UNIVERSITÀ DEGLI STUDI DI GENOVA

SCUOLA DI SCIENZE MEDICHE E FARMACEUTICHE

Scuola di Specializzazione in Pediatria

UOC Ematologia Pediatrica IRCCS Istituto Giannina Gaslini

TESI DI SPECIALIZZAZIONE IN PEDIATRIA

MACHINE LEARNING CONTRIBUTION IN THE DIAGNOSTIC WORK-UP OF TELOMERE BIOLOGY DISORDERS

Relatori:

Prof. Mohamad Maghnie

Dr. Carlo Dufour

Candidato:

Dr. Luca Arcuri

Anno accademico 2021/2022

TABLE OF CONTENTS

ABSTRACT

Telomere biology disorders (TBDs) are a heterogeneous group of diseases characterized by germline mutations in genes encoding proteins involved in telomere length (TL) homeostasis. Telomere shortening has been shown to cause serious, multifaceted degenerative diseases, including bone marrow failure, lung, liver, and vessel diseases, abnormalities of the skin, hair and annexes, oral cavity, eyes, and gastrointestinal tract, and an increased risk of cancer, primarily epithelial cancers of the head and neck. Symptoms may present in extremely variable combinations from paucisymptomatic clinical pictures to very severe conditions. Inheritance patterns may either be autosomal recessive or dominant ox X-linked with some genes (TERT, ACD, RTEL1) having double patterns. Moreover, penetrance may be very variable with phenotypes of fairly different severity across affected individual of the same family. All this makes the diagnosis of TBD very challenging.

In order to reduce the diagnostic difficulties, we turned to Artificial Intelligence (AI) to see, whether by using machine learning (ML) unsupervised analysis, specific clusters existed in which TBD patients could be located. Starting population was a heterogeneous cohort of patients (and their relatives) with cytopenia or other clinical features suggestive of TBD referred at the Hematology Unit of the Giannina Gaslini Institute. After cluster identification by using an “a posteriori analysis” we tried to identify clinical features characterizing patients associated to specific clusters with the final goal to implement the diagnostic work-up for TBDs.

In more detail, the starting cohort included 140 patients. For each patient, we evaluated the following characteristics: age, sex, genetics, TL, familiarity with TBD or suspected TBD, presence of cytopenia with distinction between bone marrow failure and peripheral cytopenia, bone marrow karyotype, any feature of immunodeficiency or autoimmunity, presence of typical TBD mucocutaneous alterations, pulmonary disease, hepatopathy, neuro-malformations, ophthalmopathy and any other malformations.

Patients with an unknown genetic status were excluded, and this reduced the size of the original dataset to 92 patients.

By using this new dataset, we performed ML unsupervised analysis ignoring genetic features and excluding the features with more than 15% missing (i.e., bone marrow karyotype, neuro-malformations, ophthalmopathy and pulmonary diseases). We performed dimensionality reduction using multiple correspondence analysis. After silhouette analysis, four clusters, numbered 1 to 4, were chosen as optimal, and clustering analysis was performed using k-means.

After the identification of these four clusters, we performed an "a posteriori analysis" aimed at evaluating the composition of these clusters and to this end we analyzed association between clusters and a number of features characterizing TBD patients.

We observed that TBD related genes were preferentially distributed in cluster n 2 and 3 (p-value ≈ 0.0002). In addition, also TBD related pathogenic variants prevailed in clusters 2 and 3, whereas the VUS and the non-TBD genes mostly fell in clusters 1 and 4 (p=0.0155).

Still in this "a posteriori" analysis, patients with TL <1st or <10th percentile were significantly associated to clusters 2 and 3.

We then assessed the distribution of clinical features evocative of TBD in the identified clusters.

Patients with lung or neurological disease, features that were excluded in the data set construction for cluster analysis, and for this reason considered independent and thus more informative, were significantly associated to cluster 2 (p=0.0355 and p=0.0384 respectively).

As for the other non-independent features, combination, BMF ± hepatopathy ± mucocutaneous triad, BMF + hepatopathy + immunodeficiency, hepatopathy + immunodeficiency, mucocuteanous + immunodeficiency, mucocutaneous + autoimmunity, were all differently distributed in the 4 clusters with a clear prevalence in cluster 2. The combinations hepatopathy + autoimmunity instead showed a tendency to be located in cluster 2.

Taken together these findings indicate that, with use of ML tools, it is possible to identify from a heterogeneous starting population with one or more feature compatible with the diagnosis of

TDB specific well-defined clusters. The “a posteriori" analysis showed a significantly different distribution of the features in the 4 clusters with a prevalence of some TBD clinical and genetic features in specific clusters. These features obviously include TL < 10^{th} centile and TBD related pathogenic variants but also “less” typical symptom either alone or in combination such as hepatic disease, autoimmunity, immunodeficiency.

Based on what above it looks advisable that patients carrying one or more of these features undergo the full diagnostic work-up for TBD.

ABBREVIATION

ML machine learning, TBD, telomere biology disorders; DC, Dyskeratosis Congenita; RS Revesz syndrome; CR, Coat’s retinopathy/plus; BMF, Bone marrow failure; AA, aplastic anemia; MDS, myelodysplastic syndrome; AML, acute myeloid leukemia; PF, pulmonary fibrosis; LC, liver cirrhosis; CM, cerebroretinal microangiopathy with calcification and cysts; PN, poikiloderma with neutropenia; RTS, Rothmund-Thomson syndrome; SCID, severe combined immunodeficiency; CVID, common variable immunodeficiency; qPCR, quantitative polymerase chain reaction; HT-STELA, high-throughput single telomere length analysis; Kb, Kilobase, FLOW-FISH, automated multicolor flow cytometry combined with fluorescence in situ; VUS, variants of uncertain significance; AI, artificial intelligence.

INTRODUCTION

At the beginning of the 20th century, it was recognized that chromosomes carry genetic information. H. Muller and B. McClintock reported that broken chromosomes are unstable and susceptible to rearrangements[1] and inferred the existence of a unique structure at the ends of chromosomes necessary for the protection of chromosome ends.

The term telomere was coined from the Greek words “telos” meaning “end” and “meros” meaning “part” and therefore “end-part.”

In 1961, Hayflick demonstrated that human fetal cells possess a maximum number of 50 to 60 potential cellular divisions, and after reaching that number, the cells stop dividing and become senescent. This limit was defined as the Hayflick limit[2]

In the 1970s, Watson and Olovnikow observed asymmetry in linear DNA replication and hypothesized that chromosomal DNA was lost from the ends of the lagging strand with each cellular division, whereas the discharge of the terminal RNA primer led to the progressive shortening of chromosomes.

In 1978, the telomeric DNA in the protozoan Tetrahymena thermophila was sequenced by Blackburn[3] and Gall and later Blackburb and Greider discovered telomerase, a terminal transferase that adds repeated DNA sequences to the chromosomal ends [4], [5].

In recent years, the biology of telomeres has been further studied and many aspects of the biological mechanisms underlying their homeostasis have been explained.

In the context of these studies, the term telomere biology disorders (TBD) has been coined.

TELOMERE PHYSIOLOGY

Telomeres are complexes of repetitive DNA sequences (TTAGGG) and associated proteins located in the terminal parts of chromosomes[6]. Their length typically ranges between 10 and 15 kilobase (Kb) pairs[7].

Their main function is to protect the end of chromosomes by discriminating natural ends from regular incidental DNA breaks. In this manner, telomeres prevent undesirable end-to-end fusion, nucleolytic degradation, and genomic instability [7] Moreover, the DNA copying machinery is incapable of copying all the way to the ends of a DNA molecule and the telomere prevents the loss of DNA material that would occur during each mitosis.

The non-coding DNA sequences of telomeres undergo destruction during DNA replication, and when telomeres are too short, cells stop to replicate and enter senescence or apoptosis. The accumulation of senescent cells, following the shortening of telomeres, leads to tissue and body aging [8].

Telomere shortness is partially avoided by “telomere maintenance mechanisms”[7], which consist of telomerase-mediated telomere maintenance (the most important) and alternative lengthening of telomeres[7]. Telomerase activity is not homogeneous, and its expression in most human tissues is present only during the first weeks of embryogenesis. From the neonatal period onward, telomerase activity is largely repressed, except in certain highly proliferative tissues such as skin, intestine, liver, bone marrow, which are thought to contain stem cell-like subpopulations, as well as in dividing lymphocytes, ovaries, and testes. Telomerase is also upregulated in most cancer cells, reflecting the need for telomere maintenance for proliferative potential [9].

Telomerase is a reverse transcriptase enzyme. It comprises two core components: a protein component with enzymatic activity (encoded by TERT) and an RNA component (encoded by TERC) with a template for telomere repeat addition. The telomerase complex catalyzes the synthesis

and extension of TTAGGG sequences to the telomeric chromosome ends [9]. Several other species-specific accessory proteins complete the telomerase complex and serve as primers and templates for extending telomeric nucleotide repeats[7]. In human cells, these proteins consist of dyskerin (DKC1), Nuclear Protein Family A, member 2 (NHP2), Nuclear Protein Family A, member 3 (NOP10), Nuclear Protein Family A, member 1(GAR1), ATPases pontin and reptin, Nuclear Assembly Factor 1 Ribonucleoprotein (NAF1), and Telomerase Cajal body protein 1 (TCAB1).

DKC1 associates with H/ACA box-containing small nucleolar RNAs (snoRNAs), including TERC. It is also involved in post-transcriptional pseudouridylation and forms a ribonucleoprotein complex with NOP10, NHP2, and GAR1[9]. All these proteins assist in telomere assembly, recruitment, trafficking, and stability[7], [10].

In addition to telomerase activity, maintenance of TL is achieved because their structure allows them not to be recognized as double-strand DNA breaks by DNA damage response machine[11]. This is possible because of their T-loop structure and the group of six proteins called the shelterin complex. Shelterin is formed by telomeric repeat binding factor 1 (TRF1, encoded by TERF1) and telomeric repeat binding factor 2 (TRF2, encoded by TERF2) that bind to duplex telomeric sequences. The complex is completed by protection of telomeres 1 (POT1, encoded by POT1), TRF1-interacting nuclear factor 2 (TIN2, encoded by TINF2), telomere protection protein 1 (TPP1, encoded by ACD), and RAP1 (encoded by TERF2IP). TRF1 and TRF2 recruited these last four proteins [11], [12].

The main functions of the shelterin complex are to generate and stabilize the T-loop structure (TRF2), prevent DNA Damage Response, DDR (TRF2, POT1, TIN2), recruit telomerase (TPP1, POT1, TIN2), and regulate telomere elongation (TRF1, POT1) [11].

Another important element in telomere maintenance is the protein complex CST. This consists of three proteins: conserved telomere protection component 1 (CTC1), suppressor of cdc thirteen 1 (STN1), and telomeric pathway with STN1 (TEN1)[13]. CST is involved in telomere capping, interacts with shelterin proteins, and modulates telomerase access to the telomere [11].

Other proteins involved in telomere homeostasis are the regulator of telomere elongation helicase 1 (RTEL1), which plays a role in regulating TL, DNA replication, and DNA repair [14], and PARN, which is a deadenylase that processes mRNAs and non-coding RNA, including TERC[11].

Finally, it has not been understood yet the role of telomeric repeat-containing RNA (TERRA), transcribed by RNA polymerase II from intrachromosomal telomeric repeats [7] but in literature it is assumed that TERRA may have a probable role in telomere regulation inhibiting telomerase.

TELOMERE BIOLOGY DISORDERS

Telomere biology disorders (TBDs) are defined as a group of illnesses involved in altered telomere maintenance that results in telomere shortening.

The first entity defined as TBD was Dyskeratosis Congenita (DC). In 1998, DKC1 was discovered as a disease-causing gene [15]. Later, the protein dyskerin, encoded by DKC1, was shown to be involved in telomerase function, and the clinical features of DC were related to an excessive shortening of telomeres[16].

In the following years, breakthroughs in discovering telomere biology and genetics led to the definition of an extended spectrum of diseases now known as TBDs, in which telomere shortening is related to different clinical presentations and multi-organ involvement. In TBDs, there are some well-defined clinical entities, such as Dyskeratosis Congenita (DC), and severe forms, such as Hoyeraal-Hreidarsson Syndrome (HH), Revesz Syndrome (RS), and Coats Plus.

CLINICAL PRESENTATIONS

As reported below, classic DC is characterized by a triad of dysplastic nails, lacy reticular pigmentation of the upper chest and/or neck, and oral leukoplakia. The triad is not always present in all patients.

Patients with DC/TBDs have an increased risk of progressive bone marrow failure (BMF), myelodysplastic syndrome or acute myeloid leukemia, solid tumors (usually squamous cell carcinoma of the head/neck or anogenital cancer), and pulmonary fibrosis. Other clinical findings may include eye abnormalities (epiphora, blepharitis, sparse eyelashes, ectropion, entropion, trichiasis), taurodontism, liver disease, gastrointestinal telangiectasias, vascular disease and avascular necrosis of the hips or shoulders. Moreover, some patients may present with a significant developmental delay, cerebellar hypoplasia (HH syndrome), bilateral exudative

retinopathy, and intracranial calcifications (Revesz syndrome and Coats plus syndrome)[17]. The onset and progression of these manifestations are unpredictable. This makes diagnosis very challenging in some cases.

Below, the most important clinical entities and findings are reported.

DYSKERATOSIS CONGENITA (DC)

DC is the prototype of the TBD. All TBDs could have various features of DC. DC is characterized by the presence of the mucocutaneous triad that is reticulated skin pigmentation, nail dystrophy, and pre-cancer lesions represented by oral leukoplakia, not always simultaneously, associated with BMF.

In addition, there could be a broad spectrum of other features involving all systems that are not always present.

Below is a table from the Telomere Biology Disorders Diagnosis and Management Guidelines [18], summarizing all possible clinical features of DC.

PHYSICAL FEATURES	
Mucocutaneous triad	Dystrophic nails
	Lacy reticulated pigmentation, especially neck and thorax
	Leukoplakia (white patches), usually oral
Additional features	
Eyes	Epiphora (tearing), lacrimal duct stenosis, blepharitis, exudative retinopathy
Hair	Early graying, loss, sparse eyelashes
Gastrointestinal	Esophageal stricture; liver fibrosis, cirrhosis; hepatopulmonary syndrome; peptic ulceration, enteropathy
Stature	Short
Dental	Caries, missing teeth, periodontitis, decreased crown/root ratio, taurodontism
Skeletal	Osteoporosis, hip avascular necrosis
Head/ Neurodevelopmental	Microcephaly, cerebellar hypoplasia (ataxia, spasticity, hypotonia), intracranial calcification
Perinatal	Low birth weight, intrauterine growth restriction
Lung	Fibrosis, restrictive; arterio-venous malformations
Males	Small testes, undescended testes; phimosis, meatal stenosis,

	urethral stricture, hypospadias, leukoplakia
Skin	Hyperhidrosis
NEURODEVELOPMENTAL	
Learning disability, developmental delay, intellectual disability, depression, anxiety	
LABORATORY FEATURES	
Blood	Anemia and/or thrombocytopenia, and/or neutropenia
	Pancytopenia
	High MCV for age
	High fetal hemoglobin (Hb F) for age
Bone Marrow	: Hypocellular for age
	Myelodysplastic syndrome: significant dyspoieses (per WHO classification) +/- cytogenetic clone
	Leukemia: > 20% blasts in marrow
Telomeres	Below first percentile for age by automated multicolor flow-FISH in three of four lymphocyte subsets (CD45 + naïve T cells, CD45 - memory T cells, CD20 + B cells, CD57 + NK/NKT cells) and granulocytes
Genes	Pathogenic variant in a DC-associated gene

TABLE 1 FROM TELOMERE BIOLOGY DISORDERS DIAGNOSIS AND MANAGEMENT GUIDELINES, 2ND EDITION, AVAILABLE AT TEAMTELOMERE.ORG [18]

HOYERAAL-HREIDARSSON SYNDROME (HH)

Hoyeraal-Hreidarsson syndrome (HH) should be considered in children with clinical features of DC associated with cerebellar hypoplasia[19]. In HH, additional features with high penetrance include intrauterine growth restriction, developmental delay and intellectual disability, microcephaly-immunodeficiency and often intractable diarrhea. The onset of HH occurs in the early years of life[18].

REVESZ SYNDROME (RS)

A diagnosis of Revesz syndrome (RS) should be considered in a child with clinical features of DC associated with bilateral exudative retinopathy (named bilateral Coats disease). RS patients may present intrauterine growth restriction, sparse hair and intracranial calcification[18].

COATS PLUS

The diagnosis of Coats plus is possible with the presence of a distinctive pattern of intracranial calcification involving the thalamus, basal ganglia, dentate, and deep cortex, with associated leukoencephalopathy and

brain cysts in association with retinal telangiectasia and exudates, osteopenia with a tendency to fracture and poor bone healing, recurrent gastrointestinal hemorrhage due to vascular ectasias in the stomach, small intestine and liver, intrauterine growth restriction, and the possible presence of dystrophic nails, sparse hair and abnormal skin pigmentation[18], [20].

Coats plus and RS share similar features. The necessity of distinguishing between the two syndromes is due to their specific features. Patients with RS frequently have severe bone marrow failure and cerebellar hypoplasia, which is rare in Coats plus. Conversely, in Coats plus, there is a specific and very distinctive pattern of intracranial calcification. As shown below, the two clinical entities are genetically different [18].

HEMATOLOGICAL MANIFESTATIONS IN TBD

The most characteristic clinical feature of TBDs is bone marrow failure (BMF)[17]. BMF is a common manifestation of TBD with a probability of approximately 80% for patients with classic DC to develop at least a single lineage cytopenia by the age of 30 years. Initially, only one cell lineage may be involved, which then progresses into severe pancytopenia and then into myelodysplasia[21].

The differential diagnosis of BMF is with isolated acquired aplastic anemia (AA). It is well established that all possible causes of inherited bone marrow failure in pediatric and adult cases should be excluded in the diagnostic algorithm for AA [22], [23].

It is common in the diagnostic work-up of AA cases to find reduced TL associated with the presence of pathogenic genetic variants of genes causing TBD in patients without any clinical

features of TBD. Distinguishing acquired AA from BMF in TBD is critical because the management is completely different. It is not easy to discriminate between the two conditions when genetics or TL are inconclusive. In these cases, it is useful to deeply investigate family history to identify any relative with a hematological problem [17].

In patients with TBDs, cases of myeloid acute leukemias and myelodysplastic syndromes have been reported [24]. Myelodysplastic syndromes are often the natural evolution of bone marrow failure [18]. American National Cancer Institute analysis of data from 15 years of follow-up in a longitudinal cohort study of patients with TBD showed that the MDS risk was 578-fold higher than that in the normal population, and the median age of MDS onset was 31 years, far earlier than that in the normal population[25]. For AML the risk was 73 times higher than that in normal individuals, whereas the median age of onset was 40 years.

Similar to patients with AA, in suspected cases of MDS or AML, it is important to investigate the presence of any familiarity with hematological problems and any possible other features related to TBD [18].

IMMUNOLOGIC IMPLICATION IN TBD

Lymphopenia is the most common immunological abnormality observed in patients with DC and HH [21], [26] The lymphopenia is characterized by a reduction in B and NK cells. The reduction of B cells determines the possible presence of hypogammaglobulinemia, which could affect all immunoglobulin subtypes (IgG, IgM, or IgA) and subsequently alter the vaccination response.

The involvement of T cells is less frequent in TBDs. The most common features of alterations that affect the T cell compartment are reduction of T cell counts (CD4 and/or CD8 counts), inversion of the CD4/CD8 ratio and a prematurely advanced naive to memory (CD45RA/CD45RO) T cell transition.

It is possible to have patients with clinical features of immunodeficiency, from severe combined immunodeficiency (SCID)-like presentation in infancy to much milder presentation in adolescents mimicking common variable immunodeficiency (CVID). In patients with T cell alterations, it is possible to find opportunistic infection (CMV or Pneumocystis Jiirovecii).

The most prominent immune-mediated clinical feature of infant-onset TBD (mostly DC or HH) is a severe, chronic, non-infectious enteropathy with mucosal inflammation and intractable diarrhea. The pathophysiology of the disease could derive from a defect in the renewal of the digestive epithelium or an impairment of the gut mucosal immune system. In some cases, it has been described as a useful treatment by anti-TNF-alpha monoclonal antibody in these patients.

TELOMERE-MEDIATED PULMONARY FIBROSIS

Telomeric shortening has been linked to pulmonary fibrosis in multiple studies[27]. Several genes involved in telomere maintenance are associated with idiopathic pulmonary fibrosis. In various studies, the percentage of mutations in telomere genes reached approximately 30 % in familiar cases and 10 % in sporadic cases [28]. In approximately 50% of sporadic cases without positive genetics, a shortening of telomeres was observed below the 10^{th} percentile for age, and in 10%, the telomeres fell below the 1^{st} percentile [29]. The absence of a telomere gene mutation in these patients is unclear, but it is assumed that environmental factors such as work exposure or smoke addition may contribute to the pathogenesis of the disease.

LIVER DISEASE-PREDOMINANT PHENOTYPE

In the literature, cases of isolated liver disease with telomere gene mutation have been described[18]. In these patients, the diagnosis is not easy, and some help could be found in looking for details in familiar history with some relative that may present more typical features of TBD. Liver manifestations can involve a wide spectrum of entities, from mildly abnormal liver tests

to fibrosis and advanced cirrhosis [30], portal hypertension, and hepatocellular carcinoma [18], [31].

MALIGNANCIES IN TELOMERE BIOLOGY DISORDERS

In TBDs, there is a significantly increased risk of developing tumors. Telomeres play a significant role in chromosomal stability, and telomere dysfunction has been implicated in cancer biology[21].

Patients with DC have a four-fold higher incidence of cancer than the general population. This incidence increases to approximately 30-times higher in patients who have undergone hematopoietic stem cell transplantation [25], [32].

The most frequent cancer types are myelodysplasia and acute myeloid leukemia [24]. Solid tumors such as head and neck, anogenital squamous cell carcinoma, and non-Hodgkin lymphoma. The risk of cervical squamous cell cancer and nonmelanoma skin cancer has also been reported. The elevated risk of head and neck carcinomas requires regular surveillance of the onset and evolution of precancerous lesions, such as oral leukoplakia[21].

VASCULAR DISEASES

In TBD, life-threatening gastrointestinal bleedings have been reported, mostly due to telangiectatic lesions and sometimes without an identified origin[33]. Other forms of vascular abnormalities include pulmonary arteriovenous malformations and retinal telangiectasias [18]. Pulmonary arteriovenous malformations can be associated with hepatopulmonary syndrome but can also represent an independent pulmonary phenotype of TBD [34].

The physiopathology of vascular disease has not yet been clarified. Some hypotheses include a link between short telomeres and impaired wound healing [35]. Other studies have indicated a possible role of Wnt/beta-catenin signaling in impaired telomere function [36], [37].

CENTRAL NERVOUS SYSTEM INVOLVMENT

Central nervous system involvement has been reported in 10–25% of patients with TBDs. As mentioned above, abnormalities in developmental delay are the primary characteristics of HH, RS, and Coats Plus. In a recent study update, it was reported that TBDs are relevant neurodevelopmental disorders or psychomotor abnormalities and psychiatric disorders such as anxiety disorders or depression[38].

SILENT CARRIERS

During the diagnostic work-up of a patient suspected of TBD, it is possible to find a pathogenic variant, which may lead to the discovery of the same variant in a family member who is asymptomatic. In these silent carriers it is not easy the management because genotype-phenotype correlation is often made highly difficult by several factors, such as the possibility of hypomorphic gene mutations, disease anticipation, and genetic and environmental modifying factors [39].

The same management problem could also affect the offspring of silent carriers because another typical phenomenon of the disease is the anticipation of symptoms. For the disease behavior, the silent carriers' offspring could have the features of the disease in the early stages of life compared with their asymptomatic parents. In these cases, genetic counseling is highly recommended.

Finally, silent carrier status would have significant implications regarding related hematopoietic cell donation because such carriers would be unsuitable donors for individuals affected by bone marrow failure.

DIAGNOSIS OF TBD

TBD has multiple clinical findings and in not classical phenotype patients, the diagnosis could be very challenging.

TELOMERE LENGTH

Over the years, various techniques have been developed to measure TL. The major methods developed to measure TL include automated multicolor flow cytometry combined with fluorescence in situ (flow-fish), Southern blotting analysis of telomere restriction fragments, telomere quantitative polymerase chain reaction (qPCR), and high-throughput single telomere length analysis (HT-STELA). At present, flow-fish is the only test that is clinically available in certified labs and validated for the diagnosis of TBD. There are only a few industrial and academic laboratories worldwide that can perform this test. The other methods find applications only for investigation use.

SOUTHERN BLOTTING ANALYSIS OF TELOMERE RESTRICTION FRAGMENTS

TL can be indirectly measured using a technique called Telomere Restriction Fragment analysis. This technique is a modified Southern blotting assay that measures the heterogeneous range of TLs in a cell population using the length distribution of the terminal restriction fragments [40].

The technique is based on DNA fragmentation by restriction enzymes that do not affect the telomeric sequences. The problem with this method is due to the sub-telomeric regions that contain non-canonical telomeric repetitions. Therefore, it is not completely predictable the amount of non-telomeric sequences undigested by restriction enzymes and those that are analyzed as telomeric sequences.

After DNA fragmentation, gel electrophoresis is performed, and telomeres are detected by Southern blotting.

The average length of telomeres is assessed by comparing the marked DNA with a DNA reference whose fragment length is known [41].

The disadvantages of the analysis of telomere restriction fragments are that it requires a large quantity of DNA and is costly in terms of both money and time. Moreover, it does not permit the measurement of very short telomeres[41].

TELOMERE QUANTITATIVE POLYMERASE CHAIN REACTION (QPCR)

The Q-PCR method is a relatively easy assay that requires small amounts of DNA. The telomere signal is measured and compared with a reference single-copy signal. The ratio of these signals is proportional to the average TL.

The qPCR technique has the advantage of being high-throughput and cheap; however, as mentioned before, it does not provide Kb value of TL but only a relative measurement[42]. It has also been reported that there is a significant result variation within and between samples[42] [43] because of the absence of standardization between laboratories [43]. Lastly, the qPCR assay is not applicable to cases of very short telomeres and in samples of cells that may not be normal diploid and karyotypically stable, such as cancer cells, because the single-gene copy number must be known to perform the assay [42].

HIGH THROUGHPUT SINGLE TELOMERE LENGTH ANALYSIS (HT-STELA)

Single telomere length Analysis (STELA) is a high-resolution single-molecule method for measuring TL distributions [44]. The technique is based on a combination of ligation, PCR-based methods, and Southern blot analysis [42], which measures TL in a single chromosome. The assay permits the exclusion of telomere-adjacent DNA; therefore, it is accurate in measuring only telomere

sequences and additionally permits. The limitation is that, due to the complexity and lack of specificity of individual chromosomal subtelomeric regions, only a small subset of chromosomes (XpYp, 2p, 11q, 12q and 17p) have the necessary criteria to design primers for successful and specific telomeric DNA amplification [42].

AUTOMATED MULTICOLOR FLOW CYTOMETRY COMBINED WITH FLUORESCENCE IN SITU (FLOW-FISH)

Flow-fish is the only clinically certified test to measure TL. The test was developed in 2003 [45], and since then, it has been the gold standard for measuring TL because of its reliable and reproducible results[42]. This assay has various advantages. Most importantly, flow cytometry is specific for telomere sequences and does not measure subtelomeric sequences. Moreover, flow-fish can measure very short telomeres, and flow cytometry may test only specific subsets of cells.

The assay consists of hybridization of chromosomes with fluorescent probes of CCCTAA sequences repeated 3 times and measurement of the amount of fluorescence emitted [46]. The TL in kilobases is obtained by comparing the fluorescent signal from the sample cells with a group of control cells whose length is as long as 1301 Cell Line human that is often used [47]. Control cells are analyzed in the same tube as the sample, and the TL of the cell subset in the sample is expressed as a fraction of the TL of internal control cells. The inclusion of internal control cells in each individual tube permits adequate correction of intra- and inter-experimental variability in hybridization efficiencies between samples.

The limitation of this assay is that it is technically demanding because of several necessary calibrations and controls [48].

For the abovementioned reasons, the flow-fish method remains the most reliable test for measuring TL because of its reproducibility and accuracy. Other methods, such as STELA and qPCR, are less technically demanding, but at the moment they are not certified for clinical use because they cannot intrinsically measure all and only telomere sequences of a nuclear cell.

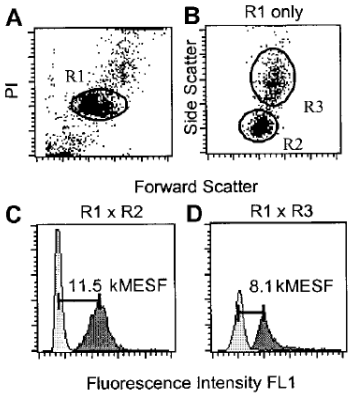

FIGURE 1 [67] FLOW-FISH ANALYSIS OF TL IN GRANULOCYTES AND LYMPHOCYTES

GENETICS OF TBD

The first gene associated with TBD was DKC1. In 1999, Heiss et al. [15] described the association between DKC1 and DC. In 1999, HH was also associated with DKC1. The understanding that these syndromes were related to an alteration of telomere biology was made when dyskerin (encoded by DKC1) was shown to affect telomerase RNA[16]. Fibroblasts and lymphoblasts from patients with DC bearing DKC1 mutations showed low levels of telomerase RNA, reduced telomerase activity, and shortening of the TL.

The discovery of pathogenic variants in TERT and TERC in patients with autosomal dominant forms of DC reinforced the idea of a link between DC and telomere biology [49], [50]. The continuation of studies on telomere biology has permitted the identification of several other genes involved in TBD. Most of these affect telomerase activity, the shelterin complex, and the CST complex, but over the years, other genes with different cellular functions, such as USB1 and MDM4, have been described as causing TBD.

The TBD gene may have an X-linked, autosomal dominant, or recessive pattern depending on the gene. There are few genes that have double inheritance, such as TERT, ACD, NPH2, PARN, and RTEL1. In this case, the specific variant can determine the inheritance pattern.

The table below shows TBD-related genes with transmission characteristics and the most characteristic clinical phenotypes.

GENE	PROTEIN	INHERITANCE	FUNCTION	PREDOMINANT PHENOTYPE
ACD	TPP1	AD, AR	Shelterin component	DC-like
CTC1	CTC1	AR	Telomere extension	DC, CP, CM
DCLRE1B	Apollo	AR	Telomere protection and DNA repair	DC, HH
DKC1	Dyskerin	XLR	Telomerase component	DC, HH
MDM4	MDM4	AD	Telomere associated regulation of p53	DC-like
NAF1	NAF1	AD	Telomerase component	PF
NHP2	NHP2	AR, AD	Telomerase component	DC, PF
NOP10	NOP10	AR	Telomerase component	DC
NPM1	NPM1	AD	No telomere association (rRNA modification)	DC
PARN	PARN	AR, AD	hTR maturation	DC, HH, PF
POT1	POT1	AR	Shelterin component	CP
RPA1	RPA70	AR	Telomere replication and protection	DC-likw, MDS
RTEL1	RTEL1	AR, AD	Telomere replication	DC, HH, PF, MDS, LC
STN1	STN1	AR	Telomere extension	CP
TERC	hTR	AD	Telomerase component	DC, AA, MDS, PF
TERT	hTERT	AD, AR	Telomerase component	DC, AA, MDS, PF, LC
TINF2	TIN2	AD	Shelterin component	DC

USB1	USB1	AR	No telomere association (snRNA maturation)	DC-like, PN, RTS
WRAP53	TCAB1	AR	Telomerase component	DC
ZCCHC8	ZCCHC8	AD	hTR maturation	PF

TABLE 2 FROM TELOMERE BIOLOGY DISORDERS DIAGNOSIS AND MANAGEMENT GUIDELINES, 2ND EDITION, AVAILABLE AT TEAMTELOMERE.ORG [18] AND GENETICS OF HUMAN TELOMERE BIOLOGY DISORDER [51].
ABBREVIATIONS: DC, DYSKERATOSIS CONGENITA; HH, HOYERAAL HREIDARSSON SYNDROME; AA, APLASTIC ANEMIA; MDS, MYELODYSPLASTIC SYNDROME; AML, ACUTE MYELOID LEUKEMIA; PF, PULMONARY FIBROSIS; LC, LIVER CIRRHOSIS; RS, REVESZ SYNDROME; CR, COAT'S RETINOPATHY/PLUS; CM, CEREBRORETINAL MICROANGIOPATHY WITH CALCIFICATION AND CYSTS; PN, POIKILODERMA WITH NEUTROPENIA; RTS, ROTHMUND-THOMSON SYNDROME.

Despite several progresses in the genetic identification of TBD, at present, in approximately 20% of patients with clinical features of TBD, it is not possible to identify a pathogenic variant of TBD.

Silent carriers and variable clinical manifestations depending on age complicate the insight into the genetic determination of TBD. TBD genetics is characterized by incomplete penetrance. The explication of this phenomenon is related to a combination of genetic, environmental, and lifestyle factors. Another possible contribution to this phenomenon has been reported in recent reports of somatic TERT promotor variants as possible modifiers of germline pathogenic mutations [52], [53].

As mentioned above, the presence of silent carriers is common; therefore, it is important to test all family members of affected patients to identify TBD-associated mutations.

Individuals carrying pathogenic mutations with no clinical features of TBD should be counseled regarding their potential risk of disease and transmission to their offspring [17].

In TBD, genetic anticipation is also described. In this case, younger members of the family have more symptoms in childhood, including the most severe, such as the classical ones of DC, whereas the older generation can be asymptomatic or with only single-organ involvement. It is remarkable that in all these reports, the offspring had shorter telomeres than the parents [17].

In TBD it has been described also the phenomenon of somatic revertant mosaicism of a germline mutation of TERC. In the presented cases, mitotic recombination of mutant variant of TERC in

blood cells was spontaneously corrected whereas the mutation was present only in skin fibroblasts [54]. Therefore, mosaicism may have diagnostic implications and can contribute to the explanation of the variable organ penetrance of the disease.

Moreover, in TBD forms due to DKC1 (X-linked pattern), there are some reports of women heterozygous for a pathogenic variant of the gene who have clinical features of the disease, including mucocutaneous triad and bone marrow failure [17]. The exact explanation of the phenomenon remains still not completely clear. In a study of 2016, Xu et al. hypnotized that in addition to X chromosome inactivation, other mechanism such as germline mosaicism or epigenetics, may contribute to TBD phenotypes present in female DKC1 mutation carriers [55].

In TBD, phenocopia is also described. This phenomenon is characterized by a dissociation between the phenotype from the genotype. Phenocopying has been observed in families with TBD, in which individuals, with wild-type genotype inherited short telomeres from a parent and, consequently, exhibited TBD-related symptoms. In these cases, TL returns normal in individuals with wild-type genotypes after two successive generations[51].

Concerning the correlation between gene and phenotype, in 2022 Niewisch et al. [56] reported that in TBD, patients with autosomal dominant inheritance patterns, excluding TINF2, show the best overall survival. Similar to previous studies in patients with autosomal dominant disease, a tendency for the development of pulmonary fibrosis was observed before stem cell transplantation.

Bhala et al. [38] assumed that neurological manifestations are more correlated with autosomal recessive or x-linked patterns except in patients carrying TINF2 (autosomal dominant disease).

Finally, another important issue regarding genetics and TBD is the increasing problem of interpreting variants of unknown significance. The problem is increasing because of the relatively easier access to next-generation sequencing technologies, and it remains one of the most important challenges for the next years in genetic diagnosis of TBD.

DIAGNOSTIC WORK-UP

As reported in the review on Dyskeratosis Congenita and Related Telomere Biology Disorders by Savage et al. [17], TBD diagnosis is established in individuals with suggestive clinical findings associated with shortened TL and pathological genetic features, i.e., biallelic pathogenic (or likely pathogenic) variants in one of the genes known to solely cause autosomal recessive TBD, a heterozygous pathogenic (or likely pathogenic) variant in one of the genes known to cause autosomal dominant TBD, a mono or biallelic pathogenic (or likely pathogenic) variant in one of the genes associated with autosomal recessive and dominant TBD, or a hemizygous pathogenic (or likely pathogenic) variant in DKC1 known to cause X-linked TBD [17].

According to the literature, TL less than the first percentile for age in lymphocytes is 97% sensitive and 91% specific for TBD [17]. TL must be measured using flow-fish. According to the literature a TL on lymphocytes below the 10th percentile must be considered suspicious for TBD [57].

Variants of uncertain significance should not be used in clinical decision-making[58]. According to American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP), it is important to make efforts to resolve if a variant is pathogenic or benign. While this effort to reclassify the variant is underway, additional monitoring of the patient for the disorder in question may be prudent[58].

DIFFERENTIAL DIAGNOSIS

The broad clinical spectrum of TBDs determines the possibility of various differential diagnostic disorders. Other inherited bone marrow failures, such as Fanconi Anemia, Shwachman-Diamond syndrome, and Diamond-Blackfan anemia, must be excluded. In 2015, Alter et al. [57] reported in a large study cohort that TL shortened is not usually in the inherited bone marrow failure mentioned above. In some patients, it is possible to find a reduction in TL but not as short as in TBD patients.

Typical features of the mucocutaneous triad can be found in other genetically determined syndromes[17].

Twenty-nail dystrophy is an autosomal recessive disorder caused by mutation of the FZD6 gene and is characterized by onychauxis, hyponychia, and onycholysis [17].

Nail-patella syndrome is an autosomal dominant syndrome caused by mutation of the gene LMX1B characterized by the classic clinical tetrad of changes in nails, knees, elbows, and presence of iliac horns[17].

Poikiloderma with neutropenia is an autosomal recessive disorder caused by a mutation in the gene USB1 that is characterized by post-inflammatory poikiloderma, and chronic noncyclic neutropenia typically associated with recurrent sinopulmonary infections and bronchiectasis. In this syndrome, there is an increased risk of myelodysplastic syndrome. Nail dystrophy and palmar or plantar hyperkeratosis may be present [17].

LIG4 syndrome is a hereditary disorder associated with impaired DNA double-strand break repair mechanisms caused by biallelic mutation of the homonymous gene. LIG 4 encodes for an ATP-dependent DNA ligase that joins double-strand breaks during non-homologous end joining. The most characteristic clinical features of this syndrome consist in growth restriction, developmental delay, microcephaly, facial dysmorphism, pancytopenia, variable immune deficiency, and an increased predisposition to leukemia[59].

MACHINE LEARNING

Machine learning is a subdomain of artificial intelligence. In recent years, its application has increased in clinical investigations to improve disease classification, risk stratification, and treatment decisions[60]. The machine learning does not work as a “classical statistic” following pre-established rules but rather learns from examples [60].

There are two main algorithms used in machine learning: supervised and unsupervised analysis. Supervised analysis is based on a set of exemplars that automatically predict specific outcomes. Unsupervised analysis identifies patterns or clusters within a dataset without starting from the initial division of data.

To the best of our knowledge, in TBDs, the machine learning approach has never been applied.

AIM OF THE STUDY

The general aim of the study was to see whether ML was able to implement the diagnostic work up of TBD that, due to the extremely variable clinical phenotype, the different and non-unique inheritance pattern and the diverse penetrance may be very challenging and may generate misdiagnoses.

The first specific aim was to evaluate the existence of statistically significant specific clusters in a heterogeneous cohort of patients (and their relatives) with cytopenia or other clinical features suggestive of TBD referred at the Hematology Unit of the Giannina Gaslini Institute.

The second specific aim was to identify clinical and biochemical features, isolated or in combination, should be considered to correctly address the diagnostic work-up of TBD.

MATERIALS AND METHODS

Patients followed at the Hematology Unit of the Giannina Gaslini Institute from 1989 to 2023 were included in the study if they met the following criteria:

1. Patients suffering from persistent cytopenia of at least one of the cellular lineages according to the international consensus [61].

Or

2. Family members of patients with telomere disease or suspected telomere disease and potential donors of hematopoietic stem cells.

Or

3. Patients with phenotypic or genetic anomalies suggestive of dyskeratosis TBD, even in the absence of cytopenia

For each patient, we evaluated the following characteristics: age, sex, genetics, TL, familiarity for TBD or suspected TBD, presence of cytopenia with distinction between bone marrow failure and peripheral cytopenia, bone marrow karyotype, any feature of immunodeficiency or autoimmunity, presence of typical TBD mucocutaneous alterations, pulmonary disease, hepatopathy, neuro-malformations, ophthalmopathy, and any other malformations.

Regarding genetic features, germline variants were identified by either whole-genome sequencing, whole-exome sequencing, targeted panels based on Next-Generation-Sequencing method, or Sanger sequencing.

Variants of only TBD genes were classified as pathogenic (P), likely pathogenic (LP), variants of uncertain significance (VUS), likely benign (LB), or benign (B) according to the current guidelines (Matthijs et al., Guidelines for diagnostic next-generation sequencing [62] and Richards S. et al., Standard and guidelines for interpretation of sequence variants [58]). Starting from the variant definition, a TBD classification was made. A TBD gene was defined “pathogenic” if it carried a biallelic P (or LP) variant in an autosomal recessive gene, a heterozygous P (or LP) variant in an

autosomal dominant gene, or a hemizygous P (or LP) variant in an X-linked gene. A TBD gene was defined "VUS" if carrying biallelic VUS in an autosomal recessive gene, a heterozygous VUS in an autosomal dominant gene, a hemizygous VUS in an X-linked gene in a male, or a biallelic VUS in an x-linked gene in a female. A TBD gene was defined as negative (NEG) in all other cases.

Cytopenia was defined as hemoglobin below the normal limit for age, leukocytes< 4000/µl, neutrophils 1500/µl or platelets < 150000/µl.

Bone marrow failure was defined according to Camitta's criteria[63]-[65]. Patients with peripheral cytopenia with normal bone marrow cellularity were defined as patients with peripheral cytopenia.

The evaluation of the average length of telomeres in granulocytes and lymphocytes was performed at the hematology laboratory in Aachen, Germany, using the modified FISH flow method as reported in the literature by the German group [66], [67].

Starting from TL, two variables were created: patients with TL on lymphocytes below 1st percentile and patients with TL on lymphocytes below 10th percentile.

The presence of immunodeficiency was defined as a reduction in at least one of the main lymphocyte subsets (B cells, T cells or NK cells) or in serum immunoglobulins.

The presence of autoimmunity was defined in patients with significant positivity for antibodies or suggestive markers of autoimmune lymphoproliferation syndrome.

The presence of mucocutaneous alterations was defined in patients with mucocutaneous features described in TBD patients according to the literature.

The presence of pulmonary disease was defined in patients with typical lung findings of TBDs according to the literature. In patients who had never undergone either a functional lung test or computed tomography scan of the thorax, the variable was defined as unknown.

The presence of hepatopathy was defined in patients with an increase in liver enzymes after the exclusion of other causes or in patients with typical liver findings of TBD on abdominal ultrasound,

FibroScan, or magnetic resonance of the abdomen. In patients who had never performed these tests, the variable was defined as unknown.

The presence of neuromalformations was defined in patients with typical cerebral malformations of TBDs or in those with psychomotor retardation. In patients with no psychomotor retardation who had never undergone cerebral magnetic resonance imaging, the variable was defined as unknown.

The presence of ophthalmopathy was defined in patients with typical ocular TBD findings. In patients who had never undergone an ophthalmic evaluation, the variable was defined as unknown.

The presence of other malformations was defined in patients with every possible other type of malformation regardless of being typical of TBDs.

For each patient, we described the following 25 categories (all categorical): “univocal patient number (UPN)”, “sex”, “gene”, “variant”, “zygosity”, “variant classification”, “gene classification”, “inheritance”, “mutations”, “TL on lymphocytes < 1^{st} percentile”, “TL on lymphocytes < 10^{th} percentile”, “age_at_sampling for TL (two classes: <18 years old and > 18 years old)”, “familiarity”, “all cytopenia”, “BMF”, “peripheral cytopenia”, “any autoimmunity”, “any immunodeficiency”, “mucocutaneous alterations”, “pulmonary disease”, “hepatopathy”, “neuro-malformations”, “ophthalmopathy”, “other malformations”, “bone marrow karyotype”.

Patients with unknown genetic status were excluded, and we created a new dataset with dimension of 92 (patients) x 25 (features).

In both the data preprocessing and the analysis, the following features have been ignored: “UPN", “gene”, “variant”, “zygosity”, “variant classification”, “gene classification”, “inheritance”, “mutations”.

In the figure below, the observed values are represented in grey.

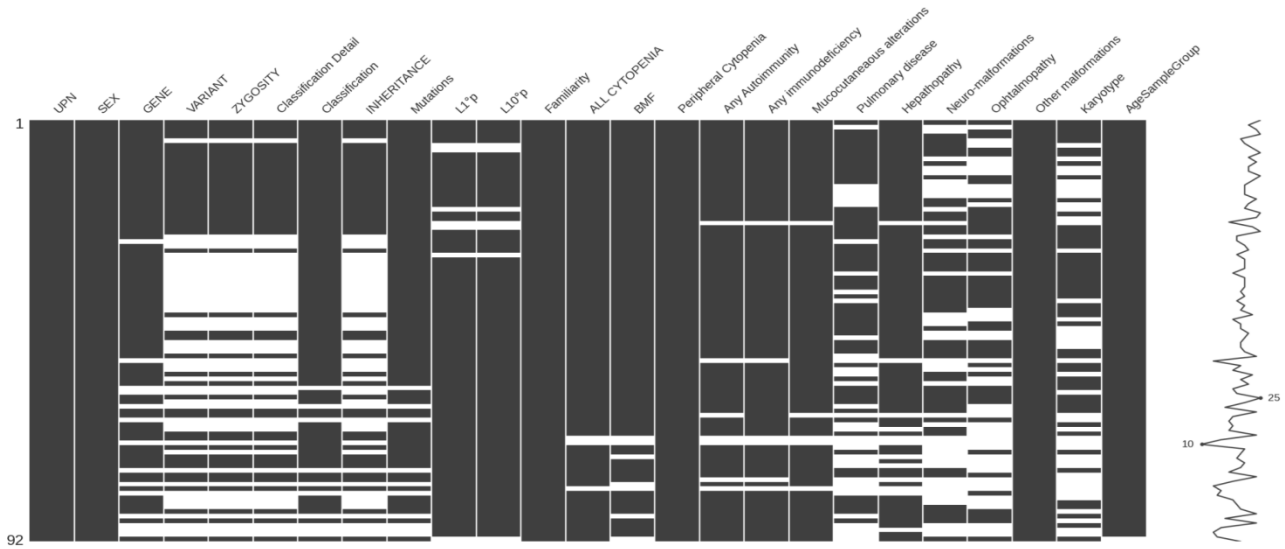

FIGURE 2 OBSERVED VALUES IN THE NEW DATASET.

Since the presence of missing values may affect the reliability of the analysis, features with more than 15% missing were excluded and the remaining missing values of each feature were imputed using the mode. The following categories were excluded: “bone marrow karyotype”, “neuro-malformations”, “ophthalmopathy”, “pulmonary disease”.

The figure below synthesizes the missing values.

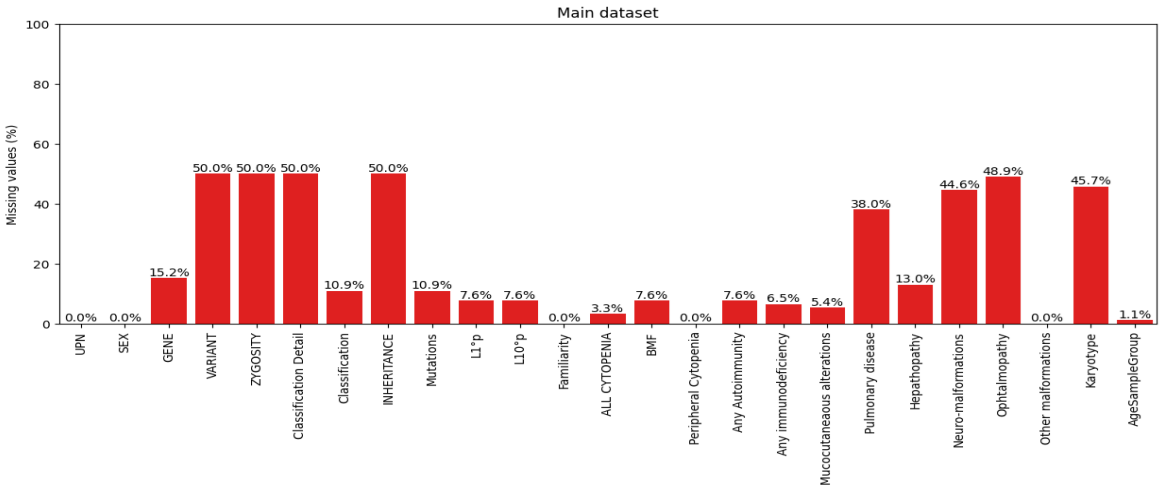

FIGURE 3 MISSING VALUES IN THE NEW DATASET

Multiple correspondence analysis was applied in order to obtain a cleaner representation of the dataset in the Euclidean space corresponding to 80% of the data variance.

Silhouette analysis was performed and the number of 4 cluster resulted as the optimal (ref. Figure 4).

Clustering analysis was performed by k-means. Analysis of the association between clusters and features was performed using the Freeman-Halton test.

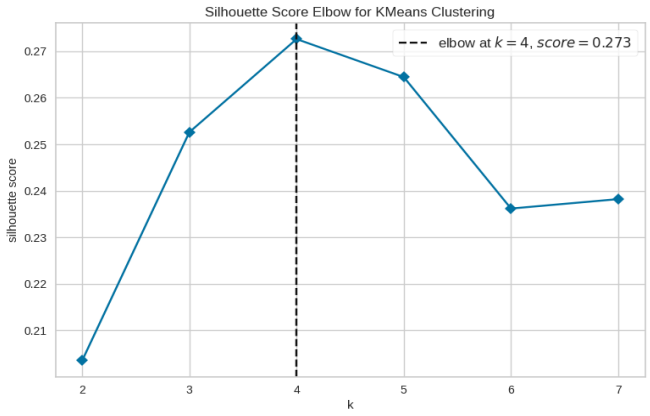

FIGURE 4 SILHOUETTE ANALYSIS

RESULTS

GENERAL CHARACTERISTICS

The analysis comprised 140 patients who met the inclusion criteria illustrated above and were referred at the Hematology Unit of the Giannina Gaslini Institute between September 2001 and April 2023.

One hundred and seventeen in one hundred and forty individuals had cytopenia according to the criteria defined in the Methods section. Eighteen in one hundred and forty individuals were relatives, including some potential donors of hematopoietic stem cells (2 of which were cytopenic). Seven in one hundred and forty individuals without cytopenia or familiarity were included in our cohort because patients with phenotypic or genetic anomalies suggestive of TBD.

In the original database, there were 85/140 males and 55/140 females. The median patient age was 13,28 years (ranging from 0 to 61).

In the new dataset there were 57/92 males and 35/92 females. The median patient age was the same as the original one (13,28 years, ranging from 0 to 61).

The missing values for every remaining category were classified as negative.

The table below summarizes the genetic characteristics of the cohort in the new dataset.

Patients with variant in TBD genes	46	"Pathogenetic" genes	20
		"VUS" genes	19
		"Negative genes"	7
Patients without variant in TBD genes	36		

TABLE 3 GENETIC FEATURES OF THE COHORT

The table below summarizes the main clinical-biochemical features of the cohort in the new dataset.

TELOMERE LENGTH	Patients	Yes	not
	TL < 1st percentile	15	77
	TL < 10th percentile (including TL < 1st percentile)	30	62
Cytopenia			
	Any type of cytopenia	80	12
	BMF	37	55
	Peripheral cytopenia	40	52
	Undefined cytopenia	3	89
Immunological features			
	Any autoimmunity	34	58
	Any immunodeficiency	57	35
Clinical features			
	Mucocutaneous disease	14	78
	Hepatopathy	14	78
	Other Malformations	25	67
Other features			
	Familiarity	34	58
	Age < 18 years	30	62

TABLE 4 GENERAL CHARACTERISTICS OF THE COHORT

CLUSTERS CHARACTERIZATION

The silhouette analysis permitted us to choose the optimal number of clusters the partition into four groups.

We analyzed the association between clusters and gene classification (TBD versus other genes), and we found that the distribution among the 4 clusters was all statistically significant (p-value ≈ 0.0002)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
TBD	13	14	11	8
Other Genes	15	1	3	17

TABLE 5 ASSOCIATION BETWEEN CLUSTERS AND CLASSIFICATION

Analyzing the association between clusters and gene pathogenicity, we found a significantly different distribution of pathogenic genes compared with "VUS" plus negative genes (p=0.0155) whereas this was not the case in the comparison between only pathogenic and negative genes (p = 0.0962).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
PATHOGENIC	2	10	6	2
NEGATIVE (+ VUS)	2 (11)	1 (4)	1 (5)	3 (6)

TABLE 6 ASSOCIATION BETWEEN CLUSTERS AND TBD GENE PATHOGENICITY

As reported in the figure below, clusters 1 and 4 showed a prevalence of “other genes” (i.e., non-TBD genes) whereas cluster 2 and 3 showed a strong prevalence of TBD genes.

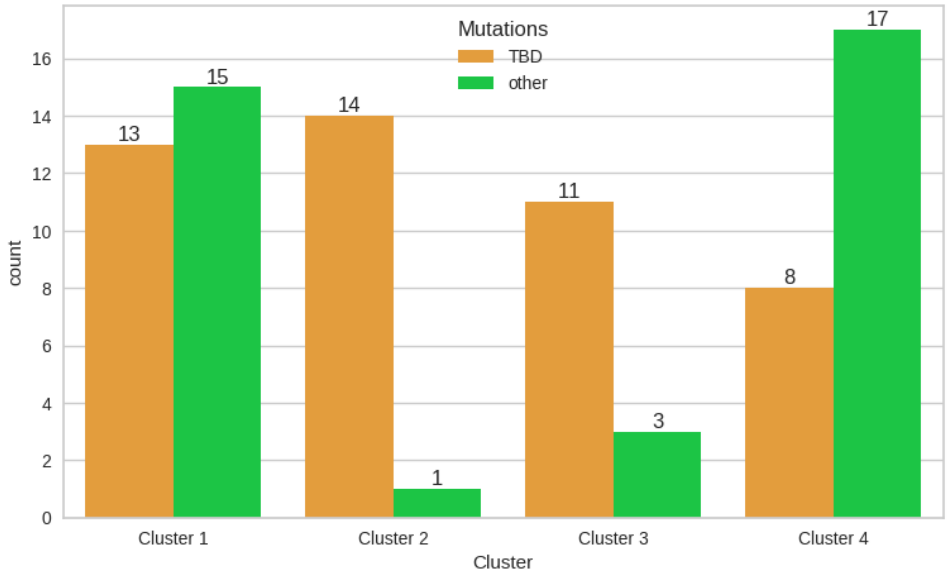

FIGURE 5 DISTRIBUTION OF TBD AND OTHER GENES

When we look at the distribution of TBD pathogenetic variants that seem to be located mainly in cluster 2 and 3.

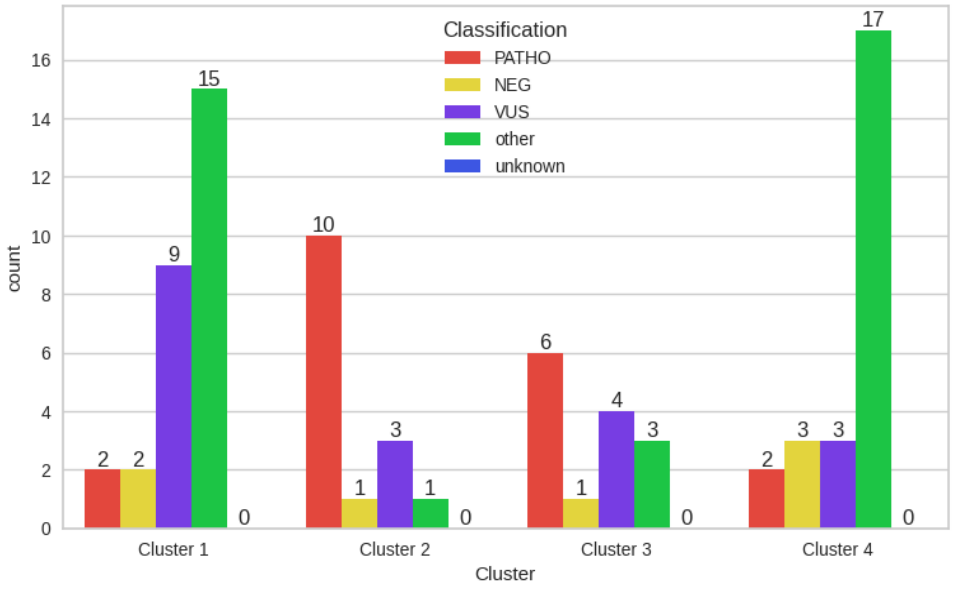

FIGURE 6 DISTRIBUTION OF PATHOGENIC TBD GENES

Therefore, based on the above shown distribution of genetic features, patients with TBD appear to belong to clusters 2 and 3.

Afterwards we conducted an "a posteriori" analyses of the distribution in the clusters of combinations of clinical and biochemical features.

We analyzed distribution of TL. Both TL <1st and < 10th percentile were significantly more distributed in clusters 2 and 3. The combination of TL (both <1st and 10th percentile) with pathogenic TBD variants was significantly more prevalent in clusters 2 and 3.

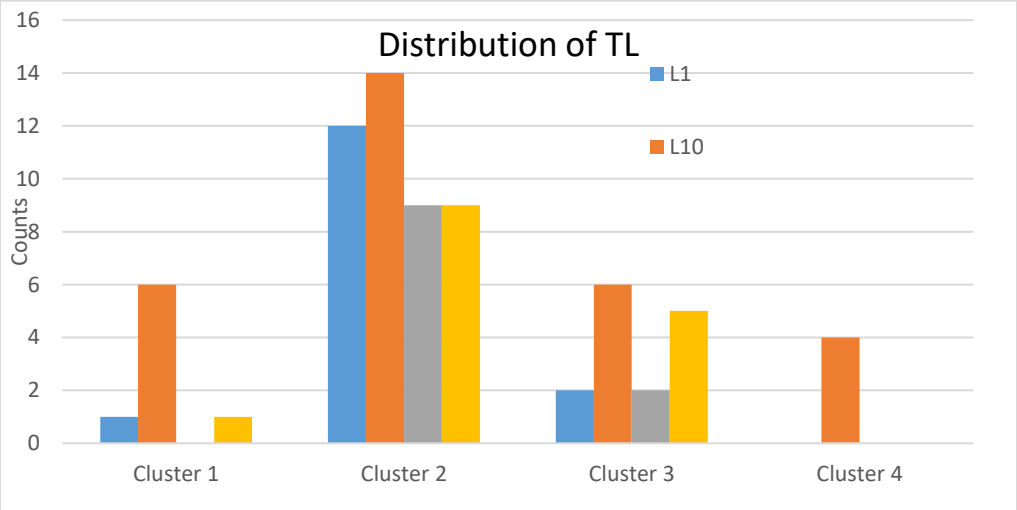

FIGURE 7 TL DISTRIBUTION

L1- Gene classification	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0 - other	14	1	3	17
0 - TBD	13	2	9	8
1- other	1	0	0	0
1 -TBD	0	12	2	0

TABLE 7 PIVOT TABLE FOR ASSOCIATION OF TL LENGTH < 1ST PERCENTILE AND GENE CLASSIFICATION

L10- Gene classification	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0 - other	11	0	3	13
0 - TBD	4	1	0	4
1- other	12	1	5	8
1 -TBD	1	13	6	0

TABLE 8 PIVOT TABLE FOR ASSOCIATION OF TL LENGTH < 1ST PERCENTILE AND GENE CLASSIFICATION

Starting from these assertions, we observed the distribution of other independent features such as those initially excluded because of excessive missing data. We removed from the analysis patients with missing values. Most of patients with neurological alterations and pulmonary disease were in cluster 2 whereas most patients without these two features were in Clusters 1 and 4, We did not find a significant distribution of patients with ophthalmopathy that is also considered a rather uncommon sign of TBD.

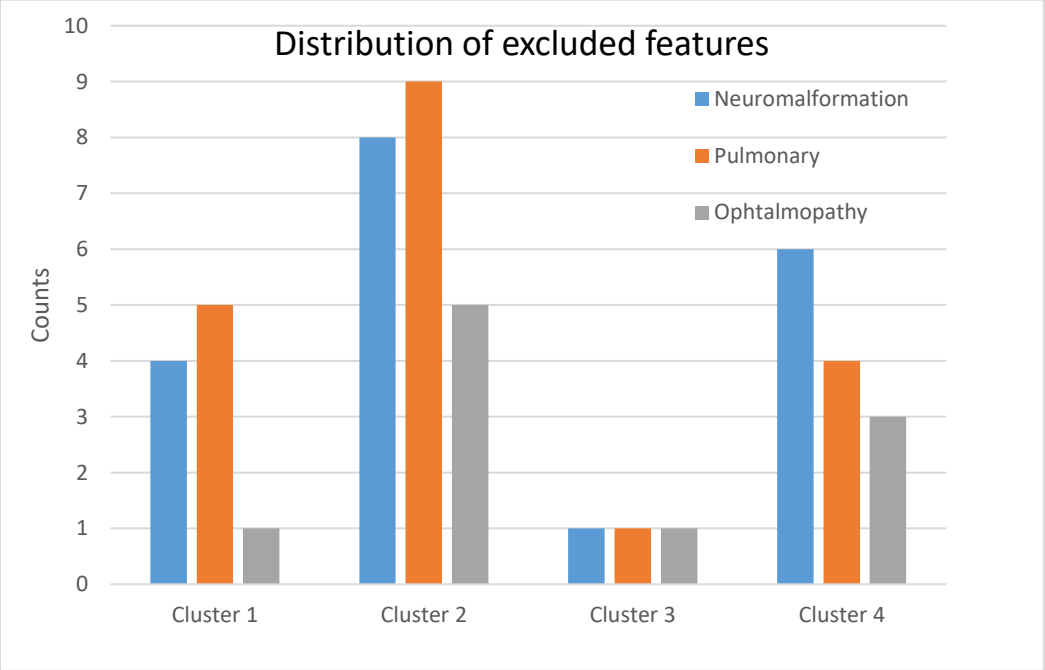

FIGURE 8 NEUROLOGICAL, LUNG AND EYE DISEASE DISTRIBUTION

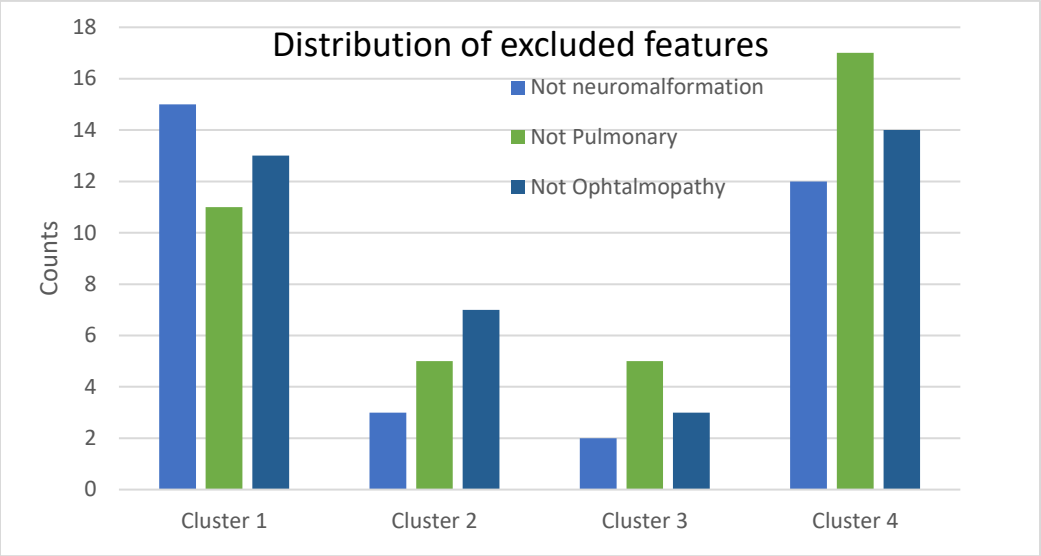

FIGURE 9 NEUROLOGICAL, LUNG AND EYE DISEASE DISTRIBUTION

Regarding the other clinical features, all the patients with the combination of bone marrow failure, hepatopathy and mucocutaneous triad were significantly associated to cluster 2. In addition, the distribution of the association between BMF and the mucocutaneous triad was significantly prevalent (8/12 patient) in Cluster 2. The association between BMF and hepatopathy was observed in all patients in Cluster 2.

The figure and the table below summarize the reported data.

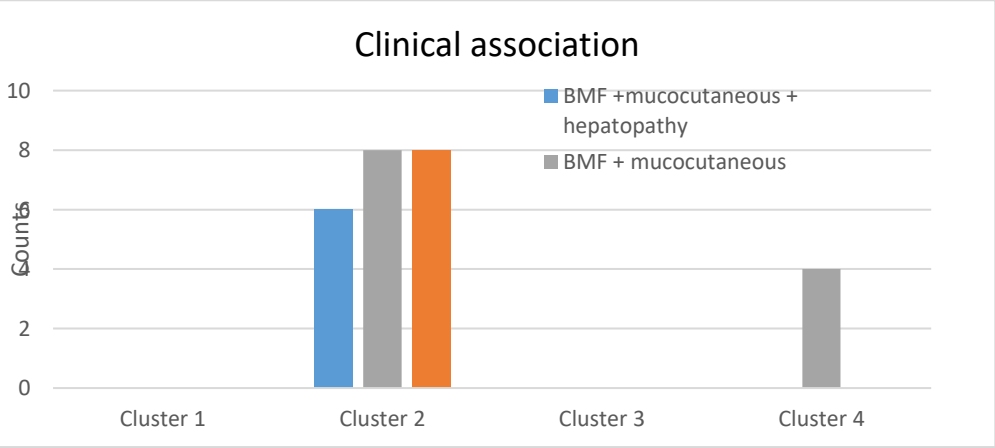

FIGURE 10 BMF, HEPATOPATHY AND MUCOCUTANEOUS TRIAD ASSOCIATION

BMF – MUC - HEP	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0 - 0 - 0	33	0	15	1
0 - 0 - 1	3	1	0	0
0 - 1 - 1	0	2	0	0
1 - 0 - 0	0	2	0	21
1 - 0 - 1	0	2	0	0
1 - 1 - 0	0	2	0	4
1 - 1 - 1	0	6	0	0

TABLE 9 PIVOT TABLE FOR ASSOCIATION OF BMF, MUCOCUTANEOUS TRIAD (MUC) AND HEPATOPATHY (HEP)

The association between the mucocutaneous triad and hepatopathy was associated to cluster 2. Moreover, patients with mucocutaneous alterations and peripheral cytopenia ± hepatopathy associated to cluster 2.

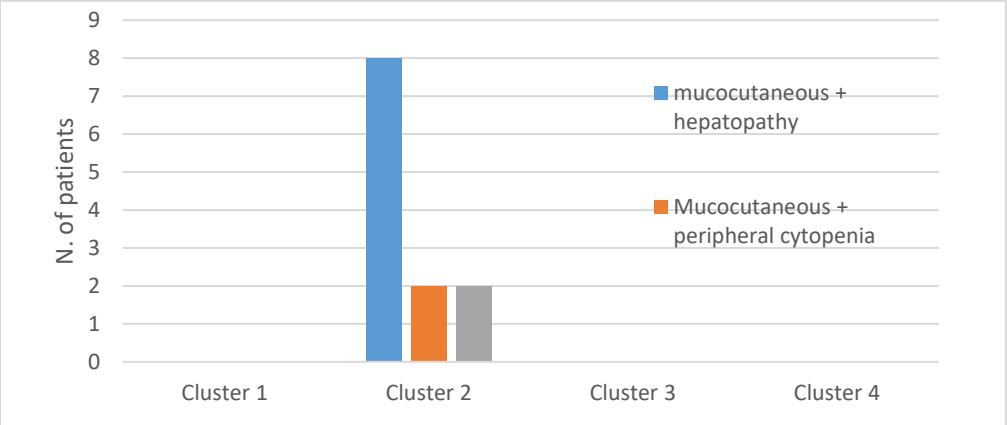

FIGURE 11 MUCOCUTANEOUS TRIAD, HEPATOPATHY AND PERIPHERAL CYTOPENIA ASSOCIATION

MUC - HEP - PER CYT	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0-0-0	0	2	14	22
0-0-1	33	0	1	0
0-1-0	0	2	0	0
0-1-1	3	1	0	0
1-0-0	0	2	0	4
1-1-0	0	6	0	0
1-1-1	0	2	0	0

TABLE 10 PIVOT TABLE FOR ASSOCIATION OF MUCOCUTANEOUS TRIAD (MUC), HEPATOPATHY (HEP) AND PERIPHERAL CYTOPENIA (PER CYT)

As for the immunological status of the patients, we found that patients with an association between BMF, hepatopathy, and immunodeficiency were significantly prevalent in cluster 2. Patients with hepatopathy and immunodeficiency without BMF were most located in Cluster 2. Mucocutaneous alteration associated with immunodeficiency, were significantly more distributed clusters 2 and 4.

Finally, we observed that the association of immunodeficiency with BMF was more distributed in cluster 4.

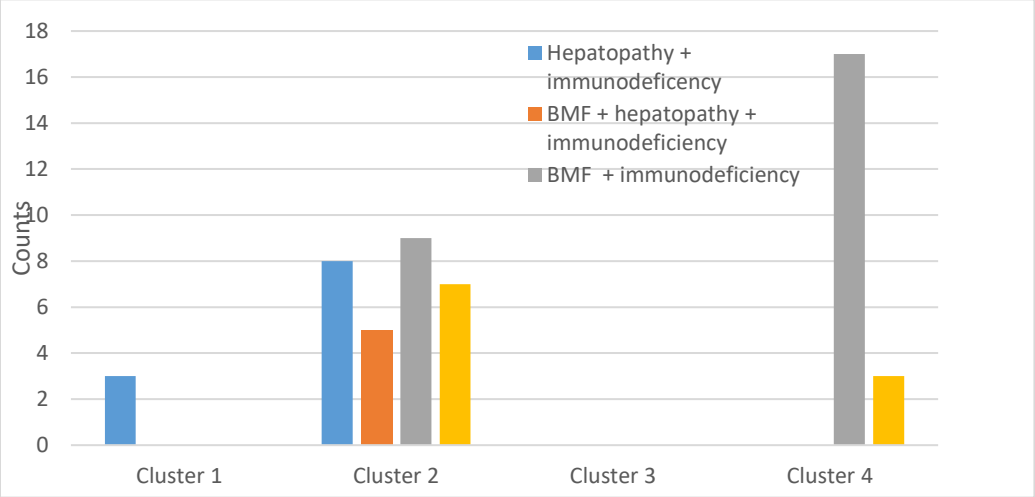

FIGURE 12 IMMUNODEFICIENCY AND CLINICAL FEATURES ASSOCIATIONS

BMF- MUC- HEP-IMM	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0-0-0-0	14	0	9	0
0-0-0-1	19	0	6	1
0-0-1-0	0	1	0	0
0-0-1-1	3	0	0	0
0-1-1-1	0	2	0	0
1-0-0-0	0	0	0	7
1-0-0-1	0	2	0	14
1-0-1-1	0	2	0	0
1-1-0-0	0	0	0	1
1-1-0-1	0	2	0	3

TABLE 11 PIVOT TABLE FOR ASSOCIATION OF BMF, MUCOCUTANEOUS TRIAD (MUC), HEPATOPATHY (HEP) AND IMMUNODEFICIENCY (IMM)

Regarding autoimmunity, we observed a statistically significant distribution in Cluster 2 of patients with mucocutaneous features and autoimmunity. The distribution of patients with autoimmunity and hepatopathy was significantly prevalent in clusters 1 and 2, with a mild prevalence of the latter.

Moreover, the association between autoimmunity and BMF was associated to cluster 4 (8/12) with some patients also in cluster 2 (4/12).

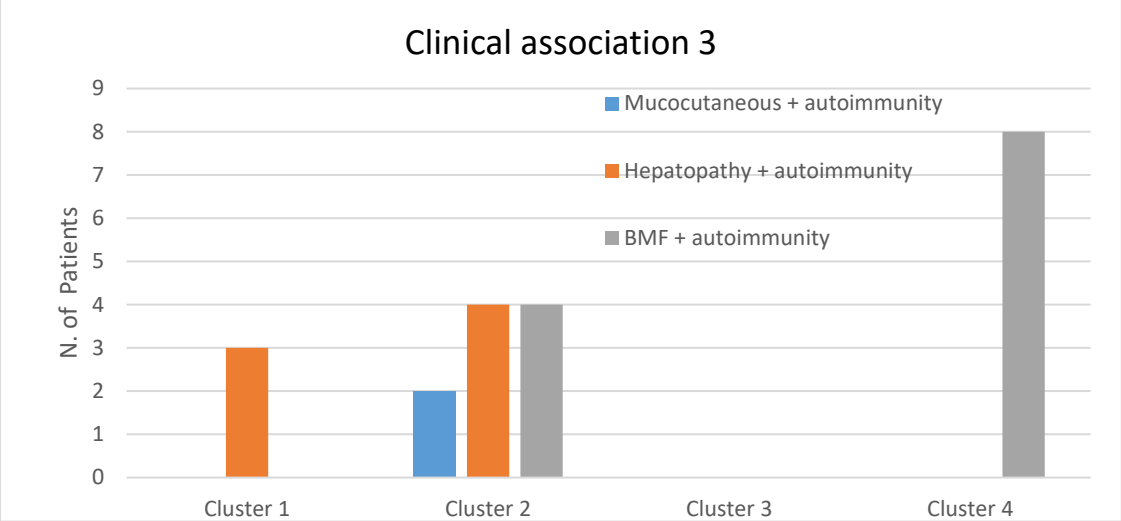

FIGURE 13 AUTOIMMUNITY AND CLINICAL FEATURES ASSOCIATION

BMF - MUC- HEP - AUTOIMM	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0-0-0-0	18	0	13	1
0-0-0-1	15	0	2	0
0-1-1-0	0	1	0	0
0-0-1-1	3	1	0	0
0-1-1-1	0	1	0	0
1-0-0-0	0	0	0	13
1-0-0-1	0	2	0	8
1-0-1-0	0	1	0	0
1-0-1-1	0	1	0	0
1-1-0-0	0	2	0	4
1-1-1-0	0	5	0	0
1-1-1-1	0	1	0	0

TABLE 12 PIVOT TABLE FOR ASSOCIATION OF BMF, MUCOCUTANEOUS TRIAD (MUC), HEPATOPATHY (HEP) AND AUTOIMMUNITY (AUTOIMM)

DISCUSSION

The broad spectrum of TBDs, the extremely variable and misleading clinical phenotype and the incomplete penetrance, makes diagnosis very challenging, especially in those who lack the classical features and whose genetics are inconclusive.

In this study, we developed an alternative model to the classical diagnostic work-up used for TBDs. Our aim was to create a helpful tool for those cryptic cases in which the correct diagnostic process is not clear.

We decided to use the unsupervised machine learning method because we considered that in a supervised algorithm, a small sample size and missing values could cause a significant bias.

In unsupervised machine learning, clustering is not based on an initial feature that works as a trainer of the clusters (supervised method), but all the features are analyzed simultaneously to create clusters. Therefore, in a small cohort with some missing values, we decided to perform unsupervised analysis.

In the initial analysis, we ignored the genetic features and excluded other features with more than 15% missing values.

This exclusion process allowed the discrimination of four statistically significant distinct clusters by comparing clusters with the genetic features of TBD mutations vs. other mutations and pathogenic mutations vs. not pathogenic mutations.

The analysis permitted the evidence of two clusters (2 and 3) of patients more suggestive of TBD than the others. In fact, most patients with genetic variants nonrelated to TBD-genes were located in Clusters 1 and 4.

Conversely, patients with TBD genes were distributed almost uniformly along the 4 clusters but the distribution of pathogenic TBD genes, was strongly associated to clusters 2 and 3.

After this step with the aim to understand which patients' characteristics were included in the clusters, we analyzed the initially excluded clinical findings usually associated with TBD. These

features must be considered independent of clustering. We observed important results that provided additional confirmation of the quality of cluster division. The distribution of ophthalmopathy a rather uncommon TBD feature, was not significant; however, we observed that both lung and neurological diseases, typical of TBD, were strongly significantly located in cluster 2 that is one of the "TBD Cluster".

Finally, we performed a posteriori analysis of the clustering features. Still withstanding the statistical limitations of this type of analysis, however we obtained interesting clinical data.

Regarding TL, we observed that most patients with TL below 1st percentile and those with TL below 10th percentile were significantly located in clusters 2 and 3. The combination of TL with pathogenic TBD genes confirmed the distribution of patients with both features in clusters 2 and 3 thus supporting the idea that clusters 2 and 3 were the real "TBD cluster" in our cohort.

Moreover, not only TL below 1st percentile had this relevant distribution but this was true also for TL below 10th. This confirms the last recommendation in the literature. In fact, in the early years, for the TBD diagnosis TL below 1st percentile was necessary. However, recent reports have shown that a TL below 10th percentile is sufficient to start the diagnostic work-up [57], [68].

Next, we analyzed which clinical findings were more associated with cluster 2 and cluster 3. We observed, as expected, that all our patients with BMF, mucocutaneous findings, and hepatopathy were located in one of the "TBD clusters" (i.e., cluster 2). In addition, the associations between BMF and either hepatopathy or the mucocutaneous triad showed a relevant location in cluster 2. Without BMF, we observed that mucocutaneous findings associated with hepatopathy, interestingly, tended to be located in the TBD clusters.

Moving forward with the analysis, we characterized one important and emerging aspect of TBDs, i.e., immune dysregulation. We observed that immunodeficiency associated only with mucocutaneous findings was mostly included in cluster 2. The same held true for the association of immunodeficient patients with BMF and hepatopathy. Regarding autoimmunity, we observed that the association of the mucocutaneous triad with autoimmunity determined the location of

patients in cluster 2. Patients with features of hepatopathy and autoimmunity also tended to be located in cluster 2.

These findings are relevant for diagnosis in two opposite directions. In fact, the link of TBD with immune dysregulation provides the important issue that is always necessary to investigate the immunological status of known TBD patients. At the same time, these results suggest that individuals with mucocutaneous typical findings associated with immunodeficiency or autoimmunity may have cryptic TBD expression. Finally, it is important to highlight that in cases of misdiagnosed hepatopathy associated with immune dysregulation, TBD may be considered.

The main limitation of our study is presence of missing values in some categories that is somehow inherent in retrospective studies.

The strength of our work is that we were able, to characterize a heterogeneous group of patients by creating well-defined clusters with typical clinical and hematological findings. Another important strength is the novelty of the ML approach that, to the best of our knowledge, has never been applied to rare diseases like TBD, and that can open new ways for analysis for these and other rare disorders.

CONCLUSION

This study showed that machine learning can contribute to the diagnostic work-up of rare, clinically variable and confounding disease such as TBDs. We also demonstrated that in a heterogeneous cohort of patients, it is possible to define specific clusters of patients in whom the diagnosis of TBD is more probable.

Moreover, it suggests that patients with only a few and isolated aspects of the classical disease, such as immune dysregulation, hepatopathy, or mucocutaneous features, must be investigated for TBD. Finally, our algorithm confirms that a patient with a TL below 10^{th} percentile must be considered a TBD patient until otherwise proven.

REFERENCES

[1] D. Chakravarti, K. A. LaBella, and R. A. DePinho, “Telomeres: history, health, and hallmarks of aging,” *Cell*, vol. 184, no. 2, pp. 306–322, Jan. 2021, doi: 10.1016/j.cell.2020.12.028.

[2] L. Hayflick, “The limited in vitro lifetime of human diploid cell strains,” *Exp Cell Res*, vol. 37, no. 3, pp. 614–636, Mar. 1965, doi: 10.1016/0014-4827(65)90211-9.

[3] E. H. Blackburn and J. G. Gall, “A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena,” *J Mol Biol*, vol. 120, no. 1, pp. 33–53, Mar. 1978, doi: 10.1016/0022-2836(78)90294-2.

[4] C. W. Greider and E. H. Blackburn, “Identification of a specific telomere terminal transferase activity in tetrahymena extracts,” *Cell*, vol. 43, no. 2, pp. 405–413, Dec. 1985, doi: 10.1016/0092-8674(85)90170-9.

[5] C. W. Greider and E. H. Blackburn, “A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis,” *Nature*, vol. 337, no. 6205, pp. 331–337, Jan. 1989, doi: 10.1038/337331a0.

[6] T. de Lange, “Shelterin: the protein complex that shapes and safeguards human telomeres,” *Genes Dev*, vol. 19, no. 18, pp. 2100–2110, Sep. 2005, doi: 10.1101/gad.1346005.

[7] K. Roka, E. E. Solomou, and A. Kattamis, “Telomere biology: from disorders to hematological diseases,” *Front Oncol*, vol. 13, May 2023, doi: 10.3389/fonc.2023.1167848.

[8] H. Banu Lateef, P. Mitta Suresh, P. Bharathi, S. Pathak, and A. Banerjee, “A Brief Overview of Telomeres and Telomerase in Aging and Cancer,” *Curr Appl Sci Technol*, vol. 23, no. 4, Dec. 2022, doi: 10.55003/cast.2023.04.23.015.

[9] S. A. Savage and A. A. Bertuch, “The genetics and clinical manifestations of telomere biology disorders.,” *Genet Med*, vol. 12, no. 12, pp. 753–64, Dec. 2010, doi: 10.1097/GIM.0b013e3181f415b5.

[10] A. Garus and C. Autexier, “Dyskerin: an essential pseudouridine synthase with multifaceted roles in ribosome biogenesis, splicing, and telomere maintenance,” *RNA*, vol. 27, no. 12, pp. 1441–1458, Dec. 2021, doi: 10.1261/rna.078953.121.

[11] M. R. Niewisch and S. A. Savage, “An update on the biology and management of dyskeratosis congenita and related telomere biology disorders,” *Expert Rev Hematol*, vol. 12, no. 12, pp. 1037–1052, Dec. 2019, doi: 10.1080/17474086.2019.1662720.

[12] R. Diotti and D. Loayza, “Shelterin complex and associated factors at human telomeres,” *Nucleus*, vol. 2, no. 2, pp. 119–135, Mar. 2011, doi: 10.4161/nucl.2.2.15135.

[13] M. Armanios and E. H. Blackburn, “The telomere syndromes,” *Nat Rev Genet*, vol. 13, no. 10, pp. 693–704, Oct. 2012, doi: 10.1038/nrg3246.

[14] C. Speckmann *et al.*, “Clinical and Molecular Heterogeneity of RTEL1 Deficiency,” *Front Immunol*, vol. 8, May 2017, doi: 10.3389/fimmu.2017.00449.

[15] N. S. Heiss *et al.*, “X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions,” *Nat Genet*, vol. 19, no. 1, pp. 32–38, May 1998, doi: 10.1038/ng0598-32.

[16] J. R. Mitchell, E. Wood, and K. Collins, “A telomerase component is defective in the human disease dyskeratosis congenita,” *Nature*, vol. 402, no. 6761, pp. 551–555, Dec. 1999, doi: 10.1038/990141.

[17] Savage SA and Niewisch MR, “Dyskeratosis Congenita and Related Telomere Biology Disorders. 2009 Nov 12 [updated 2023 Jan 19],” *GeneReviews*, Jan. 2023.

[18] Suneet Agarwal, Sharon A. Savage, and Katie Barrett Stevens, *Telomere Biology Disorders Diagnosis and Management Guidelines, 2nd Edition,* 2nd ed. 2022.

[19] G. Glousker, F. Touzot, P. Revy, Y. Tzfati, and S. A. Savage, “Unraveling the pathogenesis of Hoyeraal–Hreidarsson syndrome, a complex telomere biology disorder,” *Br J Haematol*, vol. 170, no. 4, pp. 457–471, Aug. 2015, doi: 10.1111/bjh.13442.

[20] P. Gu and S. Chang, “Functional characterization of human CTC1 mutations reveals novel mechanisms responsible for the pathogenesis of the telomere disease Coats plus,” *Aging Cell*, vol. 12, no. 6, pp. 1100–1109, Dec. 2013, doi: 10.1111/acel.12139.

[21] M. R. Niewisch and S. A. Savage, “An update on the biology and management of dyskeratosis congenita and related telomere biology disorders,” *Expert Rev Hematol*, vol. 12, no. 12, pp. 1037–1052, Dec. 2019, doi: 10.1080/17474086.2019.1662720.

[22] P. F. Fogarty *et al.*, “Late presentation of dyskeratosis congenita as apparently acquired aplastic anaemia due to mutations in telomerase RNA,” *The Lancet*, vol. 362, no. 9396, pp. 1628–1630, Nov. 2003, doi: 10.1016/S0140-6736(03)14797-6.

[23] L. Wang and H. Liu, “Pathogenesis of aplastic anemia,” *Hematology*, vol. 24, no. 1, pp. 559–566, Jan. 2019, doi: 10.1080/16078454.2019.1642548.

[24] K. E. Schratz *et al.*, “Cancer spectrum and outcomes in the Mendelian short telomere syndromes,” *Blood*, vol. 135, no. 22, pp. 1946–1956, May 2020, doi: 10.1182/blood.2019003264.

[25] B. P. Alter, N. Giri, S. A. Savage, and P. S. Rosenberg, “Cancer in the National Cancer Institute inherited bone marrow failure syndrome cohort after fifteen years of follow-up,” *Haematologica*, vol. 103, no. 1, pp. 30–39, Jan. 2018, doi: 10.3324/haematol.2017.178111.

[26] M. Knudson, S. Kulkarni, Z. K. Ballas, M. Bessler, and F. Goldman, “Association of immune abnormalities with telomere shortening in autosomal-dominant dyskeratosis congenita,” *Blood*, vol. 105, no. 2, pp. 682–688, Jan. 2005, doi: 10.1182/blood-2004-04-1673.

[27] A. Mulet and J. Signes-Costa, “Idiopathic Pulmonary Fibrosis and Telomeres,” *J Clin Med*, vol. 11, no. 23, p. 6893, Nov. 2022, doi: 10.3390/jcm11236893.

[28] C. Tirelli, C. Pesenti, M. Miozzo, M. Mondoni, L. Fontana, and S. Centanni, “The Genetic and Epigenetic Footprint in Idiopathic Pulmonary Fibrosis and Familial Pulmonary Fibrosis: A State-of-the-Art Review,” *Diagnostics*, vol. 12, no. 12, p. 3107, Dec. 2022, doi: 10.3390/diagnostics12123107.

[29] J. K. Alder and M. Armanios, “Telomere-mediated lung disease,” *Physiol Rev*, vol. 102, no. 4, pp. 1703–1720, Oct. 2022, doi: 10.1152/physrev.00046.2021.

[30] D. Hartmann *et al.*, “Telomerase gene mutations are associated with cirrhosis formation,” *Hepatology*, vol. 53, no. 5, pp. 1608–1617, May 2011, doi: 10.1002/hep.24217.

[31] R. T. Calado *et al.*, “A Spectrum of Severe Familial Liver Disorders Associate with Telomerase Mutations,” *PLoS One*, vol. 4, no. 11, p. e7926, Nov. 2009, doi: 10.1371/journal.pone.0007926.

[32] S. A. Savage, “Beginning at the ends: telomeres and human disease,” *F1000Res*, vol. 7, p. 524, May 2018, doi: 10.12688/f1000research.14068.1.

[33] R. W. Himes *et al.*, “Gastrointestinal Hemorrhage: A Manifestation of the Telomere Biology Disorders,” *J Pediatr*, vol. 230, pp. 55-61.e4, Mar. 2021, doi: 10.1016/j.jpeds.2020.09.038.

[34] P. P. Khincha *et al.*, “Pulmonary arteriovenous malformations: an uncharacterised phenotype of dyskeratosis congenita and related telomere biology disorders,” *European Respiratory Journal*, vol. 49, no. 1, p. 1601640, Jan. 2017, doi: 10.1183/13993003.01640-2016.

[35] C. Higgs *et al.*, “Understanding the evolving phenotype of vascular complications in telomere biology disorders,” *Angiogenesis*, vol. 22, no. 1, pp. 95–102, Feb. 2019, doi: 10.1007/s10456-018-9640-7.

[36] D.-H. Woo *et al.*, “Enhancing a Wnt-Telomere Feedback Loop Restores Intestinal Stem Cell Function in a Human Organotypic Model of Dyskeratosis Congenita,” *Cell Stem Cell*, vol. 19, no. 3, pp. 397–405, Sep. 2016, doi: 10.1016/j.stem.2016.05.024.[37] T.-L. B. Yang *et al.*, “Mutual reinforcement between telomere capping and canonical Wnt signalling in the intestinal stem cell niche,” *Nat Commun*, vol. 8, no. 1, p. 14766, Mar. 2017, doi: 10.1038/ncomms14766.

[38] S. Bhala *et al.*, “CNS manifestations in patients with telomere biology disorders,” *Neurol Genet*, vol. 5, no. 6, p. 370, Dec. 2019, doi: 10.1212/NXG.0000000000000370.

[39] M. Callea *et al.*, “Multisystemic Manifestations in Rare Diseases: The Experience of Dyskeratosis Congenita,” *Genes (Basel)*, vol. 13, no. 3, p. 496, Mar. 2022, doi: 10.3390/genes13030496.

[40] I. Mender and J. Shay, “Telomere Restriction Fragment (TRF) Analysis,” *Bio Protoc*, vol. 5, no. 22, 2015, doi: 10.21769/BioProtoc.1658.

[41] M. Kimura *et al.*, “Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths,” *Nat Protoc*, vol. 5, no. 9, pp. 1596–1607, Sep. 2010, doi: 10.1038/nprot.2010.124.

[42] T.-P. Lai, W. E. Wright, and J. W. Shay, “Comparison of telomere length measurement methods,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 373, no. 1741, p. 20160451, Mar. 2018, doi: 10.1098/rstb.2016.0451.

[43] J. Lin, D. L. Smith, K. Esteves, and S. Drury, “Telomere length measurement by qPCR – Summary of critical factors and recommendations for assay design,” *Psychoneuroendocrinology*, vol. 99, pp. 271–278, Jan. 2019, doi: 10.1016/j.psyneuen.2018.10.005.

[44] K. Norris *et al.*, “High-throughput STELA provides a rapid test for the diagnosis of telomere biology disorders,” *Hum Genet*, vol. 140, no. 6, pp. 945–955, Jun. 2021, doi: 10.1007/s00439-021-02257-4.

[45] G. M. Baerlocher and P. M. Lansdorp, “Telomere length measurements in leukocyte subsets by automated multicolor flow-FISH,” *Cytometry*, vol. 55A, no. 1, pp. 1–6, Sep. 2003, doi: 10.1002/cyto.a.10064.

[46] T. Wand *et al.*, “Telomere content measurement in human hematopoietic cells: Comparative analysis of qPCR and Flow-FISH techniques,” *Cytometry Part A*, vol. 89, no. 10, pp. 914–921, Oct. 2016, doi: 10.1002/cyto.a.22982.

[47] I. Schmid and B. D. Jamieson, “Assessment of Telomere Length, Phenotype, and DNA Content,” *Curr Protoc Cytom*, vol. 29, no. 1, Jul. 2004, doi: 10.1002/0471142956.cy0726s29.

[48] T. Kelesidis and I. Schmid, “Assessment of Telomere Length, Phenotype, and DNA Content,” *Curr Protoc Cytom*, vol. 79, no. 1, Jan. 2017, doi: 10.1002/cpcy.12.

[49] H. Yamaguchi *et al.*, “Mutations in *TERT,* the Gene for Telomerase Reverse Transcriptase, in Aplastic Anemia,” *New England Journal of Medicine*, vol. 352, no. 14, pp. 1413–1424, Apr. 2005, doi: 10.1056/NEJMoa042980.

[50] A. M. Aalbers, S. Kajigaya, M. M. van den Heuvel-Eibrink, V. H. J. van der Velden, R. T. Calado, and N. S. Young, “Human telomere disease due to disruption of the CCAAT box of the TERC promoter,” *Blood*, vol. 119, no. 13, pp. 3060–3063, Mar. 2012, doi: 10.1182/blood-2011-10-383182.

[51] P. Revy, C. Kannengiesser, and A. A. Bertuch, “Genetics of human telomere biology disorders,” *Nat Rev Genet*, vol. 24, no. 2, pp. 86–108, Feb. 2023, doi: 10.1038/s41576-022-00527-z.

[52] L. Maryoung *et al.*, “Somatic mutations in telomerase promoter counterbalance germline loss-of-function mutations,” *Journal of Clinical Investigation*, vol. 127, no. 3, pp. 982–986, Feb. 2017, doi: 10.1172/JCI91161.

[53] F. Gutierrez-Rodrigues *et al.*, “Pathogenic TERT promoter variants in telomere diseases,” *Genetics in Medicine*, vol. 21, no. 7, pp. 1594–1602, Jul. 2019, doi: 10.1038/s41436-018-0385-x.

[54] M. C. J. Jongmans *et al.*, “Revertant Somatic Mosaicism by Mitotic Recombination in Dyskeratosis Congenita,” *The American Journal of Human Genetics*, vol. 90, no. 3, pp. 426–433, Mar. 2012, doi: 10.1016/j.ajhg.2012.01.004.

[55] J. Xu, P. P. Khincha, N. Giri, B. P. Alter, S. A. Savage, and J. M. Y. Wong, “Investigation of chromosome X inactivation and clinical phenotypes in female carriers of DKC1 mutations,” *Am J Hematol*, vol. 91, no. 12, pp. 1215–1220, Dec. 2016, doi: 10.1002/ajh.24545.

[56] M. R. Niewisch *et al.*, “Disease progression and clinical outcomes in telomere biology disorders,” *Blood*, vol. 139, no. 12, pp. 1807–1819, Mar. 2022, doi: 10.1182/blood.2021013523.

[57] B. P. Alter, N. Giri, S. A. Savage, and P. S. Rosenberg, “Telomere length in inherited bone marrow failure syndromes,” *Haematologica*, vol. 100, no. 1, pp. 49–54, Jan. 2015, doi: 10.3324/haematol.2014.114389.

[58] S. Richards *et al.*, “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,” *Genetics in Medicine*, vol. 17, no. 5, pp. 405–424, May 2015, doi: 10.1038/gim.2015.30.

[59] A. J. Walne *et al.*, “Marked overlap of four genetic syndromes with dyskeratosis congenita confounds clinical diagnosis,” *Haematologica*, vol. 101, no. 10, pp. 1180–1189, Oct. 2016, doi: 10.3324/haematol.2016.147769.

[60] F. Gutierrez-Rodrigues *et al.*, “Differential diagnosis of bone marrow failure syndromes guided by machine learning,” *Blood*, Dec. 2022, doi: 10.1182/blood.2022017518.

[61] P. Valent *et al.*, “Definitions and standards in the diagnosis and treatment of the myelodysplastic syndromes: Consensus statements and report from a working conference,” *Leuk Res*, vol. 31, no. 6, pp. 727–736, Jun. 2007, doi: 10.1016/j.leukres.2006.11.009.

[62] G. Matthijs *et al.*, “Guidelines for diagnostic next-generation sequencing,” *European Journal of Human Genetics*, vol. 24, no. 1, pp. 2–5, Jan. 2016, doi: 10.1038/ejhg.2015.226.

[63] E. C. Guinan, “Diagnosis and Management of Aplastic Anemia,” *Hematology*, vol. 2011, no. 1, pp. 76–81, Dec. 2011, doi: 10.1182/asheducation-2011.1.76.

[64] A. Bacigalupo *et al.*, "Bone marrow transplantation (BMT) versus immunosuppression for the treatment of severe aplastic anaemia (SAA): a report of the EBMT SAA Working Party," *Br J Haematol*, vol. 70, no. 2, pp. 177–182, Oct. 1988, doi: 10.1111/j.1365-2141.1988.tb02460.x.

[65] B. Camitta *et al.*, "Severe aplastic anemia: a prospective study of the effect of early marrow transplantation on acute mortality," *Blood*, vol. 48, no. 1, pp. 63–70, Jul. 1976, doi: 10.1182/blood.V48.1.63.63.

[66] N. Rufer *et al.*, "Telomere Fluorescence Measurements in Granulocytes and T Lymphocyte Subsets Point to a High Turnover of Hematopoietic Stem Cells and Memory T Cells in Early Childhood," *J Exp Med*, vol. 190, no. 2, pp. 157–168, Jul. 1999, doi: 10.1084/jem.190.2.157.

[67] T. H. Brümmendorf, J. P. Maciejewski, J. Mak, N. S. Young, and P. M. Lansdorp, "Telomere length in leukocyte subpopulations of patients with aplastic anemia," *Blood*, vol. 97, no. 4, pp. 895–900, Feb. 2001, doi: 10.1182/blood.V97.4.895.

[68] M. Tometten *et al.*, "Identification of Adult Patients With Classical Dyskeratosis Congenita or Cryptic Telomere Biology Disorder by Telomere Length Screening Using Age-modified Criteria," *Hemasphere*, vol. 7, no. 5, p. e874, May 2023, doi: 10.1097/HS9.0000000000000874.