with their corresponding informational weight (the algorithm automatically makes a determination of the cutoff for the variables with the greatest impact on the problem of TRALI).

The algorithm that determines the cutoff point automatically is described in the next pseudo-code.

In the pseudo-code below, Total_w represents an accumulator variable that stores the total sum of the informational weights lower than 100%. Cutoff1 contains the automatic calculated cutoff point. The second cutoff point was set by expert physicians.

In the medical field as in other areas, it is very helpful to have elements for determining a cutoff point that will eliminate or include variables, when literature does not discuss them. Therefore, to have a strategy for determining the cutoff point automatically is an important contribution in the risk factors determination problem and in general, in the intelligent systems area; because expert knowledge is not always available and sometimes it is very difficult for experts to fixed it.

To propose a solution to the risk factors determination problem in the health area, will allow physicians to identify and adequately treat patients who have a predisposition to a syndrome as TRALI. Based on the results of the TRALI-GAA algorithm; medical experts made an assessment and obviously generated a medical strategy that will be implemented for the optimal management of patients in Hidalgo Hospital centenary of the city of Aguascalientes.

5.2. Accelerating operator

In the example presented in Section 3, it is shown that the first line of the basic matrix has a telltale case: 001. If we see the basic matrix rows as binary strings, then this chain constituted by only one 1 (and the remaining bits are 0) guarantees that the variable corresponding to the bit whose value is 1 must necessarily be present in any typical testor associated with the learning matrix. Note that both typical testors found, contain the variable $x_3$ (which corresponds to that bit 1).

Working with binary individuals allows us to see clearly that the identification of a single bit reduces the search space in half. If there were two rows with the described properties in the basic matrix, then the search space is divided by four and so on.

When Basic Matrix is built, an analysis is performed on it. Accelerating operator identifies special cases like the one described at the beginning of this section.

Assuming we have the following basic matrix of 3 variables Table 16; (see example in Section 3).

There is a row with only one 1 and zeros (the first row). Consequently, every typical testor has to have value 1 in variable $x_3$.

Table 16

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 17

Representation of the accelerating operator.

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$*$</td>
<td>$*$</td>
<td>1</td>
</tr>
</tbody>
</table>

$*$ Can take any value.

Table 18

Current individual.

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Thus, a mask that transforms $x_3 = 1$ for any individual whose variable $x_3 = 0$ will be applied. This mask works as an operator that looks for a better solution. The application of this operator is depicted on Table 17.

The operator shown allows any value for $x_1$ and $x_2$, but it ensures that $x_3 = 1$. For more information about this operator, the reader may consult (Torres, 2010).

5.3. Improvement mechanism

As mentioned above in Section 5.1; the improvement mechanisms will search for the typical tester inside a simple tester (reducing its number of features) or will search for a tester on an individual valued with 5 (increasing its number of features). The way this mechanism works is depicted as follows.

Let suppose we have the individual in Table 18 (before the improvement mechanism).

The individual is valued as a tester (fitness = 10). We know that the tester contains inside at least one typical tester. So, considering the mask presented in Table 17, $x_1$ and $x_2$ can take any value (0 or 1); but $x_3$ has to be 1. The improvement mechanism selects a bit from those whose value is 1 and that can be 0 (random selection between $x_1$ and $x_2$ for our example) and mutates it. This task is reducing the number of variables in the tester, looking for the typical tester we know is inside the current tester.

The mentioned process will give us any of the individuals included on Table 19.

If reader compares individuals on Table 19 with the typical testers in the example on Section 3, he will see that both individuals are typical testers (after the improvement mechanism). This mechanism does not guarantee obtaining typical testers but the searching process is well guided.

In the other hand, if the individual before the improvement mechanism is not even a tester (its fitness = 5), the improvement mechanism will add variables to the current individual.

Let suppose we have the individual presented in Table 20 (before the improvement mechanism).

The improvement mechanism will select any bit valued with 0 (randomly) and it will mutate it (generating a value 1 where the individual had a value 0). The result of this process will be any of the individuals shown in Table 19. So, a typical tester could be founded.

A complete description of the whole set of mechanisms used in this work, can be consulted in Torres et al. (2012).

6. Results and discussion

The two hits model reported in the literature was an excellent framework for the identification of groups to which variables obtained by TRALI-GAA belong. However, we do report a third group of variables named “other variables”. This last group of variables contains those TRALI risk factors that could not be listed into the first or second hit, and it is a result of this research.

Initially, we had 31 variables that were subjected to pre-processing. In this first stage 9 variables were removed, so that 22 of the 31 original variables were retained for the genetic process.

The algorithm was executed with 2 cutoff points:

1. The first was calculated by the algorithm automatically to highlight the most relevant risk factors. This cutoff was established in 54.52 and calculated as the midpoint among informational weights different to 100.00 (because an informational weight of 100.00 ensures the presence of its variable).
2. Expert physicians established the second cutoff point for accepting a variable as an important risk factor. The cutoff point they fixed was over 50.00.

Variables of the first cutoff point were established as those that must be evaluated in a first moment. If these variables indicate risk of TRALI, the second cutoff point variables needs to be checked, since its presence increases the possibility of TRALI.

The experimental results shown that even though genetic algorithms are known as rapid convergence algorithms, the set of mechanisms incorporated to TRALI-GAA, let it conveniently work.

The whole set of variables found for the algorithm according to each cutoff point is presented in Table 21.

6.1. First hit variables

Variables considered within the first hit, is the group of variables that can be detected in the original disease of a patient. Original disease is the health condition that causes the need for blood transfusion; it could be a surgery, sepsis, trauma or any other medical affection.

The resulting variables associated with the first hit, using the first cutoff point (54.52) were reduced to 5. These variables and their informational weight are presented in Table 22.

Moreover, based on the second cutoff point (over 50.00), the resulting variables for the first hit were 8 (three more than in Table 22). Table 23 shows the new three variables and its informational weight.

6.2. Second hit variables

According with the two hits model, variables belonging to the second hit are those that can be associated with the blood transfusion. Normally a blood transfusion takes place in intensive care units and operating rooms; and many of times, the surgery or

### Table 19

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 20

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 21

Distribution of the determined variables according with both cutoff points.

<table>
<thead>
<tr>
<th>Variables after TRALI-GAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cutoff point</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>First</td>
</tr>
<tr>
<td>Second</td>
</tr>
</tbody>
</table>

### Table 22

Variables associated with the first hit from the first cutoff point.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Informational weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>100.00</td>
</tr>
<tr>
<td>Base disease</td>
<td>80.06</td>
</tr>
<tr>
<td>PaO2/FiO2</td>
<td>100.00</td>
</tr>
<tr>
<td>History of heart attack</td>
<td>60.54</td>
</tr>
<tr>
<td>Hematological malignancy</td>
<td>55.08</td>
</tr>
</tbody>
</table>

### Table 23

Distribution of the determined variables according with both cutoff points.

<table>
<thead>
<tr>
<th>Variables after TRALI-GAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cutoff point</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>First</td>
</tr>
<tr>
<td>Second</td>
</tr>
</tbody>
</table>
admission to intensive therapy can be associated with a polytrauma because a crash car, another kind of accident or any disease. As shown in Table 24, these variables were reduced to 2 using the first cutoff point (54.52).

Table 24
<table>
<thead>
<tr>
<th>Variable</th>
<th>Informational weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than 3 units of fresh frozen plasma</td>
<td>57.10</td>
</tr>
<tr>
<td>Fresh frozen plasma dated with more than 2 weeks</td>
<td>57.07</td>
</tr>
<tr>
<td>More than 3 globular packages</td>
<td>50.50</td>
</tr>
<tr>
<td>Multiparous donor</td>
<td>51.47</td>
</tr>
</tbody>
</table>

Considering the second cutoff point (over 50.0), 2 more variables were conserved as TRALI risk factors. Table 25 shows the mentioned variables and their informational weight.

6.3. Other variables

Besides first and second hits variables, researchers wanted to know if some other suggestive variables could contribute to suffer TRALI; so, the genetic process searched among a group of 8 variables that were labeled as “other variables” (these variables cannot be considered within the group of variables of the first hit or the second hit). This part of the research was introduced with the intention to contribute to the knowledge of risk factors for TRALI.

According to the cutoff point of the algorithm (first cutoff point = 54.52), only one variable was strongly related with TRALI (see Table 26).

Making use of the second cutoff point (second cutoff point = 50.00), it can be seen that 2 more variables can be conserved. These 2 new variables and their informational weight are shown in Table 27.

Table 25
<table>
<thead>
<tr>
<th>Variable</th>
<th>Informational weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous surgery</td>
<td>53.31</td>
</tr>
<tr>
<td>Previous transfusion</td>
<td>52.10</td>
</tr>
</tbody>
</table>

Based on the obtained results, we can conclude that TRALI is presented mostly in patients with predisposition and presence of risk factors related with previous cardiac or lung disease, smoking habits, and previous hematological malignancies. We also can say that risk increases if the blood units used to obtain the fresh frozen plasma have more than 2 weeks in storage. “TRALI” presented in our hospitable environment is more related with the use of fresh frozen plasma than the use of globular packages. This relationship can be associated with the time of storage of the blood component (unfortunately a very frequent situation).

Blood components donated by multiparous women had less relationship with TRALI than the one expected by physicians (probably because of the deficient record of donors in the blood bank). Risk factors related to the medical history of a patient: cardiac history, previous lung disease and hematological malignancy represent a comparable risk to presenting TRALL.

The reader can see that the first cutoff point is more restrictive than the second, since between the first and the second one there is a difference of 7 variables. So the variables from the first cutoff point are those with the higher risk, and variables from the second cutoff point represent a slightly lower risk. This fact provides realistic elements to propose a care strategy based on two stages. Initially the physician closely monitors the variables of the first cutoff point, and if those variables indicate high risk in both hits, then risk factors according to the second cutoff point must be checked. Actually if a patient has all the risk factors of the first cutoff point during the first hit, transfusion should be avoided or been performed under extreme precautions.

The knowledge about results of this study, would help medical staff to avoid or to postpone elective surgeries in which a transfusion will take place (high-risk patients), or to establish the most suitable conditions when the surgery and/or transfusion is an emergency.

7. Conclusions

Nowadays, medicine is one of the most notorious areas for applying intelligent techniques. The reason is simple: when preserving health and saving life is possible, every scientist is pleasure to collaborate. Probably one of the most challenging tasks into the study of risk factors related to a medical condition is the set of the cutoff point, when reported literature does not make a suggestion. Therefore, the researchers who worked on this research decided to incorporate an automatic system (with no expert medical knowledge), that sets a cutoff point based on the relative importance of all the factors involved (except those variables whose value is 100%).

Artificial intelligence is being applied to medicine as a natural tool for discover and verification or rejection of health theories. In the case reported in this document, physicians were surprised about the fact that smoking index resulted so related with TRALI.

It can be established that in the case of TRALI in Mexico there are two critical variables of the first hit: age and the ratio PaO2/FiO2 (better known as Kirby’s index).

Regarding age it was found that there are two critical age groups: patients between 41 and 50 years old, as well as older than 71 years old. It should be noted that no one patient was 81 or more years old. 100% of patients with 71 or more years old presented the syndrome, while 63% of patients between 41 and 50 years old also presented it.

In relation to the Kirby’s index, this research confirms what is consistently mentioned in the literature, an index of less than 300 is a critical risk factor for TRALI.

The approached problem, is a challenging task for artificial intelligence area because it constitutes a very difficult to diagnose syndrome. Actually, nowadays the Centenary Miguel Hidalgo Hospital does not routinely runs the tests required to diagnose this medical condition, however, as more results are disseminated as
the reported in this paper, transfusion medicine will become stronger.

Today TRALI is recognized as the leading cause of transfusion-related death in the United States of America, and the second in the United Kingdom.

Although TRALI is not easy to diagnose, physicians can take care of critical risk factors with an insignificant cost but with very important benefits. When medical staff has relevant information about factors that imply a high risk of any disease, keep care of them is natural and easy. This is why the identification of TRALI’s risk factors is very transcendent.

We can also add that the combination of mechanisms such as those proposed in the algorithm TRALI_GAA, provided an excellent hybridization for the problem we addressed. It is well known that including as much knowledge as we can to an algorithm, produces better results. So the use of typical testors, the accelerating operator, the improvement mechanism, and the automatically computation of the cutoff point make possible to have a powerful tool to complex real life problems.

The technique applied in this work, overcame the reported problem of logistic regression when the sample size is small; problem frequently found in the analysis of rare diseases.

Finding typical testors of a dataset is an exponential problem, this is the reason why applying a meta-heuristic in the calculus is well justified. Informational weight of variables is also a very important concept, since it let us to qualify the importance of each variable, and it is also the base for the cutoff point proposed by the algorithm (the first cutoff point).

Finally, based on the second cutoff point, it is feasible to propose a strategy for the health care in two stages: the first lies in monitoring the risk factors identified by the first cutoff point, and the second in checking the factors identified by the second cutoff point.

References