UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 2611-0172, volume III, 2019

Random forest analysis: a new approach for classification of Beta Thalassemia

Massimiliano Sacco · Mariangela Sciandra · Aurelio Maggio

Abstract In recent years, Thalassemia care providers started classifying patients as transfusion-dependent-Thalassemia (TDT) or non-transfusion-dependent-Thalassemia (NTDT) owing to the established role of transfusion therapy in defining the clinical complication profile, although this classification was also based on expert opinion and is limited by reliance on patients'current transfusion status. Starting from a vast set of variables indicating severity phenotype, through the use of both classification and clustering techniques we want to explore the presence of two (TDT vs NTDT) or more clusters, in order to approaching to a new definition for the classification of Beta-Thalassemia in Thalassemia Syndromes (TS).

Keywords Random forest · Unsupervised classification · Clustering · Thalassemia

Riassunto *Negli ultimi anni, gli operatori sanitari hanno iniziato a classificare i pazienti come pazienti trasfusione-dipendenti-Talassemia (TDT) o non trasfusione-dipendenti-Talassemia (NTDT) a causa del ruolo consolidato della terapia trasfusionale nella definizione del profilo clinico delle complicanze, sebbene questa classificazione sia stata basata anche sull'opinione di esperti e sia limitata dall'attuale stato trasfusionale dei pazienti. Partendo da un vasto insieme di variabili che indicano la gravit del fenotipo, attraverso l'uso di tecniche di classificazione e di clustering vogliamo esplorare la presenza di due (TDT vs NTDT) o pi cluster, al fine di identificare una nuova classificazione della Beta-Talassemia nella Sindrome di Talassemia (TS).*

M. Sciandra
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy
E-mail: mariangela.sciandra@unipa.it

M. Sacco, A. Maggio
Campus di Ematologia "Franco e Piera Cutino", c/o A.O.O.R. Villa Sofia - Cervello, Palermo
E-mail: maxxsacco@gmail.com, aurelio.maggio@ospedaliriunitipalermo.it

Parole chiave *Random Forest - classificazione non supervisionata - raggruppamemto - Talassemia*

1 Introduction

The term *Thalassemia* (Taher et al., 2018) is derived from the greek words *Thalassa* (sea) and *Haema* (blood) and refers to disorder associated with defective synthesis of a α or β-globin subunits of Haemoglobin (Hb), inherited as pathologic alleles of one or more of the globin genes located in chromosomes 11 (β) and 16 (α). Haemoglobin is the oxygen-carrying component of the red blood cells and it consists of these two different proteins, alpha and beta. People whose haemoglobin does not produce enough alpha protein have alpha thalassemia, instead People whose haemoglobin does not produce enough Beta protein have Beta thalassemia. If the body doesn't produce enough of either of these two proteins, the red blood cells do not form properly and cannot carry sufficient oxygen. The result is anaemia that begins in early childhood and lasts throughout life.

In the past 20 years, the two important forms of this disorder, Alfa and Beta Thalassemia, resulting from the defective synthesis of the alfa and beta globin chains of haemoglobin, respectively, have become recognized as the most common monogenic disease in humans (Weatherall and Clegg, 1996). Alfa Talassemia is commonly found in Africa, the Middle East, India, Southeast Asia, southern China, and occasionally the Mediterranean region. Beta-thalassemia (Galanello and Origa, 2010) is prevalent in Mediterranean countries, the Middle East, Central Asia, India, Southern China, and the Far East as well as countries along the north coast of Africa and in South America. The highest carrier frequency is reported in Cyprus (14%), Sardinia (10.3%), and Southeast Asia. The high gene frequency of beta-thalassemia in these regions is most likely related to the selective pressure from Plasmodium falciparum malaria. Population migration and intermarriage between different ethnic groups has introduced thalassemia in almost every country of the world, including Northern Europe where thalassemia was previously absent. It has been estimated that about 1.5% of the global population (80 to 90 million people) are carriers of beta-thalassemia, with about 60,000 symptomatic individuals born annually, the great majority in the developing world. β-thalassemia is estimated to affect approximately 1 in 100,000 individuals in the general population. Thalassemia resulted in 25,000 deaths in 2013 down from 36,000 deaths in 1990. People with moderate and severe forms of thalassemia usually find out about their condition in childhood, as they have symptoms of severe anaemia early in life. Because thalassemias are inherited, the condition usually runs in families. Some people find out about their thalassemia because they have relatives with a similar condition. The signs and symptoms of thalassemia major appear within the first 2 years of life (GBD, 2015). Children developing life-threatening anaemia do not gain weight and grow at the expected rate and may develop yellowing of the skin and whites of the eyes. The hallmark of the disease process in homozygous or compound heterozygous patients is characterized by an imbalance in the alpha/beta-globin chain ratio, ineffective erythropoiesis and reduced red blood cell survival -

leading to variable degrees of chronic anaemia and transfusion requirement. For the autosomal recessive forms of the disease, both parents must be carriers for a child to be affected. If both parents carry a hemoglobinopathy trait, the risk is 25% for each pregnancy for an affected child Thalassemia can be diagnosed via a complete blood count, hemoglobin electrophoresis, and DNA testing (Vichinsky, 2005). In this paper a new classification of phenotype severity (mild, moderate, severe) in patients with beta-thalassemia is proposed. In Sec. 2 we start with a brief remaind to the standard patients'classification (TDT vs NTDT); in Sec. 3 statistical classification methods are discussed and the importance of using random forest techniques is emphasized; results on real data are shown in Sec. 4; future work and conclusion follow in Sec. 5.

2 TDT vs NTDT Razionale

Historically, patients with beta-thalassemia were classified as having beta-thalassemia major or intermedia based on few clinical parameters such as hemoglobin level, growth retardation, splenomegaly and other clinical findings or determined through expert opinion, although never formally validated (Weatherall and Clegg, 2001). In more recent years, thalassemia care providers started classifying patients as transfusion-dependent-thalassemia (TDT) or non-transfusion-dependent thalassemia (NTDT) owing to the established role of transfusion therapy in defining the clinical complication profile, although this classification was also based on expert opinion and is limited by reliance on patients' current transfusion status. Non-transfusion dependent thalassemia (NTDT) represents a group of thalassemic disorders including patients who do not require frequent blood transfusions for survival (Musallam et al., 2013). Patients with NTDT may still require occasional or more frequent red blood cell (RBC) transfusion therapy in certain circumstances including but not limited to significant infection, pregnancy, periods of rapid growth, or surgery. On the contrary, people with TDT require lifelong supportive care with regular blood transfusions - typically given every two to five weeks - to lessen the symptoms of anaemia (Cappellini MD and V., 2014). Without regular blood transfusions, people with TDT cannot survive (Rachmilewitz and Giardina, 2011). Despite the availability of supportive care, many people with TDT experience serious complications and organ damage due to iron overload. Classification of phenotype severity in patients with beta-thalassemia has so far relied mainly on expert opinion using parameters of genotype, clinical features at diagnosis, and transfusion requirement. This classification does not provide a specific definition of transfusion requirement to label patients as TDT or NTDT, and both subtypes can become interchangeable as disease progresses or with the introduction of novel therapies that alter transfusion requirement (Modell B, 1984). Starting from a vast set of variables indicating severity phenotype, through the use of both classification and clustering techniques we want to explore the presence of two (TDT vs NTDT) or more clusters, in order to approaching to a new approach for the classification of beta-thalassemia in Thalassemia Syndromes (TS). The aim is, therefore, to review and assess the IPhS importance in defining TS groups.

2.1 International Working Group (IWG) on Thalassemia

An International Working Group (IWG) on Thalassemia was formed and met in Palermo, Italy, on September 15th and 16th, 2017. A group of experts with decades of experience in managing patients with beta-thalassemia agreed on a set of clinical variables to be collected for further exploration of their merit as IPhS, primarily based on earlier work by colleagues as well as expert opinion. An International Health Repository (IHR) protocol, approved on May 25th, 2017 by the Italian Ethical Committee (EudraCT and Sponsor's Protocol Code Numbers were 2017-004457-17 and 143AOR2017) was established to allow collection of relevant data. Thirteen international thalassemia centers of excellence from seven different countries participated in retrospective data collection.

2.2 Data collection

The dataset contains information about Beta-thalassemia patients who attended the centres from 1976 to 2018. Data from 7,910 patients were collected (Beta-thalassemia diagnosis of these patients was defined based on Modell5, 1984 criteria). Patients born after prenatal diagnosis and subjects with a carrier state and with hemoglobin E/beta-thalassemia were excluded. Twenty different clinical variables were collected on each patient; they included key clinical findings at diagnosis and follow-up, management approaches, morbidities, and mortality. By using these variables, a combination of techniques of cluster and classification analyses was performed to check variable importance in clustering beta-thalassemia phenotype and to achieve the goal of this study.

Country	**NTDT***	**TDT***	**Total**
Egypt	28	902	930
Iran	710	1242	1952
Italy	1054	3295	4349
Oman	43	181	224
Pakistan	142	151	293
Saudi Arabia	1	29	30
USA	49	83	132
Total	2027	5883	7910

Table 1: Distribution of patients in the Thalassemia International Health Repository.

3 Statistical methods

Cluster analysis is a data-reduction technique designed to uncover subgroups of observations within a dataset (Fabbris, 1997). It allows to reduce a large number of observations to a much smaller number of clusters or types. A cluster is defined as a group of observations that are more similar to each other than they are to the observations in other groups. The two most popular clustering approaches are hierarchical agglomerative clustering and partitioning clustering. In agglomerative hierarchical clustering, each observation starts as its own cluster. Clusters are then combined, two at a time, until all clusters are merged into a single cluster. In the partitioning approach, the number of clusters K is specified and observations are then randomly divided into K groups and reshuffled to form cohesive clusters. Within each of these broad approaches, there are many clustering algorithms to choose from. For hierarchical clustering, the most popular are single linkage, complete linkage, average linkage, centroid, and Ward's method. For partitioning, the two most popular are k-means (only for continuous variables) and partitioning around medoids (PAM) (for mixed variables). An effective cluster analysis is a multistep process with numerous decision points. Each decision can affect the quality and usefulness of the results (Bosio A.C., 2009). On the other hand, classification analysis is used when there is the need to predict a categorical outcome from a set of predictor variables. The goal is to find an accurate method of classifying new cases into one of the two groups. So, given a predictor of variables x, and a categorical response variable y, there is the need to build a model for predicting the value of y for a new value of x or more generally, for understanding the relationship between x and y. Classification methods include linear discriminant analysis, logistic regression, nonparametric methods as the nearest neighbour classifiers and classification trees, machine learning methods as Bagging, Support Vector Machines (SVM) and Random Forest (RF).

In this work, PAM for clustering and RF both for clustering and for classification were used (Lesmeister, 2017).

3.1 Partitioning Around Medoids

PAM method has several advantages: it's not sensitive to outliers as methods based on means, it can handle large data set, and it can accommodate mixed data types when necessary. In PAM method, each cluster is identified by its most representative observation (called a medoid).

The PAM algorithm works as follows:

1. Randomly select K observations (call each a medoid).
2. Calculate the distance/dissimilarity of every observation to each medoid.
3. Assign each observation to its closest medoid.
4. Calculate the sum of the distances of each observation from its medoid (total cost).
5. Select a point that isn't a medoid, and swap it with its medoid.
6. Reassign every point to its closest medoid.

7. Calculate the total cost.

8. If this total cost is smaller, keep the new point as a medoid.

9. Repeat steps 5-8 until the medoids don't change.

A medoid is an observation of a cluster that minimizes the dissimilarity between the other observations in that cluster.

3.2 Random Forest

A random forest is an ensemble learning approaches for supervised and unsupervised learning. Supervised RF allow for classification, while unsupervised RF allow for clustering. Multiple predictive models are developed, and the results are aggregated to improve classification rates. The algorithm for a RF involves sampling cases and variables to create a large number of decision trees. Each case is classified by each decision tree.
Supervised RF is a classification analysis where the outcome need to be specified. Assume that N is the number of cases in the training sample and M is the number of variables. Then the algorithm works as it follows:

1. Grow a large number of decision trees by sampling N cases with replacement from the training set;

2. Sample $m < M$ variables at each node. These variables are considered candidates for splitting in that node. The value m is the same for each node;

3. Grow each tree fully without pruning (the minimum node size is set to 1);

4. Terminal nodes are assigned to a class based on the mode of cases in that node;

5. Classify new cases by sending them down all the trees and taking a vote-majority rules.

An out-of-bag (OOB) error estimate is obtained by classifying the cases that aren't selected when building a tree, using that tree. This is an advantage when a validation sample is unavailable. Finally, the validation sample is classified using the RF and the predictive accuracy is calculated. Random forests tend to be very accurate compared with other classification methods. Additionally, they can handle large problems (many observations and variables), can handle large amounts of missing data, and can handle cases in which the number of variables is much greater than the number of observations. The provision of OOB error rates and measures of variable importance are also significant advantages. A significant disadvantage is that it's difficult to understand the classification rules (there are 500 trees) and communicate them to others.

4 Results

Using the different clinical variables that were collected on each patient, was performed a preliminary cluster analysis to check variable importance in clustering beta-thalassemia phenotype. Results showed that the most important variables included age at diagnosis, age at first transfusion, and age at first iron chelation, commonly termed “onset variables”. To exclude the possible influence of treatment during the clinical course of the disease, only these variables that were retrieved at the time of diagnosis were considered as IPhS and used in further analysis of clustering severity classes. The remaining clinical variables were used for comparisons of clinical profile between the severity classes identified.

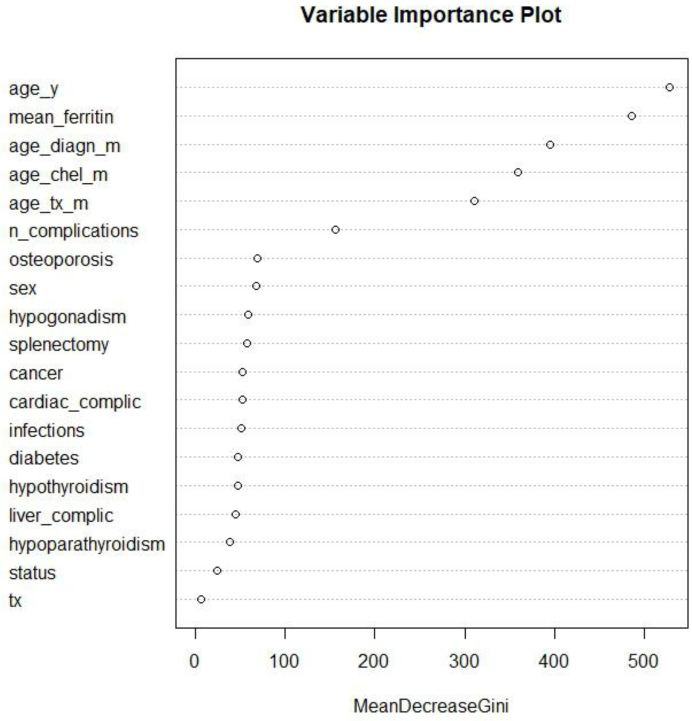
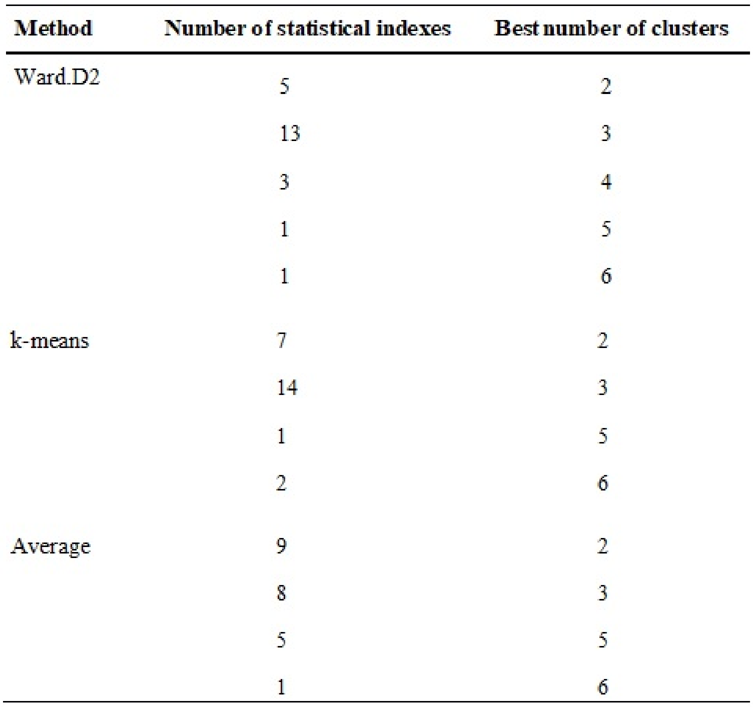


Fig. 1: Variable Importance plot of phenotype severity indicators in beta-thalassemia from unsupervised Random Forest analysis considering all available variables.

As mentioned before, in order to assess the importance of the onset variables in classifying the severity of beta-thalassemia, a combination of cluster and classification analyses were applied. The cluster analysis approach was used to find the underlying population substructure using potential IPhS data without considering prior information (Unsupervised Method). Instead, the classification works on group data based on predetermined classes, while developing criteria for distinguishing between classes (Supervised Method). The statistical approach started with the onset variables, involved the application of the following steps: 1) exploring the existence of a population substructure and for determining the best number of clusters in our data set through the use of NbClust R Package; 2) Unsupervised Random Forest (RF) clustering and the Partitioning Around Medoids (PAM) algorithm to define beta-thalassemia clusters. The different types of clustering methods using *NbClust* R Package (Charrad et al., 2014) are shown in Table 1.



Method	Number of statistical indexes	Best number of clusters
Ward.D2	5	2
	13	3
	3	4
	1	5
	1	6
k-means	7	2
	14	3
	1	5
	2	6
Average	9	2
	8	3
	5	5
	1	6

Table 1 Results of *NbClust* package using different types of clustering methods.

Data suggest that, using the onset variables as IPhS, three clusters reflecting three classes of phenotype severity emerge, can be identified, and it is true for all the majority rules (13 for Ward D2, 14 for k-means, 8 for average) inside the three used methods. Unsupervised RF and PAM algorithms were used and showed that the three obtained classes are well clustered with minimal overlapping (Fig. 2).

Supervised RF classification analysis was used to assess classification error rate considering the three obtained clusters. The classification estimated error rate was 3.3% (Table 2).

The overall accuracy of this model in predicting the three classes with different phenotype severity in Beta-Thalassemia was 96.7% (Table 2). Comparative analyses of the onset variables among the three identified classes of phenotype severity are shown in Table 3.

Age at diagnosis, at first transfusion, and at first chelation were significantly younger in severe, followed by moderate and mild patients. In Table 4 is represented the comparison of clinical findings among the three classes:

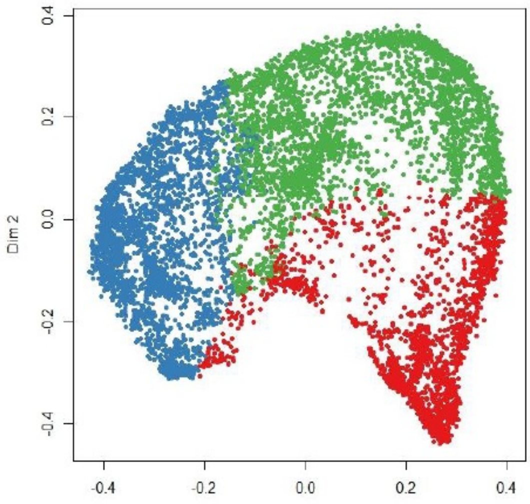
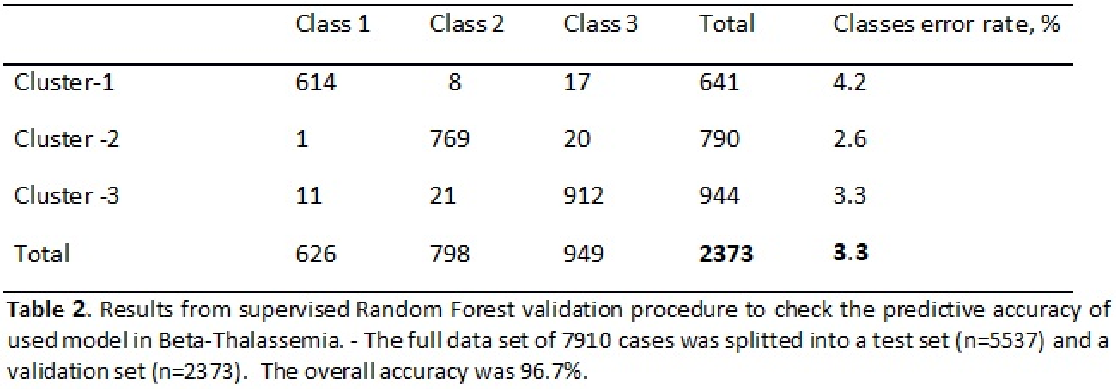


Fig. 2: The three obtained classes, clustered by Unsupervised RF and PAM algorithms

.



	Class 1	Class 2	Class 3	Total	Classes error rate, %
Cluster-1	614	8	17	641	4.2
Cluster -2	1	769	20	790	2.6
Cluster -3	11	21	912	944	3.3
Total	626	798	949	**2373**	**3.3**

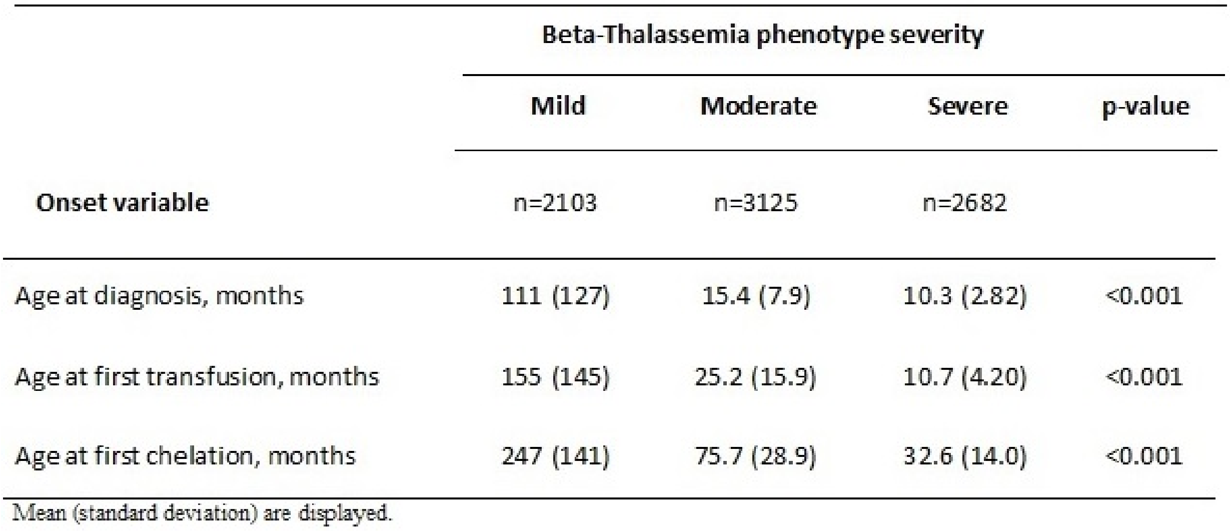
Table 2. Results from supervised Random Forest validation procedure to check the predictive accuracy of used model in Beta-Thalassemia. - The full data set of 7910 cases was splitted into a test set (n=5537) and a validation set (n=2373). The overall accuracy was 96.7%.

The distribution of TDT and NTDT within the three cluster (Mild, Moderate, Severe) is shown in the following Table 5:

The comparison between TDT vs NTDT and the distribution of them within the three cluster is represented in Figure 3:

5 Conclusions

Our findings suggests that, on the basis of onset variables, there exists three classes of phenotype severity in patients with beta-thalassemia (mild, moderate, severe). The mild class includes



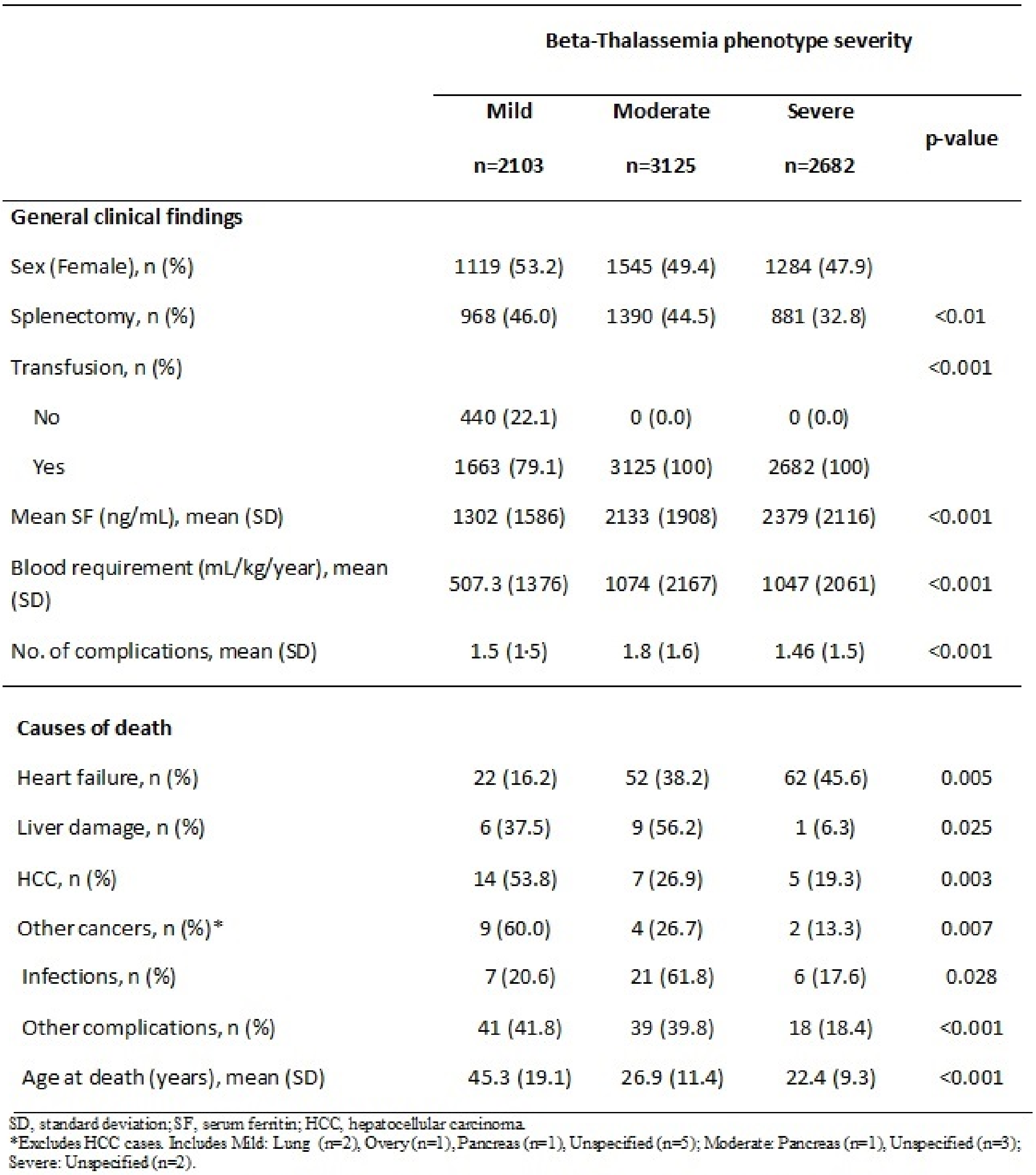
	Beta-Thalassemia phenotype severity			
	Mild	Moderate	Severe	p-value
Onset variable	n=2103	n=3125	n=2682	
Age at diagnosis, months	111 (127)	15.4 (7.9)	10.3 (2.82)	<0.001
Age at first transfusion, months	155 (145)	25.2 (15.9)	10.7 (4.20)	<0.001
Age at first chelation, months	247 (141)	75.7 (28.9)	32.6 (14.0)	<0.001

Mean (standard deviation) are displayed.

Table 3. Onset variables across the identified three classes of phenotype severity in Beta-Thalassemia.

patients with later age at diagnosis, at transfusion initiation, and at chelation initiation. On the contrary, the severe class encompasses patients with very early age at diagnosis and initiation of transfusion and iron chelation therapy chelation; while moderate class patients come in between. This classification proved robust based on a low error rate and a high predictive rate of accuracy. When looking at the clinical characteristics of the three classes, mild patients showed lower blood requirement and iron overload level and a generally lower complication rate. Mortality from heart failure was also low, and generally occurred at an older age (signifying longer survival). The higher ratio of deaths from liver damage, hepatocellular carcinoma and other cancer in this mild patient group may be attributed to the longer survival and exposure to carcinogenic risk factors including iron overload. Severe patients on the other hand were characterized by higher blood requirements and iron overload and an earlier onset of death with a high proportion attributed to heart failure. The lower prevalence of morbidity may be attributed to the earlier age at death. The moderate class shows an intermediate clinical profile between mild and severe patients, except for blood requirement which was similar between the two groups.

However, despite blood requirement similarity, the age at first transfusion was significantly earlier in the severe class (Table 5), suggesting longer natural history of iron burden. The higher prevalence of infections as a cause of death in mild and moderate classes may be attributed to the high rate of splenectomy in these two classes, which may be secondary to receiving less or delayed transfusions. As shown in figure 3 TDT vs NTDT, our analysis may suggest (Musallam et al., 2013) that the could be possible to think about 3 clusters, or as a continuum of the transfusion requirement, instead of the “dichotomous” classification TDT vs NTDT. These categories may thus be interchangeable depending on worsening of the clinical picture or intro-

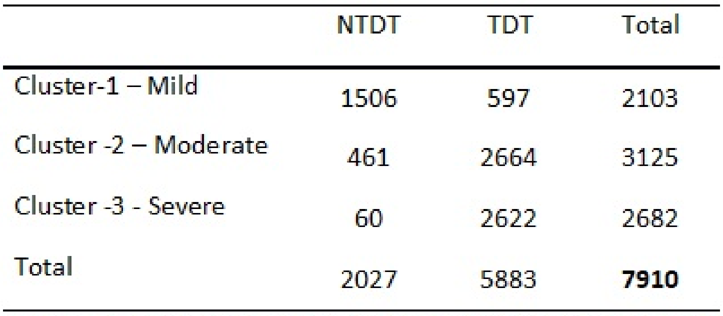


	Beta-Thalassemia phenotype severity			
	Mild n=2103	Moderate n=3125	Severe n=2682	p-value
General clinical findings				
Sex (Female), n (%)	1119 (53.2)	1545 (49.4)	1284 (47.9)	
Splenectomy, n (%)	968 (46.0)	1390 (44.5)	881 (32.8)	<0.01
Transfusion, n (%)				<0.001
No	440 (22.1)	0 (0.0)	0 (0.0)	
Yes	1663 (79.1)	3125 (100)	2682 (100)	
Mean SF (ng/mL), mean (SD)	1302 (1586)	2133 (1908)	2379 (2116)	<0.001
Blood requirement (mL/kg/year), mean (SD)	507.3 (1376)	1074 (2167)	1047 (2061)	<0.001
No. of complications, mean (SD)	1.5 (1·5)	1.8 (1.6)	1.46 (1.5)	<0.001
Causes of death				
Heart failure, n (%)	22 (16.2)	52 (38.2)	62 (45.6)	0.005
Liver damage, n (%)	6 (37.5)	9 (56.2)	1 (6.3)	0.025
HCC, n (%)	14 (53.8)	7 (26.9)	5 (19.3)	0.003
Other cancers, n (%)*	9 (60.0)	4 (26.7)	2 (13.3)	0.007
Infections, n (%)	7 (20.6)	21 (61.8)	6 (17.6)	0.028
Other complications, n (%)	41 (41.8)	39 (39.8)	18 (18.4)	<0.001
Age at death (years), mean (SD)	45.3 (19.1)	26.9 (11.4)	22.4 (9.3)	<0.001

SD, standard deviation; SF, serum ferritin; HCC, hepatocellular carcinoma.
*Excludes HCC cases. Includes Mild: Lung (n=2), Overy (n=1), Pancreas (n=1), Unspecified (n=5); Moderate: Pancreas (n=1), Unspecified (n=3); Severe: Unspecified (n=2).

Table 4. Clinical findings across the identified three classes of phenotype severity in Beta-Thalassemia

duction of therapies. Moreover, although transfusion requirement in TDT is clear (lifelong and



	NTDT	TDT	Total
Cluster-1 – Mild	1506	597	2103
Cluster -2 – Moderate	461	2664	3125
Cluster -3 - Severe	60	2622	2682
Total	2027	5883	**7910**

Table 5. distribution of TDT and NTDT within the three cluster (Mild, Moderate, Severe)

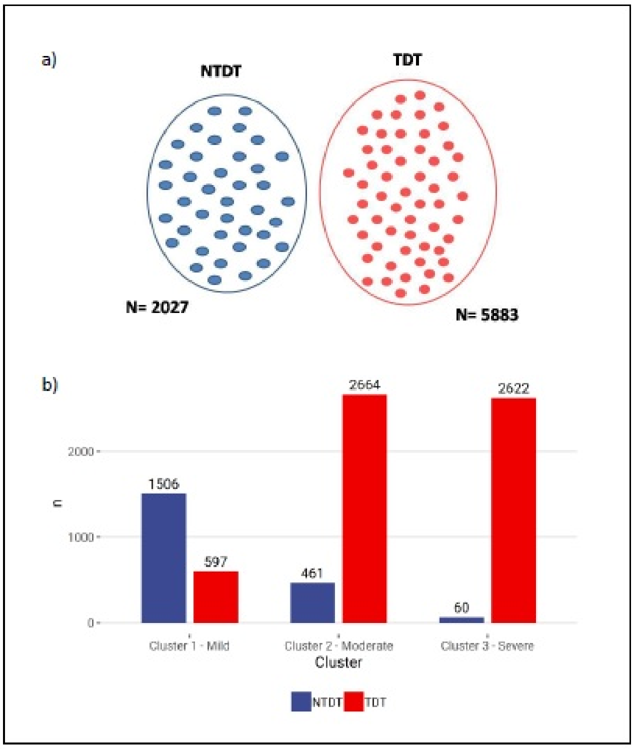


Fig. 3: (a) Classification TDT vs NTDT compared with (b) distribution of TDT vs NTDT in 3 different Cluster (Mild, Moderate, Severe).

regular), in NTDT patients transfusion requirement can be seldom, occasional or even regular (but temporary).

This classification however should be interpreted with caution because the present analysis has several limitations. It does not include patients with haemoglobin E/beta-thalassemia or alpha-thalassemia as they are mostly clustered in geographic areas not included in those centers participating in this working group. We also excluded patients with beta-thalassemia carrier state. Moreover, our findings could suggest a further evaluation in prospective studies to determine specific thresholds for these parameters that can aid physicians in assigning classes and tailoring care accordingly and could suggest to instensity the analysis through the use of techniques, such as consensus clustering, which is well suitable for the kind of data in our possession.

References

Bosio A.C., Renata Metastasio, F. C. (2009). *Procedure di analisi dei dati con il programma SPAD.* Milano: Franco Angeli.

Cappellini MD, Cohen A, P. J. T. A. and V. V. (2014). *Guidelines for the management of transfusion dependent thalassaemia (TDT)* (3 ed.). Nicosia, Cyprus: Thalassaemia International Federation.

Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software 61*(6), 1–36.

Fabbris (1997). *Statistica multivariata : analisi esplorativa dei dati.* Milano: McGraw-Hill Libri Italia.

Galanello, R. and R. Origa (2010, May). Beta-thalassemia. *Orphanet J Rare Dis 5*, 11.

GBD (2015, Jan). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet 385*(9963), 117–171.

Lesmeister (2017). *Mastering machine learning with R: Advanced prediction, algorithms, and learning methods with R* (2 ed.). Birmingham, UK: Packt Publishing.

Modell B, B. V. (1984). *The Clinical Approach to Thalassaemia.* New York and London: Grune and Stratton.

Musallam, K. M., S. Rivella, E. Vichinsky, and E. A. Rachmilewitz (2013, Jun). Non-transfusion-dependent thalassemias. *Haematologica 98*(6), 833–844.

Rachmilewitz, E. A. and P. J. Giardina (2011, 09). How I treat thalassemia. *Blood 118*(13), 3479–3488.

Taher, A. T., D. J. Weatherall, and M. D. Cappellini (2018, 01). Thalassaemia. *Lancet 391*(10116), 155–167.

Vichinsky, E. P. (2005). Changing patterns of thalassemia worldwide. *Ann. N. Y. Acad. Sci. 1054*, 18–24.

Weatherall, D. and J. Clegg (2001). *The Thalassaemia Syndromes* (4 ed.). Oxford: Blackwell Science.

Weatherall, D. J. and J. B. Clegg (1996, Aug). Thalassemia–a global public health problem. *Nat. Med. 2*(8), 847–849.