ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE E TECNOLOGIE DELLA SALUTE

Ciclo 37

Settore Concorsuale: 02/D1 - FISICA APPLICATA, DIDATTICA E STORIA DELLA FISICA

Settore Scientifico Disciplinare: FIS/07 - FISICA APPLICATA A BENI CULTURALI, AMBIENTALI, BIOLOGIA E MEDICINA

ARTIFICIAL INTELLIGENCE APPROACHES FOR PRIVACY-PROTECTED POOLING OF GENOMICS, CLINICAL AND OTHER '-OMICS' DATA ANALYSIS FOR HAEMATOLOGICAL DISEASE PROGNOSIS, PREVENTION, DIAGNOSTICS AND TREATMENT

Presentata da: Viktor Savevski

Coordinatore Dottorato	**Supervisore**
Igor Diemberger	Gastone Castellani

Esame finale anno 2025

Acknowledgement

It is humbling to reflect on the results collected in this thesis. For, when I first embarked on this journey, I was filled with optimism. The benefits of Artificial Intelligence (AI) in the healthcare sector had become apparent. In consequence, organizations were looking to foster increased healthcare data mobility. At the same time, data privacy had become a central topic in public discourse. In leveraging federated learning, I would reconcile AI innovation in healthcare with confidentiality and security.

Now, three years later, the results are truly beyond what I could ever have expected. Not only have the results given rise to a privacy-protecting AI platform for haematological diseases; they have also sparked interest with top European Healthcare organizations. It has been truly heartening seeing my work come to fruition. This would not have been possible without the support, trust, creativity, and brilliance of those around me. I would therefore like to take this opportunity to thank all those involved.

First and foremost, I would like to thank my supervisors. As I was often conceiving, developing, and implementing multiple ideas at the same time, it must have been hard to guide my progress. Nonetheless, you expressed continuous support and belief in my work.

I would also like to thank all those who participated in the *Genomics4All* consortium, of which this research is a part. It is rare for a team of this scale to be so driven, passionate, and brilliant in working towards a common goal. Our discussions were always inspiring and it is through our shared belief in the cause that we achieved our results.

Last, but not least, I would like to thank my family and friends for supporting me in this phase of my life. Even in times where my mission seemed impossible, you supported me and kept me going.

Contents

List of acronyms and abbreviations

Abbreviation	**Meaning**
AI	Artificial Intelligence
AIC	Akaike information criterion
AML	Acute Myeloid Leukemia
BP	Base pair
CI	Confidence Interval
CSF	Clinical Synthetic Fidelity
DDX41	DEAD-box RNA helicase-1 gene
del(x)	Deletion of chromosome x
DNA	Deoxyribonucleic Acid
EC	European Commission
EHA	European Hematology Association
EHR	Electronic Health Record
ERN	European Reference Network
EU	European Union
EuroMDS	European Myelodysplastic Syndromes Registry
GAN	Generative Adversarial Network
GESMD	Spanish Group of Myelodysplastic Syndromes Registry
GDPR	General Data Protection Regulation
GSF	Genomic Synthetic Fidelity
HD	Hematological disease
HIPAA	Health Insurance Portability and Accountability Act
HMA	Hypomethylating Agents
HSCT	Hematopoietic Stem Cell Transplantation
IMS	Identical Match Share
IPSS-M	Molecular International Prognostic Scoring System
IPSS-R	Revised-International Prognostic Scoring System
IQR	Inter Quartile Range
IWG-PM	International Working Group for the Prognosis of MDS
LFS	Leukemia-Free Survival

List of acronyms and abbreviations (continued)

Abbreviation	**Meaning**
MDS	Myelodysplastic Syndromes
MDS-EBk	MDS with excess of blasts, type k
MDS-MLD	MDS with multilineage dysplasia
MDS-RS-MLD	MDS with ring sideroblasts and multilineage dysplasia
MDS-RS-SLD	MDS with ring sideroblasts and single lineage dysplasia
MDS-SLD	MDS with single lineage dysplasia
MDS-U	Unclassified MDS
ML	Machine Learning
MT-GAN	Multilabel Time-series GAN
NNDR	Nearest Neighbor Distance Ratio
NOS	Not Otherwise Specified
NRM	Nonrelapse Mortality
Omics	A broad range of health data categories
OS	Overall Survival
PET	Privacy Enhancing Technology
RBAC	Role-Based Access Control
RHD	Rare Hematological Disease
RNA	Ribonucleic Acid
scRNA	Small Conditional RNA
SD	Standard Deviation
seq	Sequencing
SHAP	Shapley Additive Explanations
SLD	Single-Lineage Dysplasia
SRSF	Serine/arginine-Rich Splicing Factor
SVF	Synthetic Validation Framework
TLS	Transport Layer Encryption
TI	Transfusion Independence
WGS	Whole Genome Sequencing

Chapter 1

Introduction

1.1 Motivation

There are up to 450 Hematological Diseases (HDs), generally classified in six large groups of oncological and non-oncological diseases. HDs result from abnormalities of blood cells; lymphoid organs; and coagulation factors, and affect a substantial number of patients. For example, HDs account for about 5% of Cancers [1]. Most HDs can cause chronic health problems and many of them are life-threatening conditions requiring numerous resources for correct diagnosis, management and treatment. Recently, the European Hematology Association (EHA) evaluated the financial burden of blood disorders on European society at €22.5 billion per year [1].

Personalized or precision medicine is a medical model in which conventional medicine is combined with advanced genetic profiling, leveraging Artificial Intelligence (AI) and Machine Learning (ML). This results in tailored diagnostic, prognostic and therapeutic strategies. Personalized medicine can revolutionize hematology, improving patients' quality of life and reducing the overall financial burden of HDs. Unfortunately, this is currently underexplored: existing AI models for HDs lack patient-centricity and personalization. Furthermore, the vast amounts of relevant data produced are often inaccessible and diffused.

AI and advanced genomics on other ”-omics” research has enabled the study of personalized biomarkers in many fields of medicine, including cardiology [2]; endocrinology [3]; and oncology [4, 5]. In hematology, the analysis of genomics information recently garnered success in establishing genetic bases for myelodysplastic syndromes [6, 7]; myeloid and lymphoid neoplasms [8]; and hematopoiesis [9].

1.2 Research Objectives

1.2.1 Research Objective 1

Map the hematology data landscape, including the AI and omics opportunities; the data repositories in the European Union; and the relevant laws, regulations, and ethical guidelines.

Hematological data is diffused over multiple institutions. This makes it difficult for researchers to gain access to required information in sufficient volumes. This disproportionally affects rare diseases, which is scarce by its very nature. Our first research objective is to provide a clear overview of hematology data in the European Union and related nations, highlighting which data are collected; what institutions house them; and what data mobility protocols these institutions implement. This includes a detailed account of the data used in the remainder of the thesis.

1.2.2 Research Objective 2

Leverage increased data availability to validate existing prognostic models for hematological diseases at scale

A plethora of prognostic models has been proposed, often limited by the aforementioned data availability problems. Leveraging the substantial database amassed in addressing Research Objective 1, we will test these models at scale. This casts light on the models' statistical robustness and the scope of their applicability.

1.2.3 Research Objective 3

Develop novel precision medicine artificial intelligence models for prognosis in hematological diseases

Artificial intelligence (AI) offers novel approaches to classification, regression, and prediction. In healthcare, AI prognostic models offer a new degree of personalization. Our third research objective is to develop such models. By comparing results to those of existing models, we subsequently provide insight into the performance of the developed AI models.

1.2.4 Research Objective 4

Show that hematological disease models can be ethically and securely trained and integrated for analysis and application, without breaching patients' rights and trust.

While greater information availability fosters more advanced insights, technology, and superb models, the fundamental rights of patients should not be violated. Our fourth research objective is therefore to reconcile (AI) model development with data privacy. This is achieved by leveraging interoperability standards and a novel Privacy Enhancing Technologies (PET): synthetic data generation.

1.3 Relation to the GenoMed4All Project

The research documented in this thesis was part of the GenoMed4All project [10]. GenoMed4All is the European initiative to transform the response to Haematological Diseases by seizing the power of Artificial Intelligence. The GenoMed4All project facilitated access to data, expertise, collaboration, and European Union funding. GenoMed4All is a pan-European collaboration of 23 partners from the whole value chain, including healthcare professionals, regulatory and ethics research, academia, disruptive tech and digital service provision.

1.4 Structure of the Thesis

The remainder of the thesis is structured as follows: In Chapter 2, we address Research Objective 1. This includes a background on omics and AI-based precision medicine. We also delve into the potential of these fields in hematology. We then study the European hematology data landscape. In doing so, we delve deeper into the GenoMed4All partners that facilitate most of the remaining research in this thesis. Finally, we address data privacy and ethics. In particular, we provide an in-depth discussion of the concepts of privacy and personal data, including contemporary legal frameworks. We then describe the ethical framework adhered to in this thesis.

In Chapter 3, we address Research Objective 2, leveraging the available data to study existing prognostic frameworks. In particular, we validate the recently proposed Molecular International Prognostic Scoring System (IPSS-M) for myelodysplastic syndromes (MDS). In doing so, we also compare it the the older Revised International Prognostic Scoring System (IPSS-R). This research casts light on how wider data availability leads to more accurate and effective prognostic models, even without the use of AI.

In Chapter 4, we study the potential impact of AI and omics data for precision medicine in hematology. In particular, we use Bayesian networks and Dirichlet processes, combining mutations in 47 genes with cytogenetic abnormalities to identify genetic associations and subgroups. This allows us to identify eight MDS groups (clusters) according to specific genomic features. Each group has distinct and specific clinical features and patterns of evolution. As such, the study shows that the use of omics-based AI is highly promising in hematology disease classification and prognosis.

In Chapter 5, we show that omics-based AI research in hematology can be conducted in a safe, privacy-respectful manner. This can be achieved through a novel privacy enhancing technology (PET): synthetic data generation. Synthetic data is not collected empirically. Instead, it is algorithmically generated. This means that records in a synthetic dataset do not correspond to real patients. As such, real patients' right to privacy is protecte when synthetic data substitutes real data in studies. Recent advances in generative AI have made it possible to generate highly realistic synthetic data, combining privacy protection with analytic accuracy.

Our study is essentially a proof-of-concept of such synthetic data in hematology. We develop synthetic data through a generative AI framework (generative adversarial network, or GAN). We then describe metrics to measure how well such data preserves relevant information (fidelity), and how well it protects real patients' privacy. Applying these metrics, we find that synthetic data form a promising tool for omics-based AI and precision medicine research in hematology.

The thesis is concluded with a discussion of the obtained findings and suggestions for future research in Chapter 6.

Chapter 2

The Hematology Data and AI Landscape

Access to reliable, accurate, and current data is crucial when engaging in omics and AI research. Unfortunately, in the context of the life sciences, such data is often scarce; distributed over multiple centers; and subject to stringent privacy policies. When aiming to conduct AI-based omics studies in hematology, researchers therefore typically hindered. In this chapter, we outline the opportunities of omics data in hematology, as well as the difficulties in obtaining them. In particular, we provide a theoretical background for omics data and AI in hematology. We then outline how relevant data is collected and stored within the European Union. Subsequently, we discuss the data protection frameworks affecting the data's use. The chapter is concluded with an overview of the steps taken throughout this thesis to ensure data is treated in a responsible, secure, and ethical manner. In doing so, we address our first research objective: *Map the hematology data landscape, including the AI and omics opportunities; the data repositories in the European Union; and the relevant laws, regulations, and ethical guidelines.*

2.1 AI and Omics-Based Precision Medicine and its potential in Hematology

2.1.1 Definitions

Personalized or precision medicine is a medical model that combines already established clinical–pathological results with advanced profiling. This helps practitioners take individuals' properties into account at key stages of their patient journeys. So-called "(multi-)omics" data plays an increasing role in personalized medicine. Omics is an umbrella term for biological domains whose name ends in "omics" (for instance, genomics, proteonics, transcriptomics).

2.1.2 Impact on Life Sciences

Analysis of omics data has led to important insights, particularly when analyzed through AI methods. In a recent survey, Stanojevic et al. [11] note that computational methods for omics data offer revolutionary insights into cellular states and biological processes. Unfortunately, they note that the integration of the multiple omics data assets is computationally difficult.

Lorkowski et al. [12] further confirm the huge potential of AI for precision medicine, particularly when pooling omics data. In surveying 1572 articles on the matter, they conclude that this approach is ushering in a Fourth Industrial Revolution in medicine and healthcare. They highlight its potential for drug repurposing and new therapeutical modalities in particular.

He et al. [13] note that omics data and AI, when combined, facilitate a new degree of personalization in medicine. They prove this by pooling epigenome, transciptome, proteome, metabolome and other data. They then use AI algorithms to improve early cancer screening, diagnosis, response assessment, and prognosis prediction. Like Stanjovic at al. [11] and Lorkowski et al. [12], they expect the combination of omics data and AI to cause a paradigm shift in precision medicine.

Arjman et al. [14] provide an overview of use cases of AI paired with omics data in cancer research. For each use case, they propose a suitable AI method. Their survey shows that the approach can overcome key problems in medical research that extend beyond just oncology. These include the high dimensionality of omics data; possible data heterogeneity; class imbalance; missing data; and more. The authors conclude that ML is a promising tool in discovering effective diagnostic and therapeutic approaches to cancer growth.

2.1.3 Potential in Hematology

Shouval et al. [15] remark that machine learing (ML) and AI in general have considerable potential in hematology as well. They state that is particularly true if it can leverage both electronic health records (EHRs) and genomic data. Shouval et al. [15] also list the obstacles to successful application of AI in hematlogy. In particular, they list lack of available data as the main obstacle. Other obstacles include insufficient data quality and data ethics.

Haferlach and Wencke [16] study the efficacy of AI-based next-gen sequencing and whole genome sequencing (WGS) in hematology diagnostics. They expect that ML-based sequencing will soon outperform human diagnosis. They advocate for AI-based genome profiling in routine standard care.

An extensive review of AI in hematology was conducted by El Alaoui et al. [17]. They find that AI is highly effective in screening, diagnosis, and treatment stages of HDs. However, they note that the role of patients' data in AI use should be further explored.

These views are echoed by Lin et al. [18], who find that AI offers a clear potential for breakthroughs in hematopathology. Specific use cases they list include diagnosis, classification and treatment guidelines for HDs. They also investigate the role of bone marrow analysis. They stress that AI can reduce the turnaround time in HD diagnosis.

2.2 Hematology Data in the European Union

As outlined, AI and omics have the potential to revolutionize hematology. Unfortunately, data scarcity and ethical concerns severly limit its scope. National approaches for HDs clinical management and research are often ineffective, especially for rare conditions.

Development of infrastructures that can support collection and use of genomic information in the health-care community is therefore a research priority for HDs, as repositories of genomic and clinical information in Europe are unconnected. The number of distinct HDs is large. For individual diseases, the number of data samples is therefore typically small. This hinders the implementation and maintenance of central big data repositories as exist in other areas. for genomic profiling of HDs, researchers therefore need to acquire sufficient patient data through clinical networks.

Within the EU, information on rare diseases is managed by European Reference Networks (ERNs). These are networks of life science experts working with diseases that are rare and/or complex. *EuroBloodNet* (see [19]) is the ERN for Rare Hematological Diseases (RHD). It was approved by the European Commission (EC) in December 2016, and started its activities in March 2017. EuroBloodNet encompasses oncological and non-oncological RHDs. It brings together unique talent pool of 66 highly sophisticated and multidisciplinary healthcare teams in 15 Member States, and advanced specialized medical equipment and infrastructures [19].

This thesis will make use of the existing infrastructures and initiatives, including powerful High Performance Computing facilities, hospital registries, data processing tools, and pre-existing repositories. Data is provided by EuroBloodNet, as well as ten clinical partners (names confidential).

2.3 Privacy and Data Protection

2.3.1 The Evolution of Privacy as a Concept

The concept of privacy is multifaceted and dynamic, initially articulated as the 'right to be let alone'[20]. It has since broadened to include various dimensions such as physical, proprietary, decisional, and informational elements[21]. These dimensions collectively embody different principles of privacy, encompassing anonymity, solitude, secrecy, and the ability to control access to one's personal information[22]. Despite its diverse interpretations, privacy is universally recognized as a fundamental human right that reinforces human dignity and forms the foundation for other constitutional protections.

In our digital society, characterized by rapid information exchange [23], the boundary between public and private domains has substantially evolved [24]. These developments have reshaped our comprehension of privacy. Current understandings of privacy go beyond the notion of 'the right to be let alone', encompassing rights to control personal data, preserve confidentiality, uphold personal dignity, and safeguard intimacy [25].

2.3.2 The Concept of Personal Data

The concept of personal data varies with time and location. In the European Union (EU), the *General Data Protection Regulation* (GDPR) [26] was implemented in 2018. In Article 4.1, it defines personal data as *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier.* Regulation EU 2016/679 further details these twelve categories of such classifiers: name; identification number; location data; online identifier; and specific physical; physiological; genetic; biometric; mental; economic; cultural; and social identity information. This places hematology and omics data well within the scope of the GDPR.

Another influential data privacy legislation affecting the life sciences is the *Health Insurance Portability and Accountability Act* (HIPAA, [27]), its jurisdiction being the United States of America. HIPAA follows EU framework in requiring protection of attribute categories. Its seventeen categories span roughly the same information as those of Regulation EU2106/679. However, Genertic information is interpreted more broadly. As it is hereditary and shared between biological family members, a genetic data specimen may be personal to more than one individual. Genetic test results of family members (including embryos), manifestations of genetic diseases or disorders in family members, requests for, or receipt of genetic services on the individual or family member are therefore classified as one's personal data under HIPAA [28]. Naturally, this classification limits the options for omics research.

2.3.3 Data Anonymization Techniques

A selection of data anonymization techniques has been widely adopted in the industry [29]. These technologies typically rely on pseudonymization; the statistical distortion of data; and/or generalization of data properties. For example, an individual's location data may be generalized from their city to their wider region of residence, offering an added degree of plausible deniability [30]. Unfortunately, these methods reduce the accuracy and specificity of the data. In a 2013 lawsuit, the degree of privacy protection of these techniques was assessed. The involved parties concluded that data was only sufficiently protected once the distortion of the data was so significant, that it rendered further analysis useless [31].

More innovative approaches to reconcile data protection with analytic utility have emerged. Synthetic data [32] is entirely artificial, hence void of real patients' personal information. Through generative AI, synthetic data can preserve the patterns relevant for omics research. Importantly, the proposed AI Act [33] equates synthetic data to anonymous and non-personal data (see Article 54, point 1b). Federated learning [34] is a method used to tain overall machine learning models on multiple distinct databases. The model then leverages the complete information of the combined databases. However, the data does not have to be exchanged between centers. These techniques are discussed in detail in Chapter 5.

2.4 Ethical Approach to Data and their Use

The author acknowledges the importance of ethics, privacy and data protection and other relevant fundamental rights, societal acceptance and other regulatory requirements regarding the development, implementation and the use of technology. Participants in all studies signed informed consent forms. We consistently adhered to the EC Ethics guidelines for trustworthy AI throughout the research. This means we ensured our methods respect all applicable laws and regulations; respect ethical principles and values; and are robust with minimal bias from a technical perspective [35].

Along with demonstrating novel PETs, we take several precautions are taken to ensure: (a) the overall ethical use of AI, including its appropriate application to patient stratification and to personalized decision making, which need to uphold ethical principles such as those published by the European Commission (EC) Expert Group; (b) the ethical development and validation of the AI algorithms, so that these are trustworthy in clinical/patient use; (c) the ethical deployment and adoption of the solutions with embedded AI, so that clinical judgment and decision making retains its appropriate role in care management.

Key elements of this thesis' ethical approach include: a) an Impact Assessment framework of ethics, privacy and data protection and other relevant fundamental rights and societal acceptance; and b) continuous monitoring of ethical issues during the entire duration of the research. This includes explicit mention of the selection basis and biases in the data used for AI training and validation, on the assumption that it is most important to have transparency about bias that is perhaps impossible to fully eliminate.

Throughout the research, Transport Layer Security (TLS) cryptographic protocols with symmetric and asymmetric encryption. Besides, data flows between cluster nodes are subject to TLS cryptographic protocols to prevent access by operational resources. Once data is ingested, data at rest is encrypted at multiple-levels including storage level encryption, and if required field-level hashing, and AES/RSA encryption during ingestion together with access control policy. Data stored is subject to Strict Role-based Access Control (RBAC) mechanisms that govern access to the whole ingested dataset. This RBAC implements a model that supports both cluster level (operations) privileges and index level (table/field) privileges.

2.5 conclusion

Many authors note the revolutionary potential of AI and omics data in hematology. Unfortunately, this potential is not sufficiently realized due to a number of obstacles. Most notably among these are data scarcity and data ethics.

As such, privacy and data protection play a central role in AI-based hematology research. In the EU, privacy is safeguarded under the GDPR. A comparable law in the healthcare sector in the USA is the HIPAA. Both documents play a crucial role in formalizing concepts like personal data and privacy. Oftentimes, traditional data anonymization methods fail to meet the guidelines this puts forward. More innovative technological approaches to data protection can offer more ethically and legally sound results. These include the use of synthetic data and federated learning.

In Chapter 5, these innovative approaches are explored in detail. Additionally, we adhere to strict data ethics guidelines throughout the presented research. This includes approval by ethics committees; use of written consent forms; impact assessments; continuous monitoring of ethical issues that may arise; and top-tier cybersecurity measures.

Chapter 3

Real-World Validation of Existing Prognostic Hematology Models

In this chapter, we address the second research objective, namely: *Leverage increased data availability to validate existing prognostic models for hematological diseases at scale*. To do so, we will first provide a brief introduction to the two most common such models: the *Revised-International Prognostic Scoring System* (IPSS-R) and the *Molecular International Prognostic Scoring System* (IPSS-M). We then validate the novel IPSS-M and compares it to the IPSS-R in the context of myelodysplastic syndromes. The research was previously published as [36].

3.1 Introduction

Myelodysplastic syndromes (MDS) are heterogeneous neoplasms ranging from indolent conditions to cases rapidly progressing into acute myeloid leukemia and therefore a risk-adapted treatment strategy is needed [37]. Disease-related risk is currently assessed by the Revised International Prognostic Scoring System (IPSS-R), on the basis of bone marrow blasts, blood cytopenias, and cytogenetic abnormalities [38].2 Although IPSS-R is an excellent tool for clinical decision making, this scoring system has its weaknesses and may fail to capture reliable prognostic information at individual patient level [39, 40].

In MDS, conventional prognostic tools on the basis of clinical and hematologic features are being complemented by introducing somatic gene mutations that were shown to be valuable prognostic markers [40, 7, 41, 42, 43]. Recently, the International Working Group for Prognosis in MDS (IWG-PM) proposed a clinical-molecular prognostic model (Molecular IPSS, IPSS-M) that was developed using hematologic parameters, cytogenetic abnormalities, and mutations of 31 MDS-related genes [44]. IPSS-M improved prognostic discrimination across all clinical end points compared with IPSS-R.

In this chapter, we address the issue of clinical implementation of IPSS-M by: (i) providing an extensive validation of its prognostic value (also focusing on patients without detectable mutations); (ii) investigating the predictive and prognostic power of IPSS-M in patients receiving disease-modifying treatment (hypomethylating agents [HMA] and hematopoietic stem cell transplantation [HSCT]); and (iii) testing the accuracy in predicting IPSS-M when molecular information was missed to define a minimum set of relevant genes associated with high performance of the score.

3.2 Methods

3.2.1 Study Populations and Procedures

The study was conducted by GenoMed4All consortium [45] and supported by EuroBloodNET, the European Reference Network on rare hematologic diseases [19]. The Humanitas Ethics Committee approved the study. Written informed consent was obtained from each participant. This study was registered at ClinicalTrials.gov (ClinicalTrials.gov identifier: NCT04889729).

Inclusion criteria were age ≥ 18 years, a diagnosis of primary MDS according to WHO 2016 criteria [46], and available information on IPSS-M related variables collected at diagnosis and before starting disease-modifying treatments (if any). Patients affected with therapy-related myeloid neoplasms or incomplete information on IPSS-M variables were excluded. A total of 2,876 patients matched study criteria. A Data supplement is available with the online edition of the study, see [36].

Karyotypes were classified using the International System for Cytogenetic Nomenclature Criteria. Mutation screening of MDS-related genes was performed on DNA bone marrow mononuclear cells or peripheral blood granulocytes (Data Supplement). Patients were reclassified according to WHO 2022 and International Consensus Classification of Myeloid Neoplasms criteria [47, 48]. IPSS-M score was calculated according to the original publication [44].

3.2.2 Statistical Analysis

Survival curves were estimated with the Kaplan-Meier method and differences among groups were evaluated by log-rank test. Overall survival (OS) and leukemia-free survival (LFS) were defined as the time between diagnosis and death (from any cause) or last follow-up (for censored observations) and the time between diagnosis and acute myeloid leukemia evolution (if any) or last follow-up (for censored observations), respectively. When focusing on patient populations receiving a specific treatment, OS was calculated as the time between start of treatment and death/last follow-up. The probability of relapse after treatment was estimated according to standardized criteria.15 For patients treated with HSCT, when estimating nonrelapse mortality (NRM), any death in the absence of disease relapse was considered an event. The cumulative incidence of relapse and NRM was estimated by competing risk approach [49].

Multivariable survival analyses were performed by Cox's proportional hazards regression models (IPSS-M was incorporated as ordinal variable in the models). The discriminatory power of the models and the relative goodness of fit for the predictive score were evaluated using Harrell's concordance index [50]. To compare different statistical models, we used in addition the Akaike information criterion (AIC) [51], which allows the evaluation of a model by combining goodness of fit and complexity, with a lower AIC indicating a better trade-off between fit and complexity.

The impact of single IPSS-M factors on the prediction of clinical outcomes was evaluated by fitting a random-effects Cox's model [52, 53]. The percentage of variation of the logarithmic hazard explained by each set of variables was estimated (see Data Supplement of [36]).

The accuracy of IPSS-M in predicting the probability of survival in the presence of missing molecular data was calculated as the number of correctly classified patients divided by the size of patient's cohort. The accuracy loss was calculated as the fraction of wrongly classified patients divided by the population size.

3.3 Results

3.3.1 Clinical Characteristics of Patients and Gene Mutations

Clinical features at diagnosis of the 2,876 patients with MDS enrolled in the study are reported in Tables 3.1 through 3.5. Study participants included 1,743 men (61%) and 1,133 women (39%). Date range of diagnosis was from 1999 to

2018. Median age at diagnosis was 68 years (range, 18-96 years). Follow-up was updated on December, 2020. Median duration of follow-up was 37.5 months (95% CI, 36.2 to 38.8 months).

Demographic	**No./No. (%)/(range)**
Patients, No.	2,876
Female/male, No. (%)	1,122/1,743 (39/61)
Age, years (range)	68 (18-96)

Table 3.1: Demographic features of 2,876 patients with MDS from the GenoMed4All cohort collected at the time of diagnosis

2016 WHO Category	**No. (%)**	**2022 WHO Category**	**No. (%)**	**2022 ICC Category**	**No. (%)**
MDS-5q-	142 (5)	MDS-LB5q-	133 (4.6)	MDS-*SF3B1*	398 (13.8)
MDS-SLD	175 (6)	MDS-LB-*SF3B1*	398 (13.8)	MDS-del(5q)	133 (4.6)
MDS-MLD	649 (22.6)	MDS-bi*TP53*	153 (5.3)	MDS, NOS without dusplasia	15 (0.5)
MDS-RS-SLD	132 (4.6)	MDS-LB	867 (30.2)	MDS, NOS with SLD	173 (6)
MDS-RS-MLD	325 (11.3)	MDS-IB1	531 (18.5)	MDS, NOSS, with MLD	679 (23.6)
MDS-EB1	572 (20)	MDS-IB2	794 (27.6)	MDS-EB	531 (18.5)
MDS-EB2	864 (30)			MDS/AML	794 (27.6)
MDS-U	17 (0.6)			MDS with mutated *TP53*	83 (2.9)
				MDS/AML with mutated *TP53*	70 (2.4)

Table 3.2: Hematopathologic features of 2,876 patients with MDS from the GenoMed4All cohort collected at the time of diagnosis

Hematologic Feature	**Median (range)**
Hemoglobin, g/dL	10.0 (2.2 - 16.3)
Neutrophils, $\times 10^9$/L	1.7 (0 - 11.7)
Platelets, $\times 10^9$/L	110 (2 - 491)

Table 3.3: Hematologic features of 2,876 patients with MDS from the GenoMed4All cohort collected at the time of diagnosis

Clinical Feature	**No. (%)**
Cytogenetic risk according to IPSS-R criteria	
Very good	31 (1.1)
Good	1,909 (66.4)
Intermediate	351 (12.2)
Poor	236 (8.2)
Very poor	349 (12.1)
IPSS-R risk group No.	
Very low	293 (10)
Low	806 (28)
Moderate low	610 (21)
Moderate high	595 (21)
Very high	572 (20)
IPSS-M risk group No.	
Very low	275 (9.6)
Low	797 (27.7)
Moderate low	306 (10.6)
Moderate high	319 (11.1)
high	555 (19.3)
Very high	624 (21.7)

Table 3.4: Clinical features of 2,876 patients with MDS from the GenoMed4All cohort collected at the time of diagnosis

Treatment	**No. (%)**
Erythroid stimulating agents	356 (12.3)
Hypomethylating agents	673 (23.4)
AML-like chemotherapy	301 (10.4)
Transplantation	964 (34.0)
Other	89 (3.1)

Table 3.5: Treatment features of 2,876 patients with MDS from the GenoMed4All cohort collected at the time of diagnosis

Considering IPSS-M–related genomic features, we identified 6,749 genomic lesions at diagnosis (median, 3; range, 0-12). 2,421 patients (84.1%) presented one or more genomic alterations (mutations and/or chromosomal abnormalities). 2,369 patients (82.4%) had one or more somatic mutations on 31 IPSS-M–related genes, whereas 1,297 showed abnormal karyotype (see Figure 3.1).

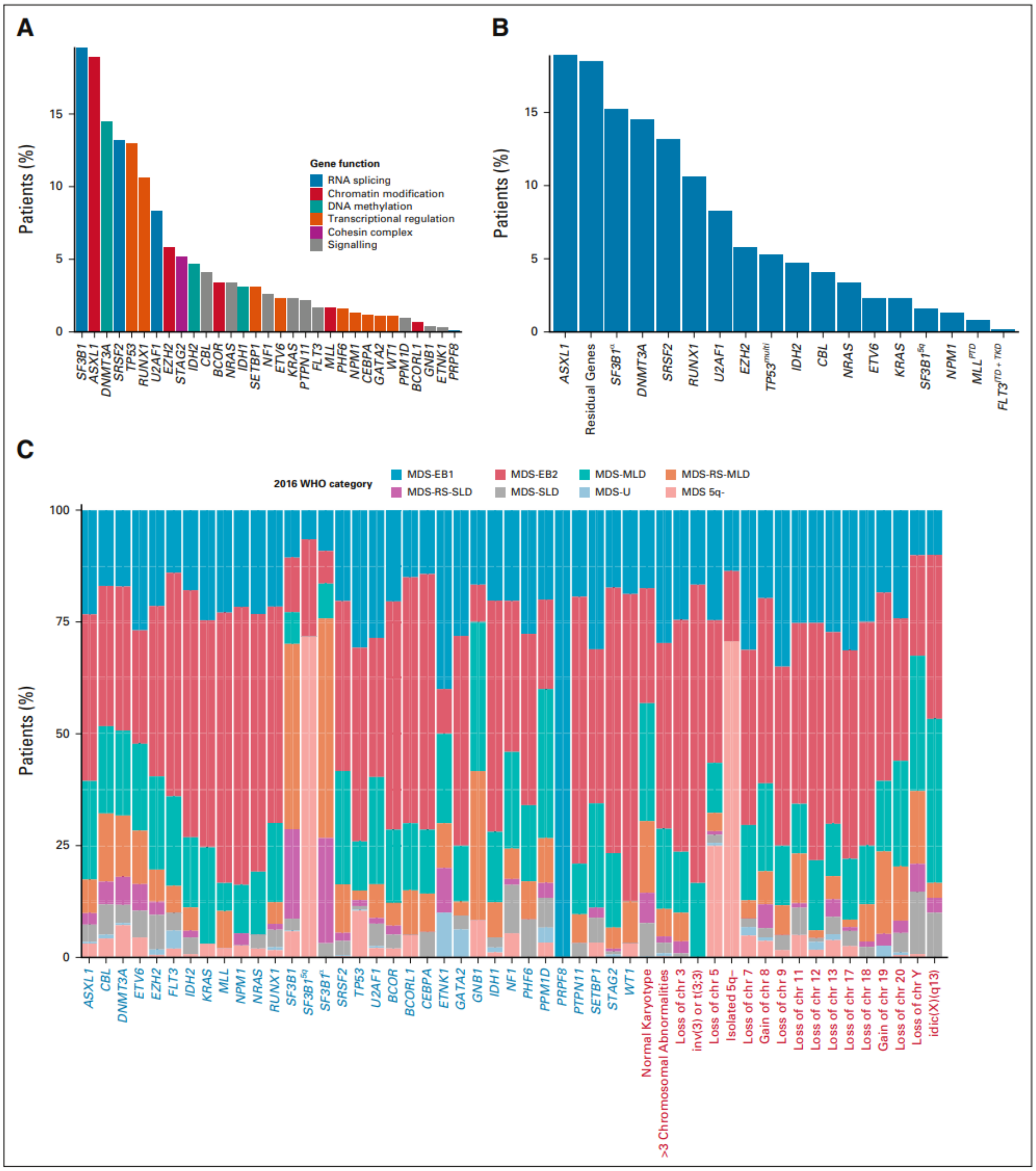
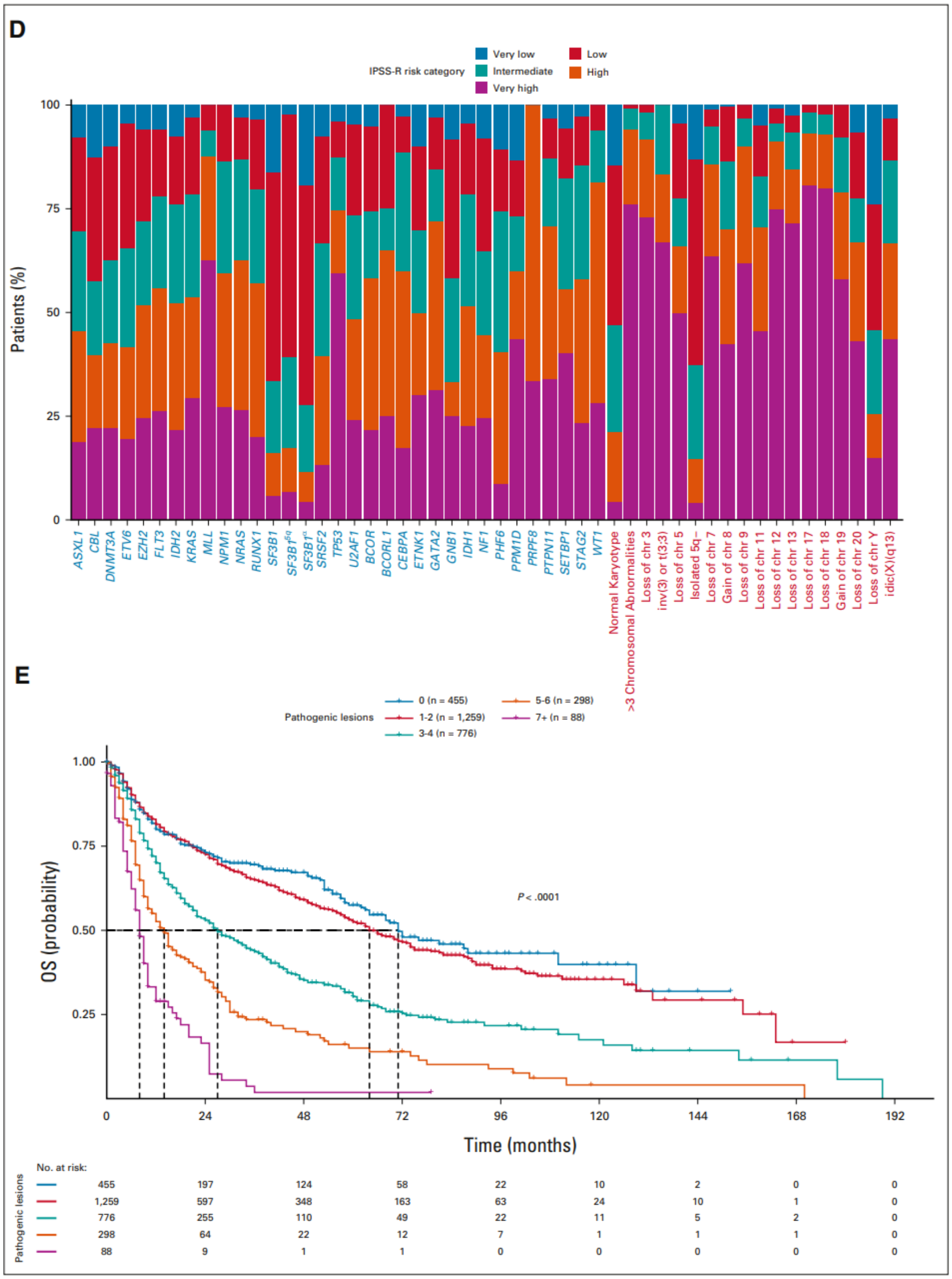


Figure 3.1: Molecular landscape of patients with MDS from the GenoMed4all cohort. (A) Frequency of mutations of the 31 genes included into IPSS-M score in 2,876 patients from the GenoMed4All cohort. Colors linked to the bars represent the gene function. (B) Frequency of mutations on genomic features grouped according to IPSS-M criteria. (C and D) Frequency of gene mutations and chromosomal abnormalities broken down by MDS subtypes according to 2016 WHO criteria and IPSS-R risk category, respectively. Mutations on genes are grouped according to IPSS-M criteria as main effect genes (gene labels are highlighted in blue) and residual genes (gene labels are highlighted in dark green). (E) Kaplan-Meier probability estimates of OS across numbers of oncogenic alterations per patient (gene mutations and cytogenetic abnormalities). P value is from log-rank test. Frequency of IPSS-M–related gene mutations and chromosomal abnormalities broken down by MDS subtypes according to 2022 WHO criteria and ICC criteria is available in the Data Supplement. ICC, International Consensus Classification of Myeloid Neoplasms and Acute Leukemia; MDS, myelodysplastic syndromes; MDS 5q-, MDS with isolated deletion of long arm of chromosome five; MDS-EB1, MDS with excess of blasts, type 1; MDS-EB2, MDS with excess of blasts, type 2; MDS-MLD, MDS with multilineage dysplasia; MDS-RS-MLD, MDS with ring sideroblasts and multilineage dysplasia; (continued on following page)



455	197	124	58	22	10	2	0	0
1,259	597	348	163	63	24	10	1	0
776	255	110	49	22	11	5	2	0
298	64	22	12	7	1	1	1	0
88	9	1	1	0	0	0	0	0

Figure 3.1 (cont.): (Continued). MDS-RS-SLD, MDS with ring sideroblasts and single-lineage dysplasia; MDS-SLD, MDS with single-lineage dysplasia; MDSU, MDS unclassifiable; IPSS-R, Revised International Prognostic Scoring System; IPSS-M, Molecular International Prognostic Scoring System; OS, overall survival.

3.3.2 Validation of the Prognostic Power of IPSS-M and Comparison With IPSS-R

We calculated IPSS-M in the study cohort at diagnosis [44]. Cytogenetic abnormalities were classified according the IPSS-R criteria [38]. Gene mutations were considered as binary variables with the exception of TP53 (not mutated, monoallelic mutation, multihit mutations) and SF3B1 (SF3B15q [SF3B1 mutation in the presence of isolated del(5q) only or with one additional aberration excluding -7/del(7q)], and SF3B1α [SF3B1 mutation without comutations in BCOR, BCORL1, RUNX1, NRAS, STAG2, SRSF2, and del(5q)]9).

Accordingly, 9.6% of patients (n = 275) were classified as very low risk, 27.7% (n = 797) as low risk, 10.6% (n = 306) as moderate low risk, 11.1% (n = 319) as moderate high, 19.3% (n = 555) as high risk, and 21.7% (n = 624) as very high risk.

We analyzed the probability of OS and LFS for all IPSS-M categories. Patients who received HSCT were censored at the time of the procedure. IPSS-M categories showed significantly different probabilities of both OS and LFS (both $p < .001$; see Figure 3.2). The independent effect of IPSS-M on clinical outcome was maintained in a multivariable model including age and sex as covariates (HR, 1.67; 95% CI, 1.61 to 1.73; $p < .001$ OS; and HR, 1.79; 95% CI, 1.73 to 1.86; $p <.001$ for LFS).

IPSS-M showed superior performance with respect to conventional IPSS-R scoring system: concordance was 0.81 (95% CI, 0.79 to 0.82) versus 0.74 (95% CI, 0.73 to 0.76) for OS and 0.89 (95% CI, 0.87 to 0.91) versus 0.76 (95% CI, 0.73 to 0.79) for LFS, respectively. In addition, to evaluate the effect of IPSS-M versus IPSS-R, we fitted two separate multivariable Cox's models including age and sex as covariates, comparing them by the AIC. AIC for the model with IPSS-M versus IPSS-R was 17,455.43 versus 17,469.33 for OS and 3,973.26 versus 4,011.64 for LFS, thus confirming the importance of accounting for gene mutations in the prognostic model.

The five-to-five comparison of IPSS-R and IPSS-M patients' distribution (in which we merged moderate low and moderate high to moderate in IPSS-M) resulted in the restratification of 46% of patients (1,324 of 2,876). Of these, 23.6% (n = 679) were upstaged and 22.4% (n = 645) were downstaged (Figure 3.2). A total of 115 patients (4%) were reclassified by more than one risk strata. Highlighting the implications of this restratification, marked differences in survival were observed between IPSS-M categories within each IPSS-R risk category; by contrast, the IPSS-R did not stratify patient outcomes within IPSS-M risk strata (Data Supplement).

We specifically studied the prognostic impact of gene mutations on main effect IPSS-M genes that were associated with adverse prognosis9 and their contribution on patients restratification from IPSS-R to IPSS-M risk categories (Data Supplement). Among restratified patients, 193 (26%) had one mutated adverse IPSS-M main effect gene, whereas 275 (37%) had two or more mutated genes. In details, in the very low + low IPSS-R category (n = 1,099), 214 patients (19.5%) were upstaged, of which 198 (93%) have more than one mu-

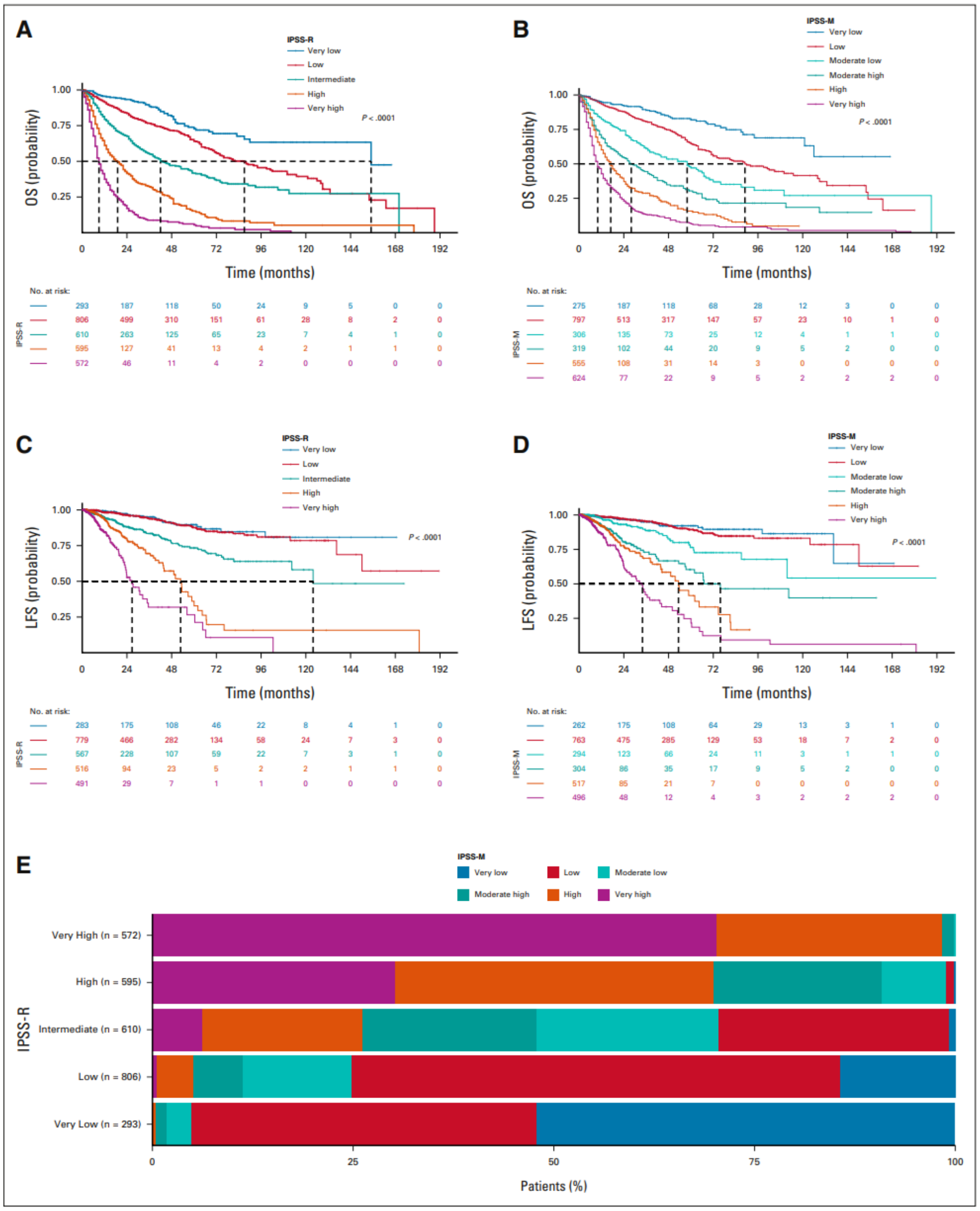


Figure 3.2: Clinical assessment of IPSS-R and IPSS-M in the GenoMed4All MDS cohort. Kaplan-Meier probability estimates of (A and B) OS and (C and D) LFS for 2,876 patients with MDS from the GenoMed4All cohort stratified by IPSS-R and IPSS-M risk categories, respectively. P values are from log-rank test. (E) Restratification of IPSS-R to IPSS-M risk groups in the MDS cohort. Each bar represents an IPSS-R category and shows the percentage of patients that is restratified in the IPSS-M categories (indicated with different colors). (F) Distribution of the restratified patients in each IPSS-R category, counting the proportion of patients who are downstaged (highlighted in blue) and upstaged (highlighted in red) with the IPSS-M classification. (G) Distribution of the restratified patients in each IPSS-R category, counting the proportion of patients with more than one shift in IPSS-M risk category (downstaged and (continued on following page)

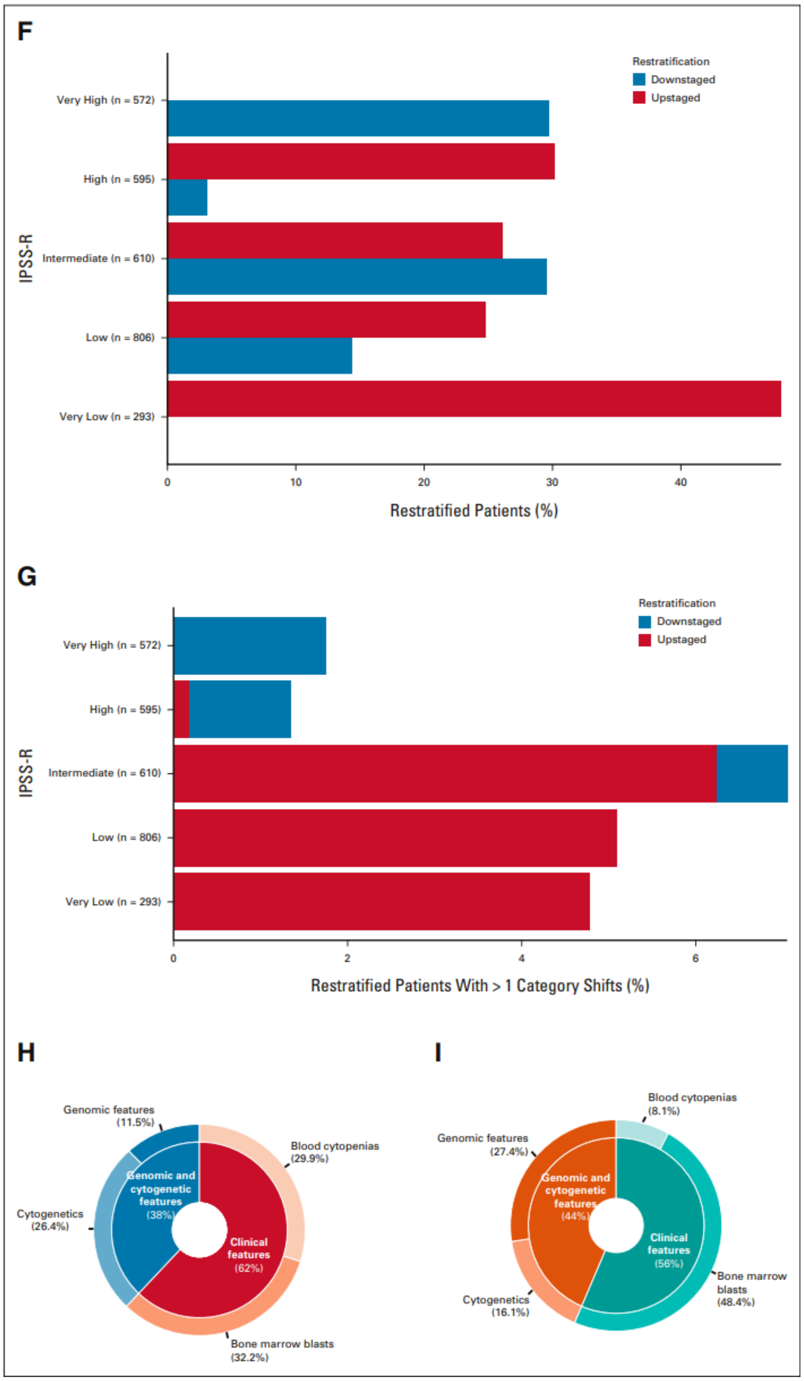


Figure 3.2 (cont.): (Continued). upstaged cases are highlighted in blue and in red, respectively). (H-I) Fraction of explained variation related to the IPSS-M prognostic factors for OS and LFS, respectively. IPSS-R, Revised International Prognostic Scoring System; IPSS-M, Molecular International Prognostic Scoring System; LFS, leukemia-free survival; MDS, myelodysplastic syndromes; OS, overall survival.

tated IPSS-M genes. Considering patients classified in the intermediate IPSS-R category (n = 610), 180 (29%) were downstaged, the majority of them had no mutations (67%); by contrast, 159 subjects (26%) were upstaged, and 69% of these patients carried two or more main effect IPSS-M mutated genes. In the very high + high IPSS-R category (n = 1,167), instead, 189 patients (16%) were reclassified in lower-risk classes and only 33% of these presented more than one mutated gene. Thus, patient restratification was not a single gene effect, but the cumulative contribution of the prognostic mutations for each subject.

Then, we addressed the issue of the prognostic value of IPSS-M in those patients without detectable mutations in the 31 IPSS-M–related genes. 507 subjects entered the analysis. IPSS-M categories maintained a significant effect on probability of both OS and LFS (both $p < .001$; Data Supplement). IPSS-M maintained superior performance to conventional IPSS-R scoring system in this patient setting: concordance was 0.89 [0.86-0.91] versus 0.73 [0.69-0.77] for OS and 0.91 [0.90-0.92] versus 0.81 [0.75-0.87] for LFS, respectively. By comparing two multivariable models including IPSS-M versus IPSS-R, AIC was 1,573.04 versus 1,590.11 for OS, respectively, and 491.91 versus 498.61 for LFS, respectively, thus confirming the best prognostic performance of IPSS-M in this population.

Finally, we evaluated the prognostic impact of IPSS-M–related variables in terms of percentage of explained variation for clinical outcomes (OS and LFS; Figure 3.2). Clinical features had a high predictive power for both OS (bone marrow blasts and cytopenias) and LFS (bone marrow blasts). IPSS-M–related genomic variables had a strong predictive power, that is increased for the LFS outcome, highlighting the impact of genomic landscape on the prediction of the risk of disease evolution.

3.3.3 Predictive and Prognostic Effect of IPSS-M in Patients Receiving Specific Treatments

In MDS, an increasing proportion of patients undergo to disease-modifying therapies, including HSCT and HMA (for high-risk subjects who are not eligible to HSCT). Therefore, it is relevant to know if IPSS-M may provide information on the probability of response to specific treatments (predictive value) and the probability of survival after treatment (prognostic value).

We therefore analyzed the predictive/prognostic value of IPSS-M in two populations treated with HSCT and HMA according to currently available guidelines on the basis of IPSS-R, age, performance stats, and donor avaialbility.1 To investigate the predictive value of IPSS-M, the risk of disease relapse in patients treated with HSCT and the overall response rate (including the achievement of complete response (CR), partial response, marrow CR, and stable disease with hematologic improvement according to 2006 IWG criteria)15 for patients treated with HMA were used as primary end points, while the prognostic value of IPSS-M was tested on the probability of OS since the start of treatment in both cases.

Nine hundred sixty-four patients receiving HSCT entered the analysis, in which clinical and genomic information for IPSS-M calculation was available at the time of transplant in patients who were transplanted upfront and before chemotherapy/HMA in those receiving treatment before transplantation (Data Supplement). Patients receiving HSCT were reclassified according to IPSS-M criteria: 126 (13.1%) patients were classified as low-risk, 108 (11.2%) patients as moderate low, 136 (14.1%) as moderate high, and 290 (30.1%) and 304 (31.5%) as high and very high risk, respectively. As illustrated in Figure 3.3, the 5-year OS probability was 61% in low-, 55% in moderate low-, 46% in moderate high-, 33% in high-, and 27% in very high-risk patients ($p < .0001$). In these risk groups, by competing risk analysis, the 5-year cumulative incidence of relapse was 14%, 14%, 15%, 20% and 29%, respectively ($p < .001$; Fig Fig3).3). A five-to-five mapping between the IPSS-R and IPSS-M categories resulted in the restratification of 45% ($n = 433$) of the patients. Of these, 21% ($n = 204$) were upstaged and 24% ($n = 229$) were downstaged (Figure 3.3).

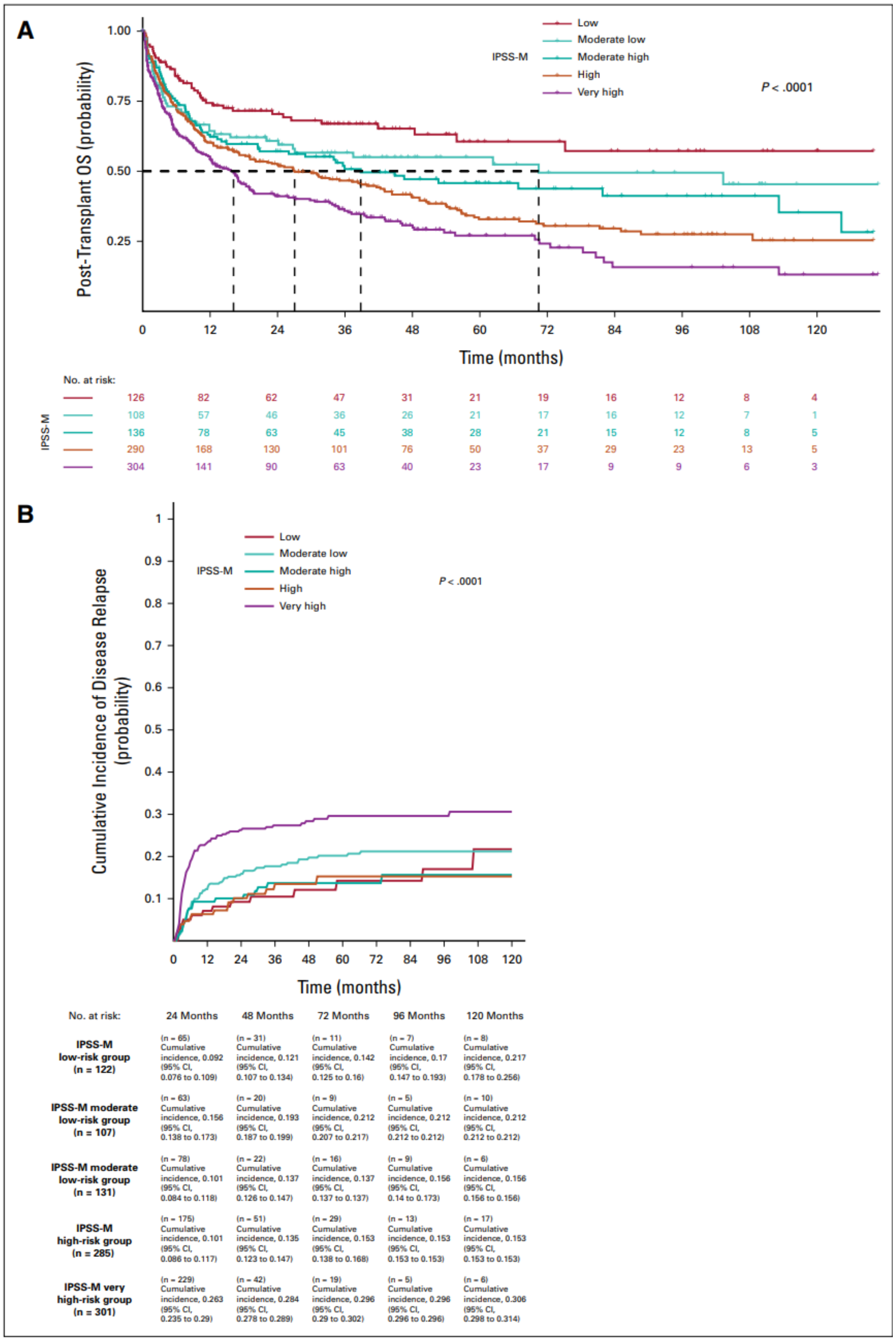
We analyzed the prognostic effect of the IPSS-M score by a multivariable model, including recipient age and sex, time from diagnosis to transplantation, source of hematopoietic stem cells, type of donor, disease status at transplant (active/progressive disease v complete remission), and conditioning regimen (reduced-intensity v standard conditioning). The IPSS-M score was significantly associated with OS (HR, 1.18 [95% CI, 1.08 to 1.27]; $p < .001$) and probability of relapse (HR, 1.38 [95% CI, 1.21 to 1.56]; $p < .001$)

IPSS-M showed superior performance to conventional IPSS-R in predicting both OS and probability of relapse after HSCT (concordance was 0.76 [95% CI, 0.73 to 0.78] v 0.60 [95% CI, 0.57 to 0.64] for OS, and 0.89 [95% CI, 0.87 to 0.91] v 0.70 [95% CI, 0.65 to 0.74] for probability of relapse, respectively). By comparing two multivariable models including IPSS-M versus IPSS-R, AIC was 6,545.87 versus 6,559.96 for OS, respectively, and 2,404.97 versus 2,416.78 for probability of relapse, respectively, thus confirming the best prognostic performance of IPSS-M in predicting post-transplantation outcomes.

Recipient age was a significant risk factor for OS and NRM (HR, 1.01 [95% CI, 1.00 to 1.02]; P = .028, and HR, 1.01 [95% CI, 1.01 to 1.02]; $p < .001$, respectively). Lack of complete remission after pretransplantation treatment (induction chemotherapy/HMA) showed an independent effect on relapse (HR, 1.78 [95% CI, 1.32 to 2.41]; $p < .001$). Patients receiving standard conditioning regimens showed a reduced probability of relapse (HR, 0.63 [95% CI, 0.49 to 0.82]; $p < .001$). With respect to donor-recipient HLA match, patients receiving transplant from mismatched unrelated donors showed a significantly reduced OS (HR, 1.2 [95% CI, 1.082 to 1.33]; $p = .012$) and a significantly increased NRM (HR, 1.33 [95% CI, 1.08 to 1.63]; $p = .007$) than those transplanted from a HLA-matched donor.

We then investigated the predictive/prognostic effect of IPSS-M in a cohort of high-risk patients with MDS ineligible for HSCT who received HMA. Inclusion criteria were bone marrow blasts $\geq 10\%$ and availability of clinical and genomic information before starting treatment. 268 patients entered the analysis.

Patients were reclassified according to IPSS-M criteria: 39 patients (15%)



No. at risk:											
IPSS-M	126	82	62	47	31	21	19	16	12	8	4
	108	57	46	36	26	21	17	16	12	7	1
	136	78	63	45	38	28	21	15	12	8	5
	290	168	130	101	76	50	37	29	23	13	5
	304	141	90	63	40	23	17	9	9	6	3

No. at risk:	24 Months	48 Months	72 Months	96 Months	120 Months
IPSS-M low-risk group (n = 122)	(n = 65) Cumulative incidence, 0.092 (95% CI, 0.076 to 0.109)	(n = 31) Cumulative incidence, 0.121 (95% CI, 0.107 to 0.134)	(n = 11) Cumulative incidence, 0.142 (95% CI, 0.125 to 0.16)	(n = 7) Cumulative incidence, 0.17 (95% CI, 0.147 to 0.193)	(n = 8) Cumulative incidence, 0.217 (95% CI, 0.178 to 0.256)
IPSS-M moderate low-risk group (n = 107)	(n = 63) Cumulative incidence, 0.156 (95% CI, 0.138 to 0.173)	(n = 20) Cumulative incidence, 0.193 (95% CI, 0.187 to 0.199)	(n = 9) Cumulative incidence, 0.212 (95% CI, 0.207 to 0.217)	(n = 5) Cumulative incidence, 0.212 (95% CI, 0.212 to 0.212)	(n = 10) Cumulative incidence, 0.212 (95% CI, 0.212 to 0.212)
IPSS-M moderate low-risk group (n = 131)	(n = 78) Cumulative incidence, 0.101 (95% CI, 0.084 to 0.118)	(n = 22) Cumulative incidence, 0.137 (95% CI, 0.126 to 0.147)	(n = 16) Cumulative incidence, 0.137 (95% CI, 0.137 to 0.137)	(n = 9) Cumulative incidence, 0.156 (95% CI, 0.14 to 0.173)	(n = 6) Cumulative incidence, 0.156 (95% CI, 0.156 to 0.156)
IPSS-M high-risk group (n = 285)	(n = 175) Cumulative incidence, 0.101 (95% CI, 0.086 to 0.117)	(n = 51) Cumulative incidence, 0.135 (95% CI, 0.123 to 0.147)	(n = 29) Cumulative incidence, 0.153 (95% CI, 0.138 to 0.168)	(n = 13) Cumulative incidence, 0.153 (95% CI, 0.153 to 0.153)	(n = 17) Cumulative incidence, 0.153 (95% CI, 0.153 to 0.153)
IPSS-M very high-risk group (n = 301)	(n = 229) Cumulative incidence, 0.263 (95% CI, 0.235 to 0.29)	(n = 42) Cumulative incidence, 0.284 (95% CI, 0.278 to 0.289)	(n = 19) Cumulative incidence, 0.296 (95% CI, 0.29 to 0.302)	(n = 5) Cumulative incidence, 0.296 (95% CI, 0.296 to 0.296)	(n = 6) Cumulative incidence, 0.306 (95% CI, 0.298 to 0.314)

Figure 3.3: Clinical assessment of IPSS-R and IPSS-M in 964 MDS patients from the GenoMed4All cohort who received allogeneic stem-cell transplantation (HSCT). (A) Kaplan-Meier probability estimates of OS and (B) cumulative (continued on following page)

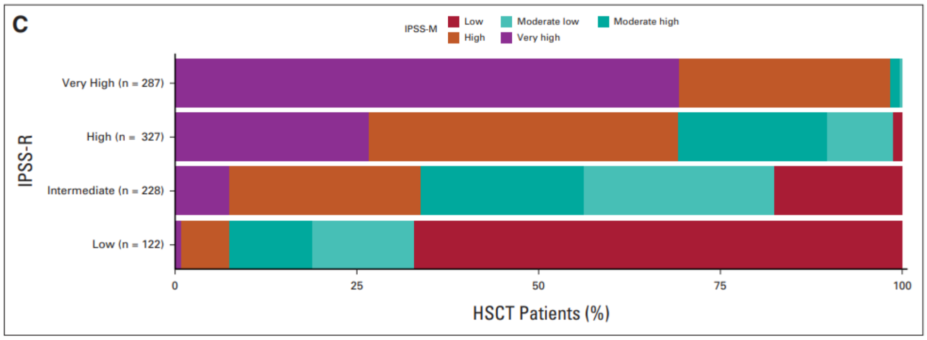


Figure 3.3 (cont.): (Continued). incidence of disease relapse (estimated with a competing risk approach including NRM) for 964 patients from the GenoMed4All cohort who received HSCT, stratified by IPSS-M risk categories. P values are from log-rank test. (C) Restratification of IPSS-R to IPSS-M risk groups in the MDS cohort. Each bar represents an IPSS-R category and shows the percentage of patients that is restratified in the IPSS-M categories (indicated with different colors). HSCT, hematopoietic stem-cell transplantation; IPSS-M, Molecular International Prognostic Scoring System; IPSS-R, Revised International Prognostic Scoring System; LFS, leukemia-free survival; MDS, myelodysplastic syndromes; NRM, nonrelapse mortality; OS, overall survival.

had moderate high, 87 (32%) had high, and 142 (53%) had very high risk. Median duration of MDS before the onset of HMA was 5 months (range, 1-11 months). Patients received HMA for a median of six cycles (range, 1-32 cycles) without significant difference among IPSS-M categories ($p = .41$). The probability of overall response (CR, marrow CR, partial response, and stable disease with hematologic improvement) evaluated after 4-6 cycles of treatment was 42%, without significant difference among IPSS-M categories ($p = .19$).

Median OS in the whole population treated by HMA was 13.9 months. As illustrated in Figure 3.4, the estimated median OS was 20.7 months in moderate high-, 17.9 months in high-, and 12.7 in very high-risk patients ($p < .001$; HR, 1.34 [95% CI, 1.08 to 1.65]; $p = .006$ in a multivariable model adjusted by age and sex).

3.3.4 Accuracy of IPSS-M Prediction when Molecular Information was Missed

We analyzed the loss of accuracy of IPSS-M prediction when one or more IPSS-M–related molecular features are missing.

GenoMed4all and IWG-PM9 populations were used as learning and validation cohorts, respectively. We first evaluated the impact of a missing information from each of the IPSS-M genomic features [44]. Figure 3.5 shows the accuracy loss of IPSS-M prediction for each missing genomic variable. Then, we evaluated the IPSS-M prediction accuracy in the presence of a combination of missing genomic features (starting from missing information on residual genes and then considering the main effect genes ordered by their prognostic weights estimated on probability of LFS; Figure 3.5) [46]. Information on mutational status of a set of 15 genes (ASXL1, CBL, DNMT3A, ETV6, EZH2, FLT3, IDH2, MLLPTD, NPM1, NRAS, RUNX1, SF3B1, SRSF2, TP53$^{\mathrm{multihit}}$, and U2AF1) was required to have an accuracy of IPSS-M prediction of 80% and 70% in the GenoMed4all and IWG-PM cohorts, respectively, while considering a set of 10 and seven genes, the accuracy of IPSS-M prediction decreased to $< 70\%$ versus $< 60\%$, respectively, and to $< 60\%$ versus $< 50\%$, respectively (Figure 3.5).

3.4 Discussion

A more precise risk score is essential to improve precision medicine strategies for patients with MDS. These in turn can identify patient groups that may respond better versus do not benefit from current treatment approaches[37, 38, 39, 40, 7, 41, 42, 43, 44]. In this study, we provided an extensive validation of the recently developed IPSS-M [44] and we confirmed that the molecular score performed better than the conventional IPSS-R. This was also true in patients without detectable mutations, thus suggesting that the statistical model used to develop IPSS-M is more efficient per se in capturing prognostic information with respect to conventional Cox's model [44].

The precise definition of the probability of leukemic evolution is particularly important in the lower-risk groups, which represent the majority of patients with MDS, and in whom new treatment approaches, including HSCT, may be addressed in a refined manner [37, 54]. In this context, our findings confirmed that in the very low-low intermediate IPSS-R risk group, 20% of patients were reclassified into a less favorable prognostic category, $> 90\%$ of which had one or more mutated main effect IPSS-M genes. Thus, the clinical implementation of IPSS-M is expected to result in a more effective selection of candidates to disease-modifying therapies (including HSCT) among patients with early-stage disease [54, 55, 56]. Transplantation performed early after the diagnosis is associated with the most favorable outcome [54], and therefore, patients with higher risk according to IPSS-M should be considered to receive a transplant procedure earlier than the conventional scoring system (IPSS-R) would dictate [57, 58]. We observed in addition that, in patients with MDS treated with HSCT, IPSS-M significantly improved the prediction of the probability of OS with respect to IPSS-R. In particular, IPSS-M was able to efficiently capture the probability of relapse, thus potentially refining the choice of the optimal conditioning regimen at individual patient level [59] (a myeloablative conditioning should be preferable in eligible subjects who are at higher risk of relapse according to genomic features) and improving the identification of patients with high risk of transplantation failure that can be considered for preemptive treatments of disease recurrence [55, 56].

HMA are the only class of drugs approved for the treatment of higher-risk MDS not eligible for HSCT. However, only 40%-50% of patients experience hematologic improvement, and CR occurs in 10%-15% of cases [37, 60]. Effective methods for identifying patients who are most likely to respond to HMA would be of immediate clinical utility. Models on the basis of clinical features are not sufficiently conclusive to deny eligible patients a trial of HMA based on their predictions alone [61, 62].

In our study, IPSS-M failed to stratify individual probability of response; however, response duration and probability of OS were inversely related to IPSS-M risk. This is in line with observation that the IPSS-M is a very good tool to reflect the disease biology and the aggressiveness of MDS subtypes [61, 62, 63]. Additional factors other than gene mutations can be involved in determining sensitivity to HMA [64, 65].

Molecular testing is not yet routine globally because of cost, infrastructure, and reimbursement considerations [37, 44]. We analyzed the accuracy of IPSS-M prediction in both GenoMed4All and IWG-PM cohorts when one or more molecular features are missing. Considering a minimum data set of 15 relevant genes, the accuracy of IPSS-M prediction was 80% and 70%, respectively, while reducing the number of available genes to 10 or less, the accuracy of IPSS-M prediction was significantly lower in both cohorts. These findings may facilitate the clinical implementation of the score into a real-world clinical setting and may help clinicians to define the robustness of the prognosis prediction according to the amount of available information.

Our study may present some limitations, mainly because of the retrospective nature of the data. However, we were able to analyze a large population of patients with MDS, and the collection of DNA for genomic screening was provided independently from disease diagnosis, risk category, and treatment, thus limiting the risk of a selection bias effect and improving the generalizability of the results.

Despite the improved prognostication provided by IPSS-M, we observed that demographic features have a high predictive prognostic power, and clinical parameters (bone marrow blasts and anemia) still retain a strong predictive effect on survival, suggesting that these variables reflect important features of the disease state that are not captured by genomic landscape.4 Accordingly, including sex and age information and combining gene mutation with gene expression data33 might further improve outcome prediction in MDS in next future.

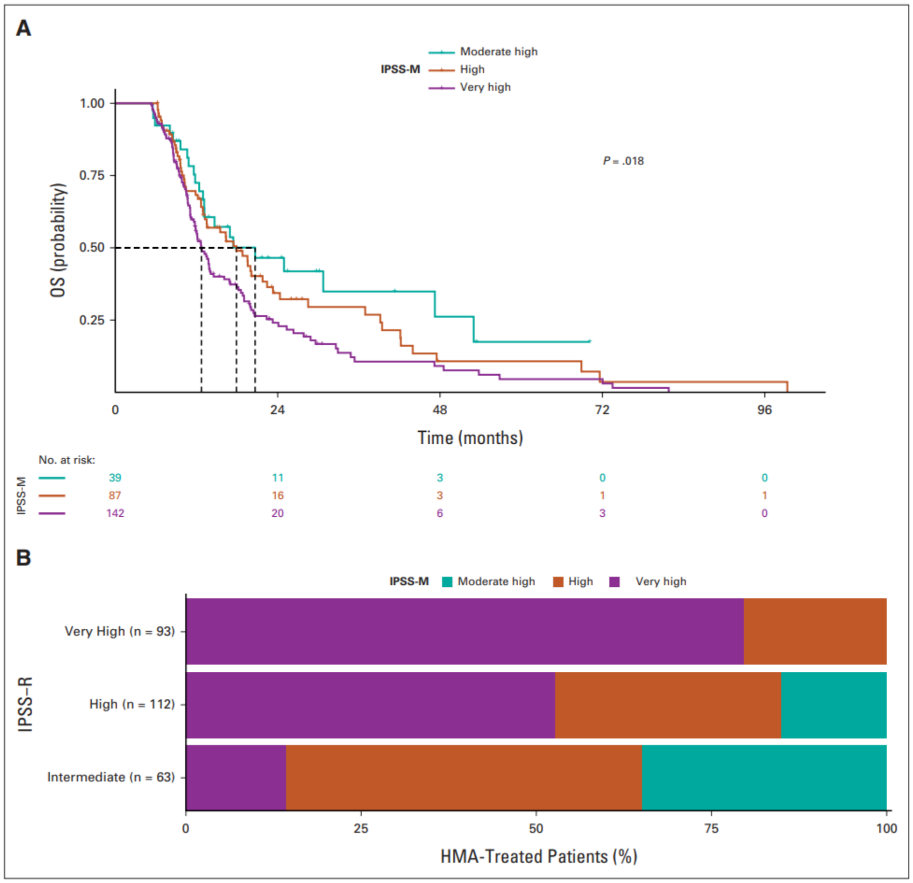


Figure 3.4: Clinical assessment of IPSS-R and IPSS-M in patients with MDS from the GenoMed4All cohort who received HMA. (A) Kaplan-Meier probability estimates of OS of patients with MDS from the GenoMed4All cohort who received HMA (n 5 268) stratified by IPSS-M risk categories. P values are from log-rank test. (B) Restratification of IPSS-R to IPSS-M risk groups in the HMA-treated MDS patients. Each bar represents an IPSS-R category and shows the percentage of patients that is restratified in the IPSS-M categories (indicated with different colors). HMA, hypomethylating agents; IPSS-M, Molecular International Prognostic Scoring System; IPSS-R, Revised International Prognostic Scoring System; LFS, leukemia-free survival; MDS, myelodysplastic syndromes; OS, overall survival.

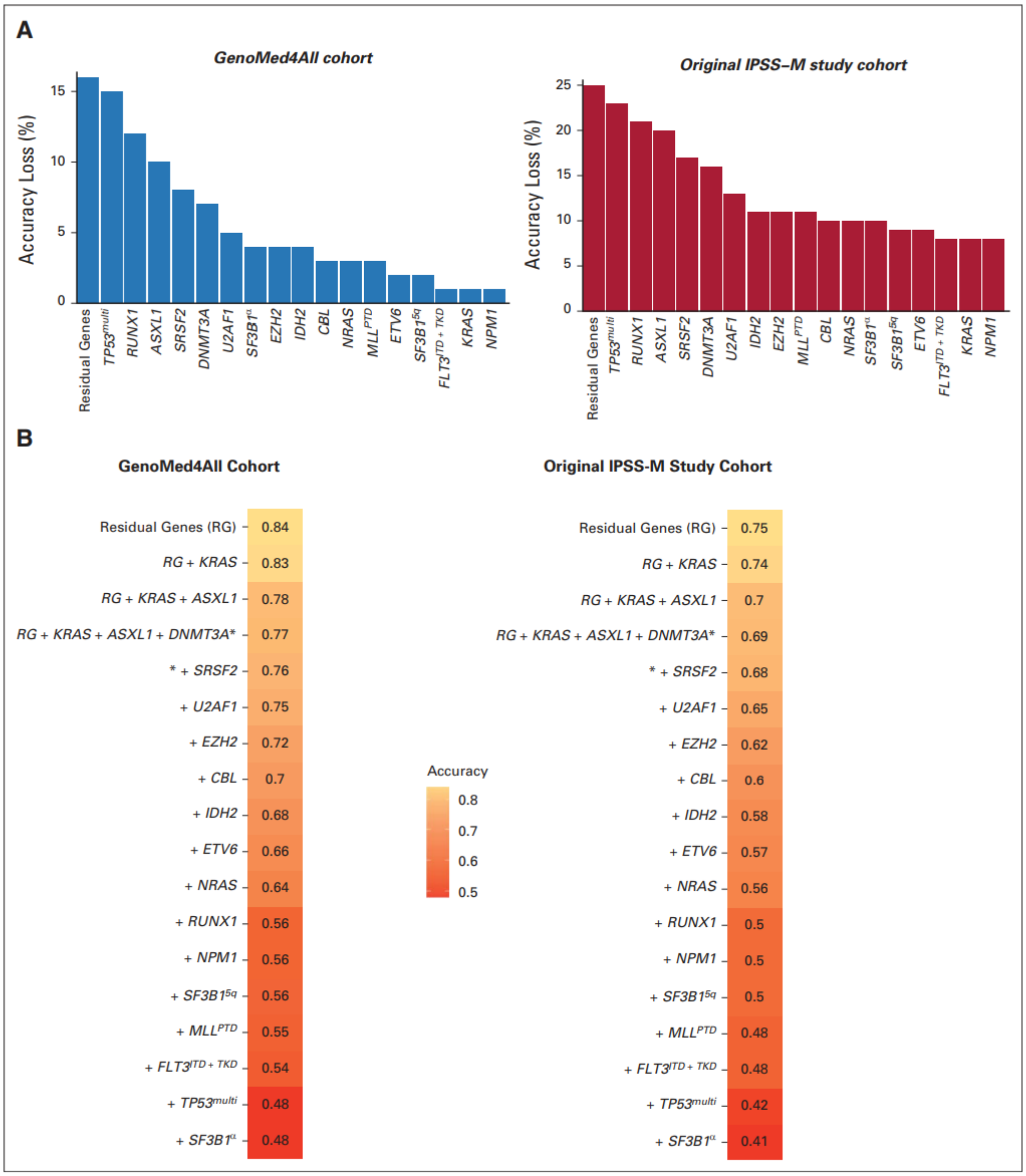


Figure 3.5: Accuracy of IPSS-M prediction when molecular information was missed. (A) Impact of a missing information from each of the IPSS-M genomic feature on the accuracy of IPSS-M risk prediction in GenoMed4all (n = 2,876) and IWG-PM cohorts (n = 2,957). The height of the bar is proportional to the accuracy loss in IPSS-M prediction in the presence of a missing genomic feature, estimated as the percentage of patients classified in the wrong risk category. (B) Prediction of IPSS-M accuracy in the presence of a combination of missing genomic features in the GenoMed4all and IWG-PM cohorts (starting from missing information on residual genes and then considering the main effect genes ordered by their prognostic weights estimated on probability of LFS). The accuracy was estimated as the number of patients classified in the correct risk category divided by the patient population's size. IPSS-M, Molecular International Prognostic Scoring System; IPSS-R, Revised International Prognostic Scoring System; IWG-PM, International Working Group for Prognosis in MDS; LFS, leukemia-free survival.

Chapter 4

Artificial Intelligence for Precision Medicine Prognostic Modeling in Hematology

In preceding chapters, we established the potential of omics and AI in hematology. We also studied existing prognostic hematology models that do not leverage these novel fields. In this chapter, we address the third research objective, namely: *Develop novel precision medicine artificial intelligence models for prognosis in hematological diseases.* In particular, we develop computational approaches to define genotype-phenotype correlations in MDS and to measure combined prognostic information of gene mutations and clinical variables. By comparing these methods to (age-adjusted) IPSS-R risk groups, we obtain a nuanced understanding of omics and AI in hematology. This study was previously published as [40].

4.1 Introduction

Myelodysplastic syndromes (MDS) are heterogeneous clonal hematopoietic disorders characterized by peripheral blood cytopenia and increased risk of evolution into acute myeloid leukemia (AML) [66]. Current disease classification provided by WHO mainly uses morphological features to define MDS categories, leading to a clinical overlap between subtypes and to low interobserver reproducibility in the evaluation of marrow dysplasia [46, 67, 68].

MDS range from indolent conditions to cases rapidly progressing into AML [37, 69]. Disease-related risk is assessed by International Prognostic Scoring System (IPSS) on the basis of percentage of bone marrow blasts, number of peripheral blood cytopenias, and presence of specific clonal cytogenetic abnormalities [38]. In 2012, a revised version of IPSS (IPSS-R) was proposed by introducing five cytogenetic risk groups together with refined categories for bone marrow blasts and cytopenias [38]. Although IPSS and IPSS-R are excellent tools for clinical decision making, these scoring systems have their own weaknesses and may fail to capture reliable prognostic information at individual patient level. In particular, cytogenetics (which is the only biological parameter included in these scores) is not informative in a large proportion of patients and chromosomal abnormalities mostly refer to secondary, late genomic events occurring in the natural history of the disease [39].

The development of MDS is driven by mutations on genes involved in RNA splicing, DNA methylation, chromatin modification, transcriptional regulation, and signal transduction [7, 70, 71, 41]. Chromosomal abnormalities also contribute to MDS pathophysiology.13 Despite recent progress in understanding the disease biology, MDS with isolated 5q deletion is the only category defined by a specific genomic abnormality in the WHO classification2 and only few genotype-phenotype associations have been reported until now, mainly referring to the close relationship between mutations in SF3B1 gene and MDS subtypes with ring sideroblasts [46, 7, 70, 71, 41]

In myeloid malignancies, classifications on the basis of clinical and morphological criteria are being complemented by introducing genomic features that are closer to the disease biology and better capture clinical-pathological entities [46, 72, 52, 73].2,14-16 In this study, we aim to define a new genomic classification of MDS and to improve individual prognostic assessment moving from systems on the basis of clinical parameters to models including genomic information.

4.2 Methods

4.2.1 Study Populations

The Humanitas Research Hospital Ethics Committee approved the study. Written informed consent was obtained from each participant. The study was conducted by EuroMDS consortium (ClinicalTrials.gov identifier: NCT04174547).

We analyzed an international retrospective cohort of 2,043 patients affected with primary MDS according to 2016 WHO criteria [46] and an independent cohort of 318 patients prospectively diagnosed at Humanitas Research Hospital, Milan, Italy (see Data Supplement 1 of [40]).

4.2.2 Genomic Screening

At diagnosis, cytogenetic analysis was performed using standard G-banding and karyotypes were classified using the International System for Cytogenetic Nomenclature Criteria. Mutation screening of 47 genes related to myeloid neoplasms was performed on DNA from peripheral blood granulocytes or bone marrow mononuclear cells (Data Supplement 1 of [40]).

4.2.3 Statistical Methods

Detailed methods are reported in the Data Supplement 1 of [40]. Bradley-Terry models are used to estimate timing of mutation acquisition and to assess the prognostic value of clonal versus subclonal mutations [41].

Bayesian network analysis and hierarchical Dirichlet processes are used to identify genomic associations and subgroups as a basis to define a molecular classification of MDS [52, 72, 73]. Bayesian networks allow to infer the structure of conditional dependencies among mutations, that is, how the presence of a given mutation influences the probability of the others (causality). Dirichlet processes are applied to define clusters capturing broad dependencies among all gene mutations and cytogenetic abnormalities [41, 52, 72, 73]. Patients are clustered based on genomic components identified by Dirichlet processes. Multivariate logistic regression analysis is applied to compare clinical and hematological characteristics among different groups. Survival analyses are performed with Kaplan-Meier method, and differences between groups are evaluated by log-rank test. To carry out the analysis, R package available online [74] is used.

Random-effects Cox proportional hazards multistate modeling was used for developing innovative prognostic tools including clinical parameters and genomics [75, 76]. With the aim to help clinicians to be familiar with such a next-generation prognostic tool, we have created a prototype Web portal that allows outcome predictions to be generated based on this data set for user-defined constellations of genomic features and clinical variables.

All the analyses were carried out on EuroMDS cohort. The Humanitas cohort was used to independently validate models for patient prognostication.

4.3 Results

4.3.1 Genomic Landscape in Myelodysplastic Syndromes

For more detailed results, figures, and supplementary materials, we refer to the sata supplement and appendix of [40] throughout. We studied 2,043 patients with MDS from EuroMDS consortium (Data Supplement 1). Normal karyotype is reported in 1,195 patients (59%), whereas 651 (32%) showed chromosomal abnormalities. Mutations are identified in 45 of 47 genes. A total of 1,630 patients (80%) present one or more mutations (median, 2; range, 1-17). Only six genes are mutated in >10% of patients, with five additional genes mutated in 5%-10%, and 36 mutated in <5% of patients. This is visualized in Figure 4.1.

4.3.2 Mutation Acquisition Order and Prognostic Value of Clonal Versus Subclonal Mutations

By using Bradley-Terry modeling, we calculate a global ranking of MDS genes reflecting how early in disease natural history they are mutated. Mutations in genes involved in RNA splicing and DNA methylation occur early, whereas mutations in genes involved in chromatin modification and signaling often occur later.

A total of 14 genes are associated with worse prognosis if mutated, whereas one gene (SF3B1) is associated with better outcome. Variant allele fractions are used to estimate the proportion of tumor cells carrying a given mutation and identify clonal or subclonal mutations. Accordingly, 58% of patients show only clonal mutations, whereas 42% have evidence for both clonal and subclonal mutations. No significant differences in survival between clonal and subclonal mutations for the majority of the investigated genes are observed, highlighting the importance of including information on subclonal mutations in the predictive model.

4.3.3 Identification of Genomic Associations and Subgroups in Myelodysplastic Syndromes

Pairwise associations among genes and cytogenetic abnormalities reveal a complex landscape of positive and negative associations. Bayesian networks are applied to define in a more comprehensive way the relationships between genomic abnormalities. Accordingly, mutations of splicing genes are mutually exclusive. SF3B1 mutations are mutually exclusive with TP53 mutations, whereas they co-occur with JAK/STAT pathway mutations. SRSF2 mutations co-occur with TET2, ASXL1, CBL, IDH1/2, RUNX1, and STAG2 mutations. U2AF1 mutations co-occur with abnormalities of chromosome 7 and 20 and NRAS mutations. TET2 mutations co-occur with SRSF2 and ZRSR2 mutations. DNMT3A mutations are mutually exclusive with ASXL1 mutations, whereas they co-occur with BCOR, IDH1, and NPM1 mutations. 5q deletion is frequently present as a single genomic abnormality, whereas a co-occurrence with TP53 mutations and with several single cytogenetic components of complex karyotype is observed.

4.3.4 Definition of a Genomic Classification of Myelodysplastic Syndromes

Dirichlet processes are used to identify genomic subgroups among MDS. We identify six components, each describing a specific distribution of variables included in the model (ie, cytogenetic abnormalities and gene mutations). Each patient is characterized by a weight vector indicating the contribution of each component to its genome. By performing hierarchical agglomerative clustering, we obtain eight groups (clusters) defined according to specific genomic features.

One group includes patients without specific genomic profiles (ie, without recurrent mutations in the study genes and/or chromosomal abnormalities); strikingly, all the remaining groups are deeply characterized by a single (in some cases two) component of Dirichlet processes. In many groups, dominant genomic features include splicing gene mutations. We identify two groups (1 and 6) in which dominant features are SF3B1 mutations, presence of ring sideroblasts, and transfusion-dependent anemia. Group 6 includes patients with ring sideroblasts and isolated SF3B1 mutations (except for co-mutation patterns including TET2, DNMT3A, and JAK/STAT pathway genes) characterized by isolated anemia, normal or high platelet count, single or multilineage dysplasia, and low percentage of marrow blasts (median, 2%). Group 1 includes patients with SF3B1 with co-existing mutations in other genes (ASXL1 and RUNX1) characterized by anemia associated with mild neutropenia and thrombocytopenia, multilineage dysplasia, and higher marrow blast percentage with respect to group 6 (7% v 2%, P < .0001).

In two groups (3 and 5), dominant genomic features are represented by SRSF2 mutations. In these groups, the most frequently reported chromosomal abnormality is trisomy 8. Group 3 includes patients with SRSF2 and concomitant TET2 mutations. Patients present single cytopenia (anemia in most cases) and higher monocyte absolute count with respect to the other groups (P ¡ .0001). Bone marrow features include multilineage dysplasia and excess blasts (median, 8%). Group 5 is characterized by SRSF2 mutations with co-existing mutations in other genes (ASXL1, RUNX1, IDH2, and EZH2). Patients present two or more cytopenias, multilineage dysplasia, and excess blasts (median, 11%; significantly higher with respect to group 3; P = .0031).

Group 4 dominant features include U2AF1 mutations associated with 20q deletion and chromosome 7 abnormalities (Appendix, Data Supplement 1). Patients present a higher rate of transfusion-dependent anemia with respect to the other groups (P from .023 to ¡ .0001). Marrow features include multilineage dysplasia and excess blasts in most cases.

Group 2 is characterized by TP53 mutations and/or complex karyotype. In most patients, two or more cytopenias (with high rate of transfusion dependency) and excess blasts are present (Appendix, Data Supplement 1).

Group 7 includes patients with AML-like mutation patterns (DNMT3A, NPM1, FLT3, IDH1, and RUNX1 genes). Patients are characterized by two or more cytopenias (with high rate of transfusion dependency) and excess blasts, in most cases (83%) ranging from 15% to 19%.

Finally, group 0 includes MDS without specific genomic profiles. These patients are characterized by younger age, isolated anemia, normal or reduced marrow cellularity (with respect to age-adjusted normal ranges), absence of ring sideroblasts, and low percentage of marrow blasts (median, 2%).

A heterogeneous distribution of 2016 WHO disease subtypes is observed through the new groups defined by genomic features (P < .0001, Appendix). Interestingly, this new classification accounts for genomic heterogeneity of patients stratified according to WHO criteria. This is evident for MDS with isolated 5q deletion. Patients with none or one mutation (mainly including SF3B1 gene) are clustered into group 6, whereas those with two or more mutations or TP53 mutations are classified into group 1 (Appendix). MDS with 5q deletion included into group 6 show lower rate of transfusion dependency and lower percentage of marrow blasts with respect to patients classified into group 1 (P = .0043 and P < .0001).

These findings provide the proof of concept for a new classification of MDS on the basis of entities defined according to specific genomic features. In the Appendix, we define a diagram to classify patients in the appropriate category on the basis of individual genomic profile.

4.3.5 Clinical Relevance of Genomic Classification of Myelodysplastic Syndromes in Predicting Survival and Response to Specific Treatments

Genomic-based MDS groups present different probability of survival (Appendix, P < .0001), suggesting that the integration of genomic features may improve the capability to capture prognostic information. Groups 1 and 6 characterized by SF3B1 mutations show better survival with respect to groups 2, 3, 4, 5, and 7 (P from < .0001 to .0093), isolated SF3B1 (group 6) being associated with better outcome with respect to SF3B1 with co-mutated patterns (group 1, P = .0304). Group 0 including patients without specific genomic abnormalities is associated with good prognosis as well (P from < .0001 to .012 with respect to groups 2, 3, 4, 5, and 7). Groups defined by splicing mutations other than SF3B1 show worse survival; among them, group 5 (SRSF2 mutations with co-existing mutations in other genes) is associated with dismal outcome (P from < .0001 to .0177 with respect to groups 0, 1, 4, and 6). Group 2 including patients with TP53 mutations and complex karyotype shows the poorest outcome (P from < .0001 to .0473). Group 7 including patients with AML-like mutations shows high rate of leukemic evolution and worse prognosis as well (P < .0001 with respect to groups 1, 3, and 6). Finally, among patients with isolated 5q deletion, cases with none or single mutation are associated with a better prognosis with respect to those with two or more mutations or TP53 mutations (P = .0432).

Then, we tested whether grouping MDS patients according to genomic features may provide information about response to specific treatments. We focused on 424 cases who underwent allogeneic transplantation and on 221 cases treated with hypomethylating agents. With the limit to analyze a retrospective cohort of selected patients, MDS groups on the basis of genomic features do not identify different probability of survival after hypomethylating agents (not shown), whereas they are able to significantly stratify post-transplantation outcome. This is visualized in Figure 4.2. SF3B1-related groups (groups 1 and 6), MDS with AML-like mutations (group 7), and MDS without specific genomic abnormalities (group 0) show a better outcome after transplant, whereas groups defined by TP53 mutation and/or complex karyotype (group 2) and by U2AF1 mutations (group 4) are associated with a high rate of transplantation failure (Fig 4.2).

4.3.6 Personalized Prognostic Assessment on the Basis of Clinical and Genomic Features

Random-effects Cox multistate model incorporating 63 clinical and genomic variables are developed to estimate personalized probability of survival.

First, we determined the fraction of explained variation for clinical outcome that was attributable to different prognostic factors. This is visualized in Figure 4.3. Demographic features (age and sex) have a high predictive prognostic power. Gene mutations and co-mutation patterns increase the prognostic power of cytogenetics. Clinical features (percentage of marrow blasts and anemia) still retain a strong independent predictive power for survival, suggesting that these variables reflect important features of the disease state that are not captured by genomic landscape (Fig 4.3).

We then explored whether Random-effects Cox multistate model could generate accurate survival predictions for individual patients and if the obtained predictions are more informative than conventional age-adjusted IPSS-R.

Random-effects Cox multistate model is able to generate a prediction for survival that correlated well with the observed outcomes in EuroMDS cohort (Tables 4.1 and 4.2). Internal cross-validation shows a concordance of 0.74 and 0.71 for survival in training (67% of patients) and test (33% of patients) subsets, respectively. This model shows superior performance to conventional scoring systems (age-adjusted IPSS-R concordance is 0.62 and 0.65 in training and test subsets of EuroMDS cohort, respectively). Interestingly, the concordance of Dirichlet process components is similar to that of age-adjusted IPSS-R (0.65 and 0.62, respectively), thus underlying the relevance of accounting for genomic features into the prognostic model.

	Training (66% of EuroMDS Patients)		Test (33% of EuroMDS Patients)	
Statistical Model and Variable Selection	Concordance	SD	Concordance	SD
Cytogenetics IPSS-R risk groups	0.576	0.012	0.567	0.016
Age-adjusted IPSS-R risk groups	0.620	0.015	0.659	0.019
Dirichlet processes	0.649	0.014	0.629	0.020
CoxRFX_Clinical, demographics, Dirichlet	0.729	0.015	0.713	0.021
CoxRFX_Clinical, demographics, genomics	0.742	0.015	0.709	0.021

Table 4.1: Concordance Comparison Between Random-Effects Cox Proportional Hazards Multistate Models (CoxRFX) and IPSS-R on Training-Test Approach.

In Figure 4.4, we illustrate an example of the calculations to obtain a personalized prediction of survival by using patients from EuroMDS cohort; in two patients with same clinical phenotype and similar predicted prognosis according

	Training (66% of EuroMDS Patients)		Test (33% of EuroMDS Patients)	
Statistical Model and Variable Selection	Concordance	SD	Concordance	SD
CoxRFX_Clinical, demographics, Dirichlet	0.715	0.012	NA	NA
CoxRFX_Clinical, demographics, genomics	0.737	0.012	0.753	0.037

Table 4.2: Concordance of CoxRFX Models and Age-Adjusted IPSS-R on Training-Validation Approach

to age-adjusted IPSS-R, Random-effects Cox multistate model is able to capture additional prognostic information and efficiently predicts clinical outcome.

Because the underlying survival model is complex, specific information technology support is needed to combine all the information at individual patient level and to translate it into a personalized outcome prediction. With the aim to help clinicians to be familiar with such a next-generation prognostic tool, we have created a prototype Web portal [77] that allows outcome predictions to be generated based on EuroMDS data set for user-defined constellations of genomic features and clinical variables.

4.3.7 Independent Validation of Personalized Prognostic Assessment

An independent validation of Random-effects Cox multistate model is performed on Humanitas cohort (a single-center prospective population of 318 patients showing significantly different hematological features with respect to EuroMDS cohort). Concordance for survival in Humanitas cohort was similar to that observed in EuroMDS cohort (0.75 and 0.74, respectively), suggesting that the model provides considerable discriminatory power that accurately generalizes to other real-world populations (Tables 4.1 and 4.2).

4.4 Discussion

We developed computational approaches to define genotype-phenotype correlations in MDS and to measure combined prognostic information of gene mutations and clinical variables.

RNA splicing is the most commonly mutated pathway in MDS [70, 71, 41] and occurs early in disease evolution. These mutations play a major role in determining the disease phenotype, with differences in morphological features and survival [41]. Splicing mutations may also influence the subsequent genomic evolution of the disease because the patterns of cooperating mutations are different between SF3B1, SRSF2, and U2AF1 genes [41, 78]. Overall, these findings suggest that a genomic classification in MDS is advisable.

We identify eight subgroups of MDS based on specific genomic features. WHO subtypes are heterogeneously distributed across these new genomic categories, suggesting that the current classification is unable to capture distinct MDS biological features.

SF3B1 mutations define a specific MDS subtype characterized by ring sideroblasts, low blast count, and favorable outcome [46, 41, 70, 71, 79]. Among SF3B1-mutated patients, JAK/STAT pathway coexisting mutations can induce the acquisition of a myeloproliferative phenotype [80]. A distinct disease subtype includes patients with SF3B1 mutations and co-existing mutations in other genes (RUNX1 and ASXL1), characterized by multilineage dysplasia [79]. This disease subgroup is associated with poorer outcome. SRSF2 and U2AF1 mutations identify distinct disease subtypes with specific co-mutation patterns, hematological phenotype, and reduced probability of survival with respect to SF3B1-defined categories [81, 82, 83, 84, 85].

The subgroup with TP53 mutations and complex karyotype has very poor outcomes [86]; this same subgroup has been identified in AML and myeloproliferative neoplasms [52, 72, 73]. We identify an MDS subtype including cases with mutations that are recurrently described in de novo AML [46, 72]; this category shows a very high risk of leukemic transformation and poor outcome, suggesting that the current threshold of 20% marrow blasts might be not suitable to recognize different disease entities from a biological point of view. Moreover, we notice a high percentage of patients with marrow hypocellularity in the group without specific genomic features; these MDS show overlapping clinical features with aplastic anemia [46, 87]. Overall, these findings suggest that a genomic classification could transcend the boundaries of MDS and help categorization of cases bordering with other myeloid conditions where current morphological criteria are often inadequate.

Moving to prognostication, we have built statistical models that can generate personally tailored survival prediction using information from both clinical and genomic features [52]. We show that the inclusion of gene mutations and comutational patterns significantly improves patient prognostication with respect to IPSS-R, which considers only cytogenetics abnormalities. Although conventional prognostic systems provide an outcome prediction based on the median survival of patients with similar clinical features, our new prognostic model is based on individual patient genotype and phenotype, thus improving the capability of capturing prognostic information in such a heterogeneous disease. Finally, genomic features are relevant for predicting survival after transplantation, supporting the rationale to include this information to support transplantation decision making in MDS [55, 56].

The most critical issue for this novel prognostic model is sample size, which is particularly relevant in MDS showing a long tail of genes mutated in a low proportion of cases. According to previous data, for a gene mutated in 5%-10% of patients, a training set of 500-1,000 patients would suffice, but for a gene mutated in $< 1\%$ of patients, a cohort of $> 5{,}000$ would be needed [52]. Additional cooperative efforts are therefore needed to improve the reliability and generalizability of these models.

The integration of clinical data with diagnostic genome profiling in MDS may provide prognostic predictions that are personally tailored to individual patients. Such information will empower the clinician and support complex decision-making process in these patients.

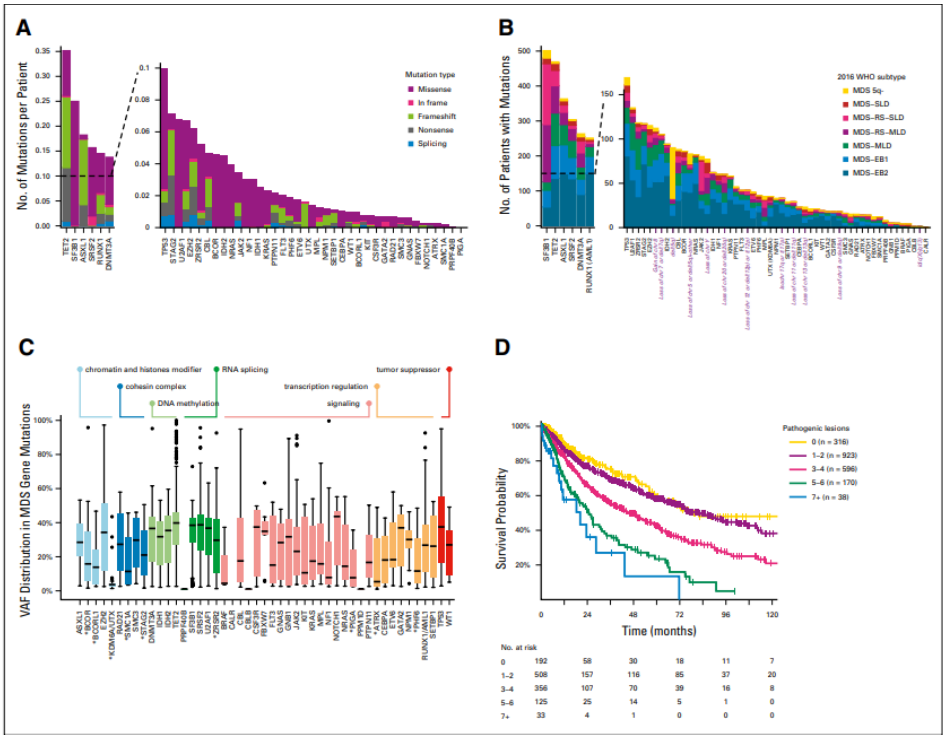
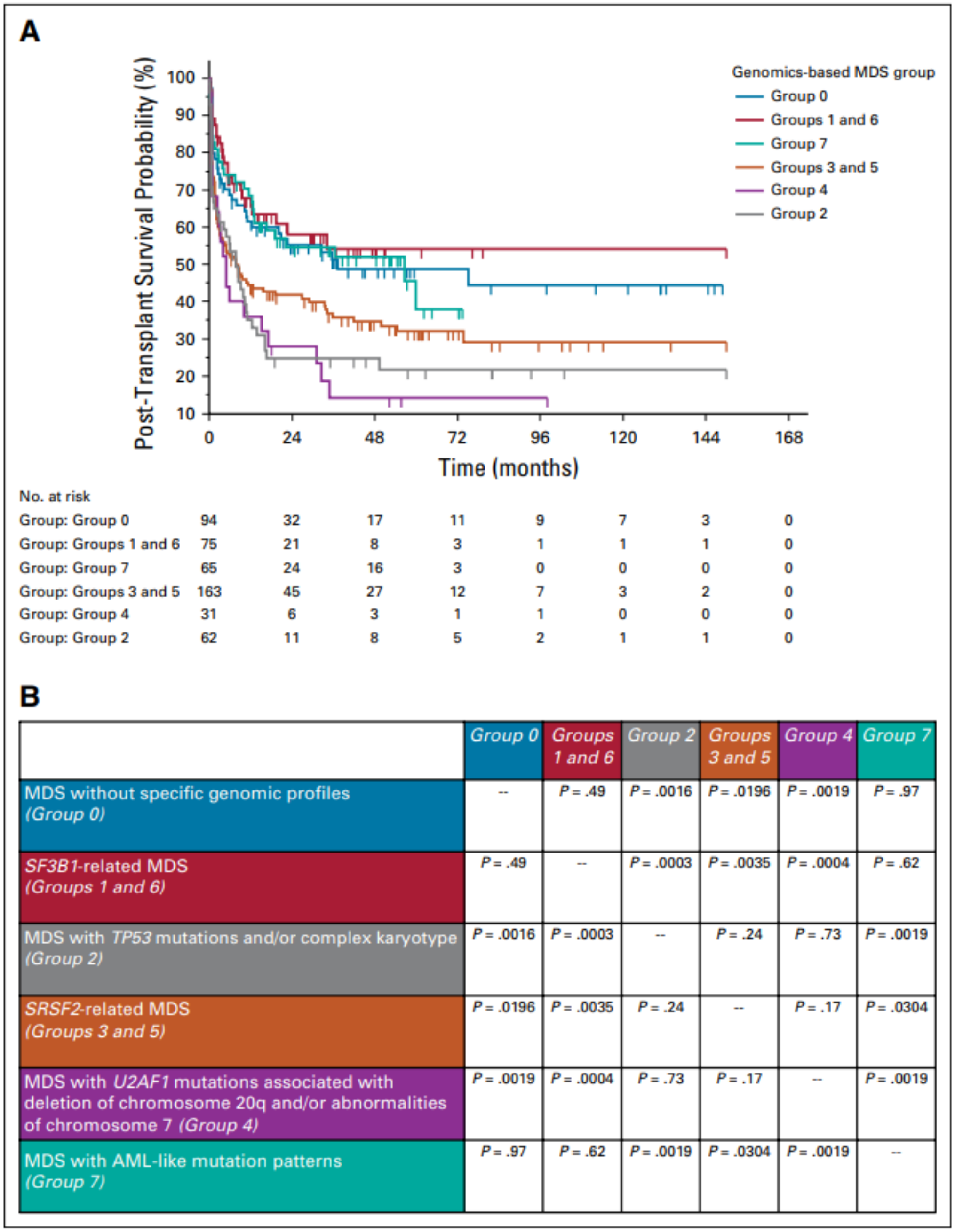


Figure 4.1: (A) Frequency of mutations and chromosomal abnormalities in the EuroMDS cohort (n = 2,043), stratified according to the type of mutation (missense, nonsense, affecting a splice site, or other). Insertions and deletions (del) were categorized according to whether they resulted in a shift in the codon reading frame (by either 1 or 2 base pairs [bp]) or were in frame. Splicing factor genes were the most frequently mutated (49%), followed by DNA methylation–related genes (37.9%), chromatin and histone modifier genes (31.3%), signaling genes (28.5%), transcription regulation genes (24%), tumor suppressor genes (11.1%), and cohesin complex genes (7.6%). (B) Frequency of recurrently mutated genes and chromosomal abnormalities in the EuroMDS cohort, broken down by MDS subtype according to 2016 WHO criteria. (C) VAF of driver mutations in the EuroMDS cohort, broken down by gene and gene function (boxplots reporting median, 25-75 percentiles, and ranges); VAF of X-linked genes (ATRX, BCOR, BCORL1, PHF6, PIGA, SMC1A, STAG2, UTX, and ZRSR2, highlighted by asterisk in the figure plot) was halved in male patients. (D) Relationship between the number of genomic abnormalities (mutations and chromosomal abnormalities) and outcome (overall survival). MDS, myelodysplastic syndromes; MDS 5q-, MDS with isolated deletion of long arm of chromosome 5; MDS-EB1, MDS with excess of blasts, type 1; MDS-EB2, MDS with excess of blasts, type 2; MDS-MLD, MDS with multilineage dysplasia; MDS-RS-MLD, MDS with ring sideroblasts and multilineage dysplasia; MDS-RS-SLD, MDS with ring sideroblasts and single-lineage dysplasia; MDS-SLD, MDS with single-lineage dysplasia; VAF, variant allele frequencies.



No. at risk

	0	24	48	72	96	120	144	168
Group: Group 0	94	32	17	11	9	7	3	0
Group: Groups 1 and 6	75	21	8	3	1	1	1	0
Group: Group 7	65	24	16	3	0	0	0	0
Group: Groups 3 and 5	163	45	27	12	7	3	2	0
Group: Group 4	31	6	3	1	1	0	0	0
Group: Group 2	62	11	8	5	2	1	1	0

	Group 0	*Groups 1 and 6*	*Group 2*	*Groups 3 and 5*	*Group 4*	*Group 7*
MDS without specific genomic profiles *(Group 0)*	--	*P* = .49	*P* = .0016	*P* = .0196	*P* = .0019	*P* = .97
SF3B1-related MDS *(Groups 1 and 6)*	*P* = .49	--	*P* = .0003	*P* = .0035	*P* = .0004	*P* = .62
MDS with *TP53* mutations and/or complex karyotype *(Group 2)*	*P* = .0016	*P* = .0003	--	*P* = .24	*P* = .73	*P* = .0019
SRSF2-related MDS *(Groups 3 and 5)*	*P* = .0196	*P* = .0035	*P* = .24	--	*P* = .17	*P* = .0304
MDS with *U2AF1* mutations associated with deletion of chromosome 20q and/or abnormalities of chromosome 7 *(Group 4)*	*P* = .0019	*P* = .0004	*P* = .73	*P* = .17	--	*P* = .0019
MDS with AML-like mutation patterns *(Group 7)*	*P* = .97	*P* = .62	*P* = .0019	*P* = .0304	*P* = .0019	--

Figure 4.2: (A) Probability of overall survival after allogeneic transplantation in the EuroMDS cohort. Patients were stratified according to specific genomic features. A total of 424 cases with complete information about transplant procedures and clinical outcome entered the analysis. (B) Comparison of probability of survival among different genomic-based MDS groups (P values of log-rank test were reported). AML, acute myeloid leukemia; MDS, myelodysplastic syndromes.

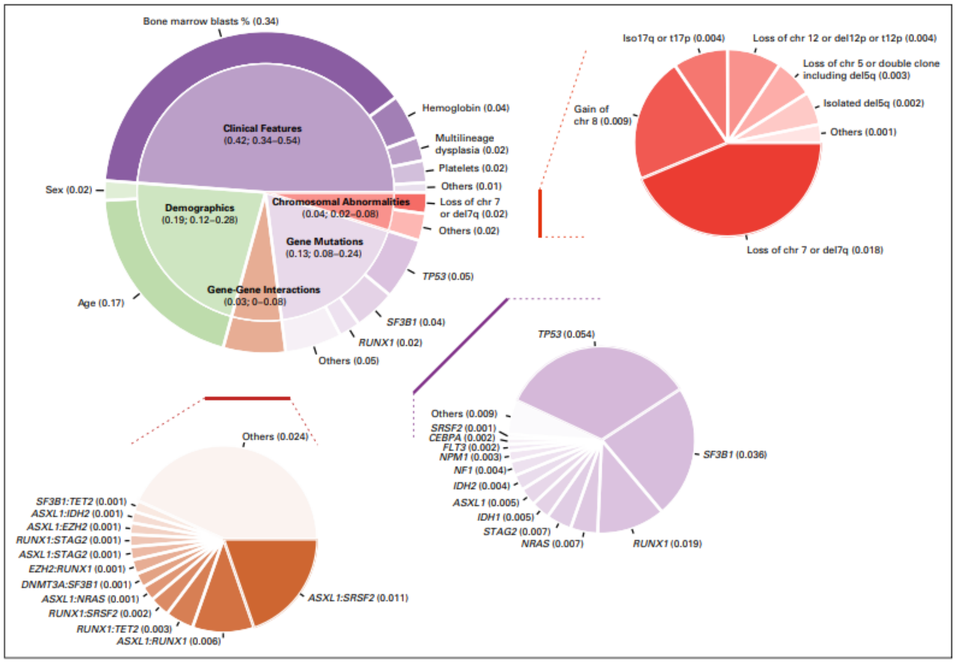


Figure 4.3: (A) Probability of overall survival after allogeneic transplantation in the EuroMDS cohort. Patients were stratified according to specific genomic features. A total of 424 cases with complete information about transplant procedures and clinical outcome entered the analysis. (B) Comparison of probability of survival among different genomic-based MDS groups (P values of log-rank test were reported). AML, acute myeloid leukemia; MDS, myelodysplastic syndromes.

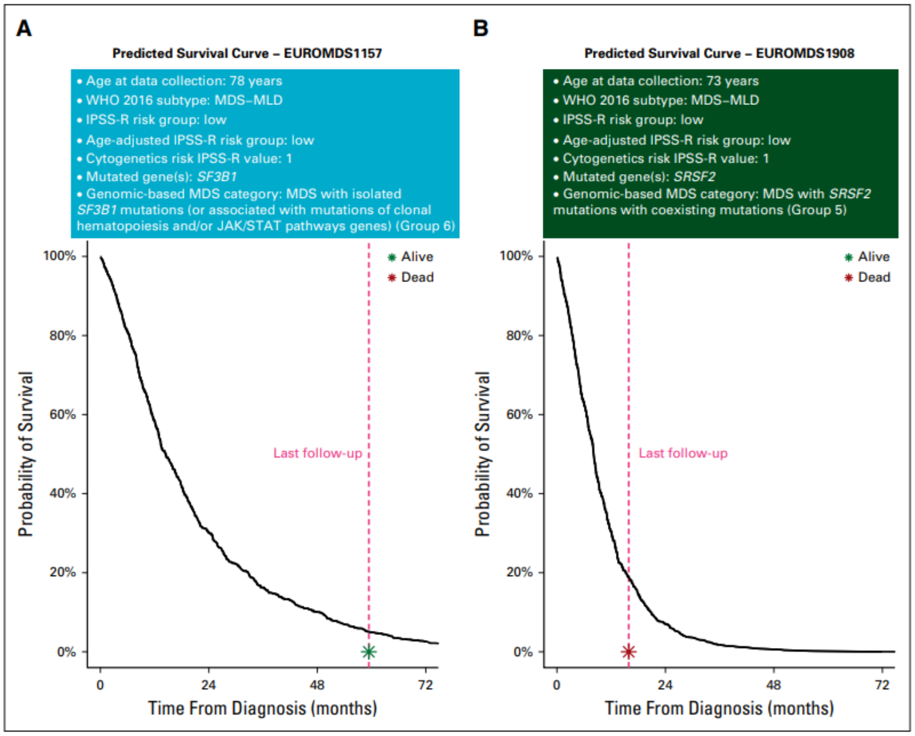


Figure 4.4: Personalized prediction of overall survival using a multistate prognostic model including clinical and genomic features and their interactions in two patients from the EuroMDS cohort (labeled as patient A and patient B), both classified as MDS with multilineage dysplasia according to 2016 WHO classification and belonging to low-risk group according to age-adjusted revised version of International Prognostic Scoring System (IPSS-R). Using currently available prognostication, both patients are predicted to have an indolent clinical course without significant risk of disease evolution and death (in the EuroMDS cohort, Kaplan-Meier curves show a median survival of 79 months for low-risk age-adjusted IPSS-R). When looking at mutational profile, driver mutations involved different splicing factor genes in these patients: patient A carries SF3B1 mutation, whereas patient B presents SRSF2 mutation. We then calculated expected survival by using the novel genomic-based prognostic model (exponential survival curves are reported in the figure). Patient A was classified into genomic-based group 6, and patient B was classified into group 5. Accordingly, the estimation of life expectancy is now significantly different in these two patients, as underlined by the slope of the two exponential curves. The model predicts a better probability of survival for patient A (with SF3B1 mutation) with respect to patient B (with SRSF2 mutation), thus reflecting more precisely the observed clinical outcome. In fact, patient B died 16 months after the diagnosis as a result of leukemic evolution, whereas patient A was still alive without evidence of disease progression after 60 months of follow-up. IPSS-R fails to capture such a difference in clinical outcome. The interpretation of the predicted survival curves by genomic-based predictive model is meaningful also considering that we are in the context of a cohort of elderly patients: patient A (age 78 years) has a 30% survival probability at the age of 80, whereas patient B (age 73 years) has a 30% survival probability at the age of 74.

Chapter 5

Privacy Enhancement through Synthetic Data in Hematology Research

In the preceding chapters, we have established that omics data and AI have a considerable potential in hematology model development. What remains is to show that this potential can be realized in a manner that respects individuals' privacy. In this chapter, we study ML-based *synthetic data*, a new paradigm in data protection. In doing so, we address research question four: *Show that hematological disease models can be ethically and securely trained and integrated for analysis and application, without breaching patients' rights and trust.* We find that synthetic data mimic real clinical-genomic features and outcomes, and anonymize patient information. The implementation of this technology allows to increase the scientific use and value of real data, thus accelerating precision medicine in hematology and the conduction of clinical trials. This study was previously published as [88].

5.1 Introduction

Personalized medicine combines established clinical-pathologic parameters with advanced genomic profiling to develop innovative diagnostic, prognostic, and therapeutic strategies [89]. Hematology has been rapidly transformed by genome characterization and is the forefront to reap the benefits of personalized medicine for patient management [89].

The clinical implementation of personalized medicine requires the availability of a great amount of real-world data, including clinical features, genomics, treatments, and outcomes [90, 91, 92]. Collecting such information in large patient populations is challenging, especially when facing rare diseases with heterogeneous clinical/molecular background. Additionally, real data often have imbalances or lack/incomplete information [93, 94]. Finally, there are many issues concerning patient privacy that may prevent use of data outside specific contexts and that are to be accounted for [95].

One approach that can circumvent these issues is the creation of synthetic data. Synthetic data are artificial data generated by a model trained to learn the essential characteristics of a real source data set [96, 97]. Synthetic data building techniques attempt to ensure that the generated data are neither a copy nor a representation of the real data, setting the grounds to data sharing without violating the current legislation on privacy [96, 97]. Moreover, synthetic data allow to increase insufficient information obtained from real patients by data augmentation and data integration, thus potentially solving issues related with small sample size and clinical/molecular class imbalance [98].

Overall, synthetic data may overcome many of the pitfalls of real data, allowing for faster, less expensive, and more scalable access to information that is representative of the underlying source and privacy-preserving.8-11 Synthetic data is a growing technology [96] and it is expected that in the next 2-3 years, >60% of the data used in research and development process across different domains (including life sciences) will be synthetically generated [99].

In this project, we addressed the issue of clinical validation and research utility of synthetic data in hematology. To this purpose, we aimed to (1) apply innovative synthetic data generation methods to real-world data sets of different hematologic malignancies including comprehensive clinical and genomic information; (2) develop a synthetic validation framework (SVF) to evaluate data fidelity and privacy preservability; and (3) test the capability of synthetic data to accelerate translational and clinical research.

As a paradigmatic use case, we focused on myeloid malignancies, which are rare neoplasms with high clinical heterogeneity and complex genomic background and that include patients with unmet clinical needs [48].

5.2 Methods

5.2.1 Study Populations

The study was conducted by GenoMed4All and Synthema European consortia and supported by EuroBloodNET, the European Reference Network on rare hematologic diseases. Written informed consent was obtained from each participant. The Humanitas Ethics Committee approved the study. This study was registered at ClinicalTrials.gov (identifier: NCT04889729).

All the study procedures were compliant with the 2021 WHO guidance on ethics and governance of artificial intelligence for health [100].

Inclusion criteria were age ≥18 years, a diagnosis of myeloid neoplasm (either myelodysplastic syndromes [MDS] or AML) according to WHO 2016 criteria [46], and information available on demographics, clinical features, mutational screening/chromosomal abnormalities, treatment, and survival. Overall, 7,133 patients were included.

5.2.2 Generative Model for Synthetic Data

Artificial intelligence (AI)–based generative models are characterized by multilayer neural networks that are able to generate samples (patients) by learning the distribution of a set of real data [101].16 In this context, generative adversarial networks (GANs) [102] create simulation scenarios where models and processes interact to create completely new data sets of events. GANs consist of two networks: the generator and the discriminator. These two networks are trained adversarially. The generator creates artificial outputs that are passed to the discriminator along with real data, while the discriminator is tasked to identify which outputs were real and which were fake. The final goal here is to reach equilibrium, in which the generated samples follow the same distribution as the real data. When this happens, the discriminator can do no better than random guessing.16 Conditional GANs are variants of GANs where a label is added as a parameter to the input of the models to create more realistic data by learning specific correlations [103]. In this study, we implemented a conditional Wasserstein's tabular GAN18 with gradient penalty [104] that ensures high performance in modeling large data sets with complex distribution and interactions among different features. We adopted different preprocessing steps and training strategies to properly prepare the input data and optimize the training steps.

5.2.3 Development of a Synthetic Validation Framework

A SVF was developed to evaluate fidelity and privacy preservability of the newly generated synthetic data.

We assessed the quality of the following data types: demographics, clinical features, genomics (evaluated as categorical variables), and clinical outcomes (probability of overall survival and leukemia-free survival). Distribution, correlation, and principal component analysis evaluation were then assessed on all data types. Descriptive statistics and pairwise association analyses were carried out. We calculated a clinical synthetic fidelity (CSF) and a genomic synthetic fidelity (GSF) as the average of multiple metric tests adopted; optimal threshold was considered ≥85% in both systems.

Real and synthetic patients were stratified by hierarchical Dirichlet clustering [105] to identify genomic associations and subgroups. Survival analyses were performed with Kaplan-Meier curves. We implemented Cox proportional hazard and L1-penalized Cox regression models to define features with significant impact on survival probability [38, 105]. Model discrimination was assessed using Harrell's concordance index [106].

To assess the privacy preservability and evaluate the risk associated with synthetic datasets of resampling a patient from a synthetic record, we first measured the exact matches between synthetic and original data (identical match share [IMS]). Moreover, we calculated the distance to closest record that measures the absolute distances between synthetic records to their nearest original records, and we then calculated the nearest neighbor distance ratio (NNDR), that is, the ratio of the distances of each synthetic record to the nearest and to the second nearest neighbors, that allows to compare inliers and outliers in the population on an equal base [107]. Optimal range for NNDR was considered from 0.60 to 0.85 (value closer to 0.50 indicating a significant loss of similarity of the synthetic patients compared with the real ones that can affect the fidelity of synthetic data; value closer to 1.00 indicating an excess of similarity of synthetic data with respect to the real ones, thus possibly affecting the privacy preservability) [107].

Explainability of AI algorithms was assessed by Shapley Additive Explanations (SHAP), a method to explain individual predictions on the basis of the game theoretically optimal Shapley values [108].

5.2.4 Experimental Setup

We tested synthetic data generation process in different experimental settings, summarized in Figure 5.1.

In *setting A*, we investigated the capability of the generative model to create a synthetic reproduction of real data with high grade of fidelity on clinical/genomic features, clinical outcomes, and with high privacy preservability. We used 2043 patients with MDS from GenoMed4All cohort [105] to train and test the model.

In *setting B*, we tested the capability of the model to overcome lack/incomplete information in real data and to allow data augmentation; moreover, we assessed the generalizability of the model's performances across different clinical settings. We considered three different populations: 2,043 MDS from GenoMed4All cohort [105]; 2,957 MDS from the International Working Group for Prognosis in MDS (IWG-PM) cohort [44], and 1,002 AML from GenoMed4All cohort [105]. In all experiments, we calculated fidelity and privacy metrics.

In *setting C*, we investigated if the generation of synthetic data can accelerate translational research. Starting from a MDS cohort available in 2014 (n =

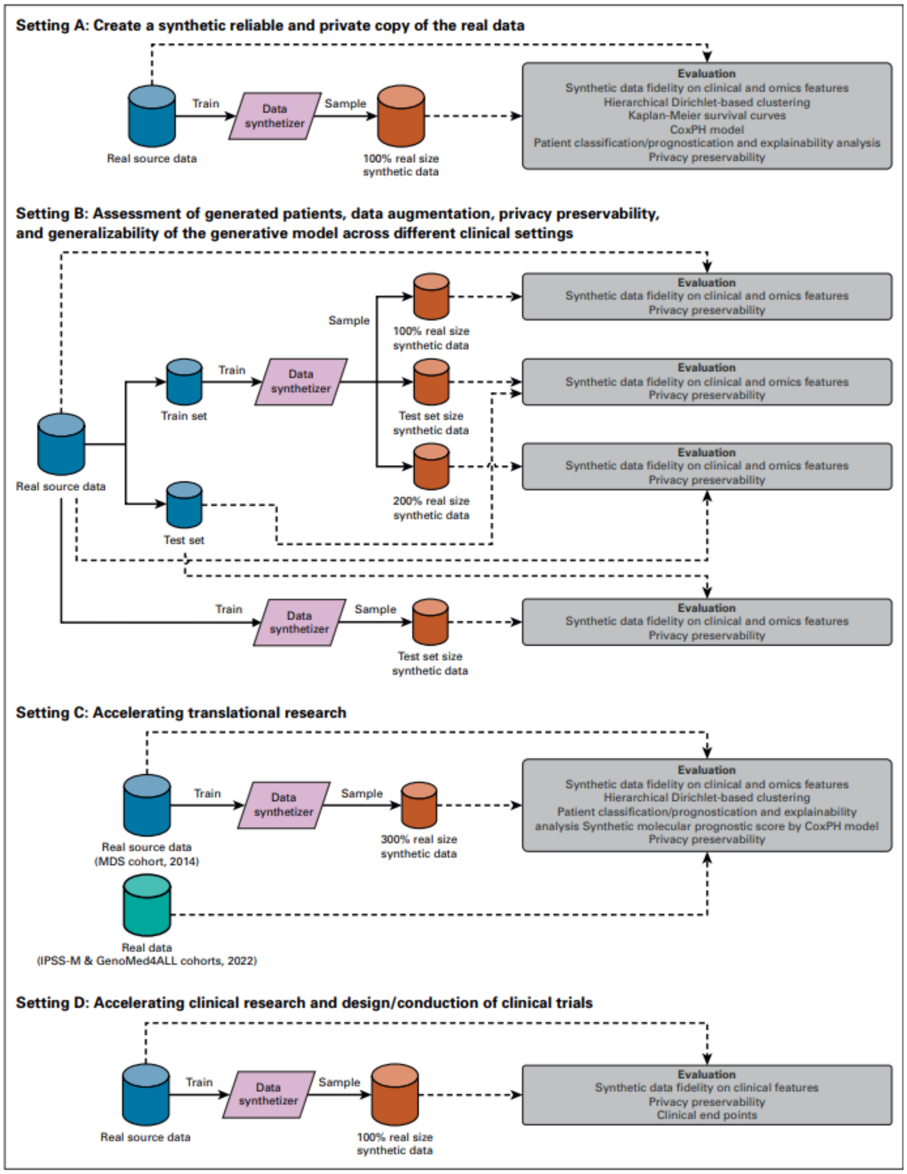

Setting A: Create a synthetic reliable and private copy of the real data
Real source data
Train
Data synthetizer
Sample
100% real size synthetic data
Evaluation
Synthetic data fidelity on clinical and omics features
Hierarchical Dirichlet-based clustering
Kaplan-Meier survival curves
CoxPH model
Patient classification/prognostication and explainability analysis
Privacy preservability
Setting B: Assessment of generated patients, data augmentation, privacy preservability, and generalizability of the generative model across different clinical settings
Real source data
Train set
Test set
Train
Data synthetizer
Sample
100% real size synthetic data
Test set size synthetic data
200% real size synthetic data
Evaluation
Synthetic data fidelity on clinical and omics features
Privacy preservability
Evaluation
Synthetic data fidelity on clinical and omics features
Privacy preservability
Evaluation
Synthetic data fidelity on clinical and omics features
Privacy preservability
Train
Data synthetizer
Sample
Test set size synthetic data
Evaluation
Synthetic data fidelity on clinical and omics features
Privacy preservability
Setting C: Accelerating translational research
Real source data (MDS cohort, 2014)
Train
Data synthetizer
Sample
300% real size synthetic data
Evaluation
Synthetic data fidelity on clinical and omics features
Hierarchical Dirichlet-based clustering
Patient classification/prognostication and explainability analysis Synthetic molecular prognostic score by CoxPH model
Privacy preservability
Real data (IPSS-M & GenoMed4ALL cohorts, 2022)
Setting D: Accelerating clinical research and design/conduction of clinical trials
Real source data
Train
Data synthetizer
Sample
100% real size synthetic data
Evaluation
Synthetic data fidelity on clinical features
Privacy preservability
Clinical end points

Figure 5.1: Overview of experimental settings to validate synthetic data. Setting A: Create a synthetic reliable and private copy of the real data. Setting B: Assessment of generated patients, data augmentation, privacy preservability, and generalizability of the generative model across different clinical settings. Setting C: Accelerating translational research. Setting D: Accelerating clinical research and design/conduction of clinical trials. IPSS-M, Molecular International Prognostic Scoring System; MDS, myelodysplastic syndromes.

944) [109], we generated a 300% augmented synthetic data set. We aimed to recapitulate and anticipate in this cohort of synthetic patients the most relevant and recent insights in personalized medicine (ie, the definition of a new molecular MDS classification and of a molecular scoring system, developed on 2,043 and 2,957 real patients in 2022, respectively) [44, 105].

In *setting D*, we generated synthetic patients to be used as a control arm in clinical trials, thus possibly accelerating clinical development of new drugs/new indications of existing drugs. Starting from 187 MDS treated with luspatercept into a multicenter clinical trial [110], we generated a new synthetic cohort of the same size. Then, we tested the capability of newly generated synthetic patients to recapitulate all the clinical end points of the original study.

5.3 Results

5.3.1 Creation of a Synthetic, Reliable, and Private Reproduction of Real Data (Setting A)

We used 2,043 real MDS from GenoMed4All cohort [105] to generate a new cohort of 2,043 synthetic patients. The model showed high-fidelity performances for both clinical and genomic features (CSF = 93%; GSF = 90%; see Figure 5.2). We then applied Dirichlet processes to compare complex interactions and broad dependencies among genomic features in real versus synthetic patients and we obtained highly comparable results; explainability analysis (SHAP) showed that similar features drive patients' classification in both data sets.

Synthetic patients had comparable survival outcomes with respect to the real ones. When applying the reference scoring system for MDS prognostication (Revised International Prognostic Scoring System), the probability of survival of the five risk categories between synthetic and real patients was comparable (Figure 5.3).

We build a CoxPH model including all features of prognostic relevance with a unique binary covariate (indicating the belonging of the patient to the real or the synthetic data set) that obtained a P value of .742, suggesting that there is no significant difference in the survival probability between the two cohorts in a multivariable setting (Figure 5.3). Concordances obtained for the different category included in the model (demographics, clinical, and genomics) were comparable in both cohorts. Considering the global concordance of the model, we obtained similar results with the model fitted on real versus synthetic data (0.736 ± 0.012 v 0.769 ± 0.012; Figure 5.3).

In terms of privacy metrics, the IMS analysis showed that none of the real patients were copied in the synthetic dataset; moreover, we obtained good results for NNDR (0.64), indicating adequate distance to real data and poor privacy risk [107].

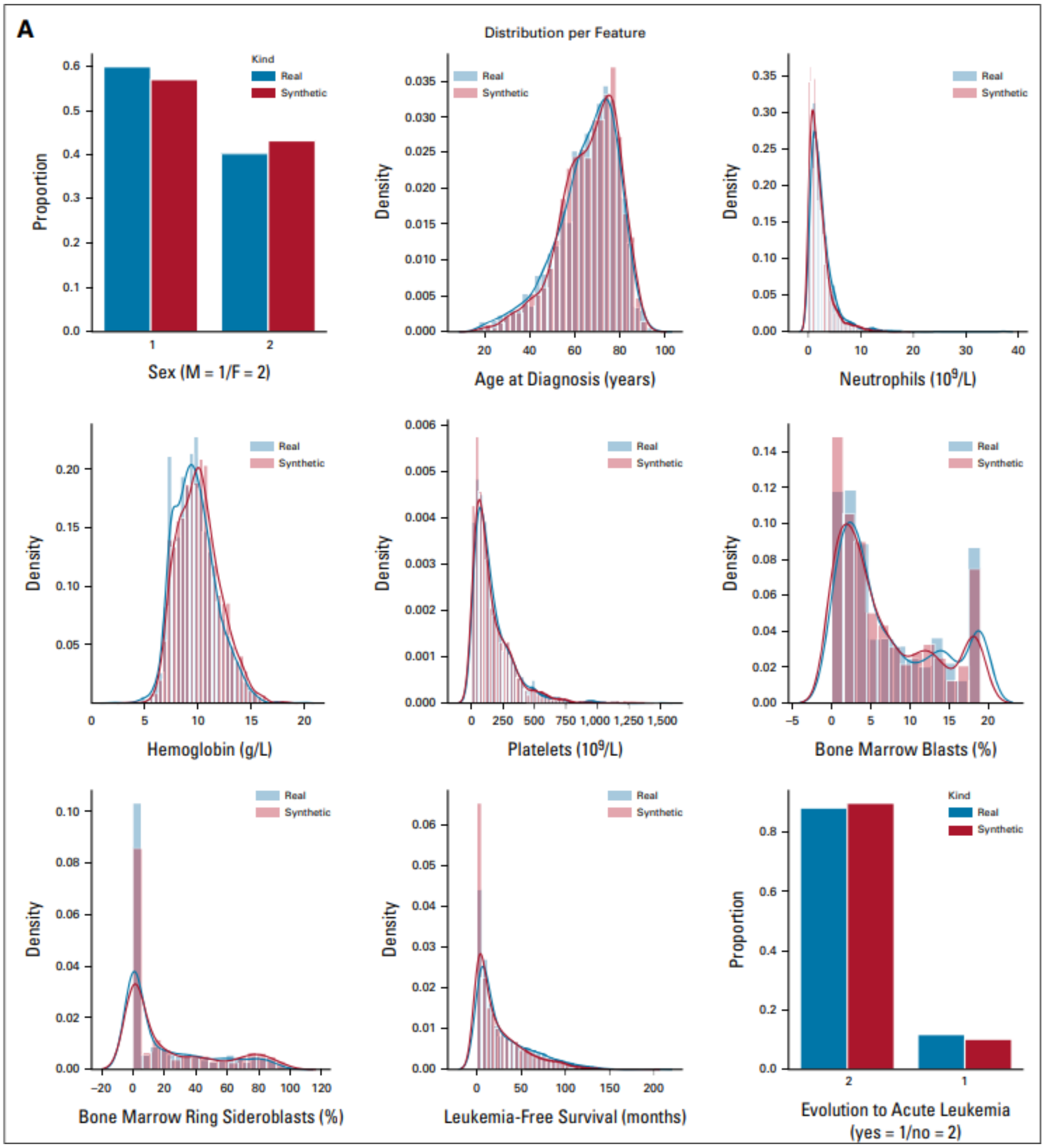


Figure 5.2: SVF on synthetic MDS cohort (n = 2,043), as performed in setting A. (A) Distributions for clinical, demographic, and survival features. Blue illustrates the real data, while red illustrates the synthetic data. (B) Frequency of recurrently mutated genes and chromosomal abnormalities. (C) Pairwise association among genes and/or cytogenetics abnormalities. In the upper triangle, for each couple of genomic abnormalities, the numbers of patients showing mutation co-occurrences are illustrated using a blue and white color scale. In the lower triangle, the gene-gene co-occurrence and mutual exclusivity is assessed using odds ratio, illustrated using a green and yellow color scale according to odds ratio values. All results in (A), (B), and (C) are referring to one MDS synthetic data set of 2,043 patients generated. Detailed results are reported in the Data Supplement. (D) Synthetic data fidelity calculated by SVF on clinical, demographic, and genomic features and patient survival. Average over three training and sampling replications on MDS cohort of 2,043 patients. MDS, myelodysplastic syndromes; SVF, synthetic validation framework.

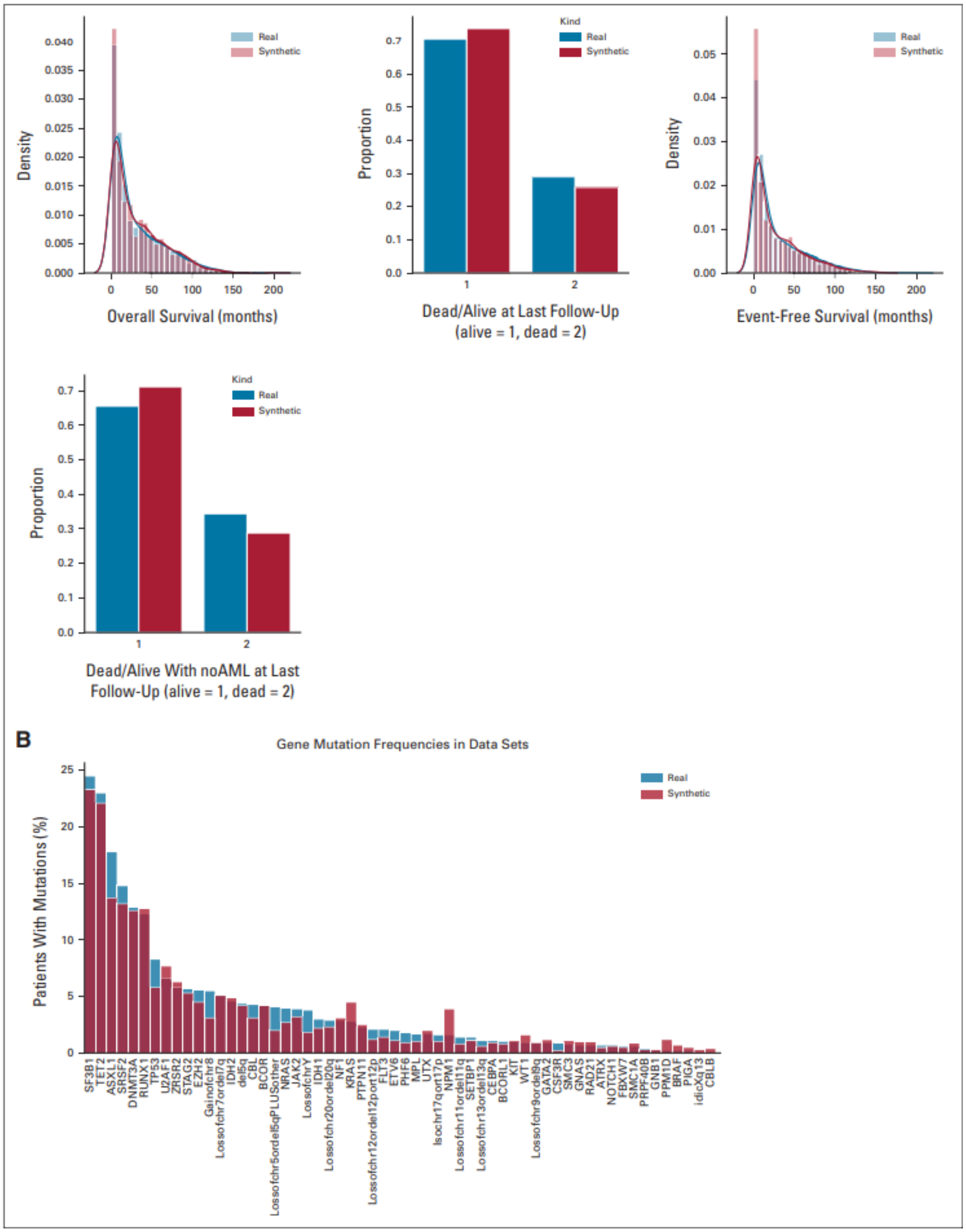


Figure 5.2 (cont.)

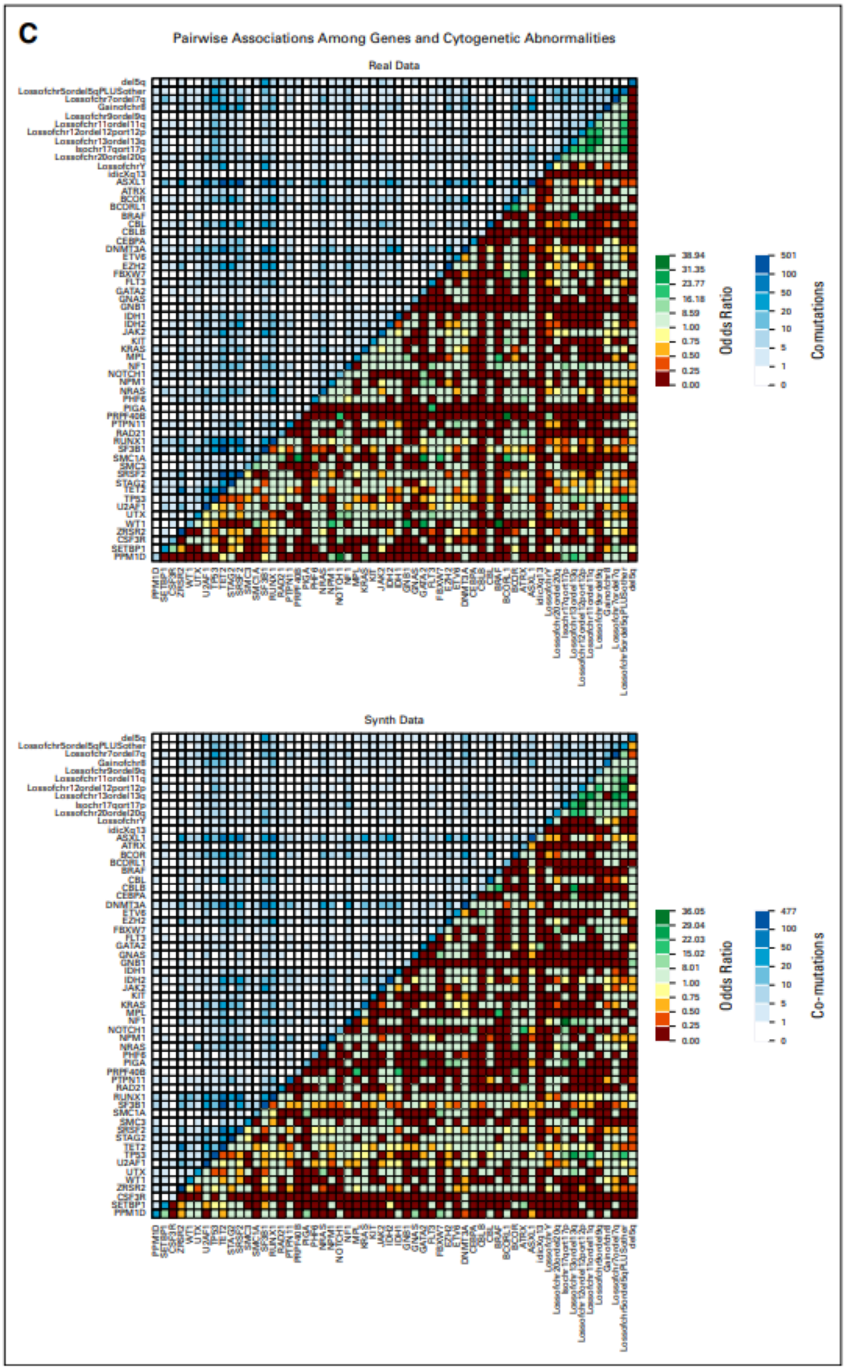
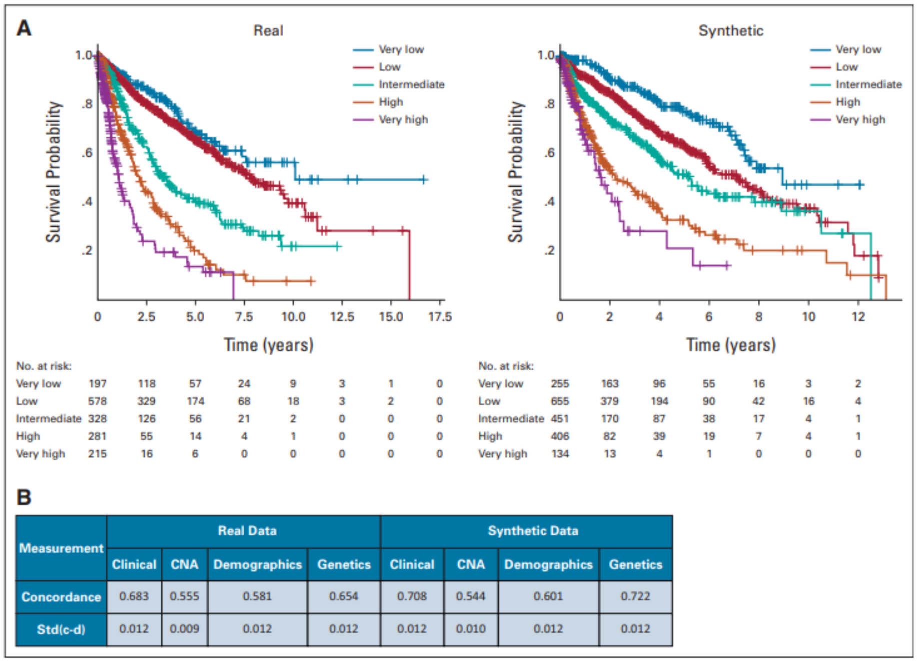


Figure 5.2 (cont.)



No. at risk:								
Very low	197	118	57	24	9	3	1	0
Low	578	329	174	68	18	3	2	0
Intermediate	328	126	56	21	2	0	0	0
High	281	55	14	4	1	0	0	0
Very high	215	16	6	0	0	0	0	0

No. at risk:							
Very low	255	163	96	55	16	3	2
Low	655	379	194	90	42	16	4
Intermediate	451	170	87	38	17	4	1
High	406	82	39	19	7	4	1
Very high	134	13	4	1	0	0	0

Measurement	Real Data				Synthetic Data			
	Clinical	CNA	Demographics	Genetics	Clinical	CNA	Demographics	Genetics
Concordance	0.683	0.555	0.581	0.654	0.708	0.544	0.601	0.722
Std(c-d)	0.012	0.009	0.012	0.012	0.012	0.010	0.012	0.012

Figure 5.3: Patient classification and survival analysis on the synthetic MDS cohort ($n = 2{,}043$), as performed in setting A. (A) Kaplan-Meier survival probability curves obtained from the real (left) and synthetic (right) populations, stratified according to IPSS-R risk categories. The P values of the log-rank test are calculated, confirming the hypothesis of no difference in survival probabilities between real and synthetic patients for every IPSS-R risk group. (B) Partial concordance and standard error for each category of variables obtained from the mixed-effect CoxPH models fitted on the real and synthetic cohorts. CNA, copy number alteration; IPSS-R, Revised International Prognostic Scoring System; MDS, myelodysplastic syndromes.

5.3.2 Resolution of Lack/Incomplete Information, Data Augmentation, Privacy Preservability, and Generalizability of the Model Across Different Clinical Settings (Setting B)

Starting from the MDS GenoMed4all cohort (n = 2,043) [105], we trained the model with a set of a smaller size (including 70% of the patients) and then with a set with 30% of missing information across all features. We obtained the same high-fidelity performances as in setting A, in which synthetic patients were generated form the whole real data set (CSF and GSF were >90% in both experiments).

Then we generated a 200% augmented data set of synthetic MDS patients, resulting into a high fidelity of the model (CSF = 91%; GSF = 89%) that was maintained when comparing the synthetic data sets with the real test set never seen by the model during the training phase (CSF = 90%; GSF = 88%).

When considering a more complex data set (IWG-PM MDS cohort, N = 2,604) including a higher number of genomic features (245 v 65), we obtained comparable fidelity performances to the previous experiments (CSF = 93%; GSF = 93%).

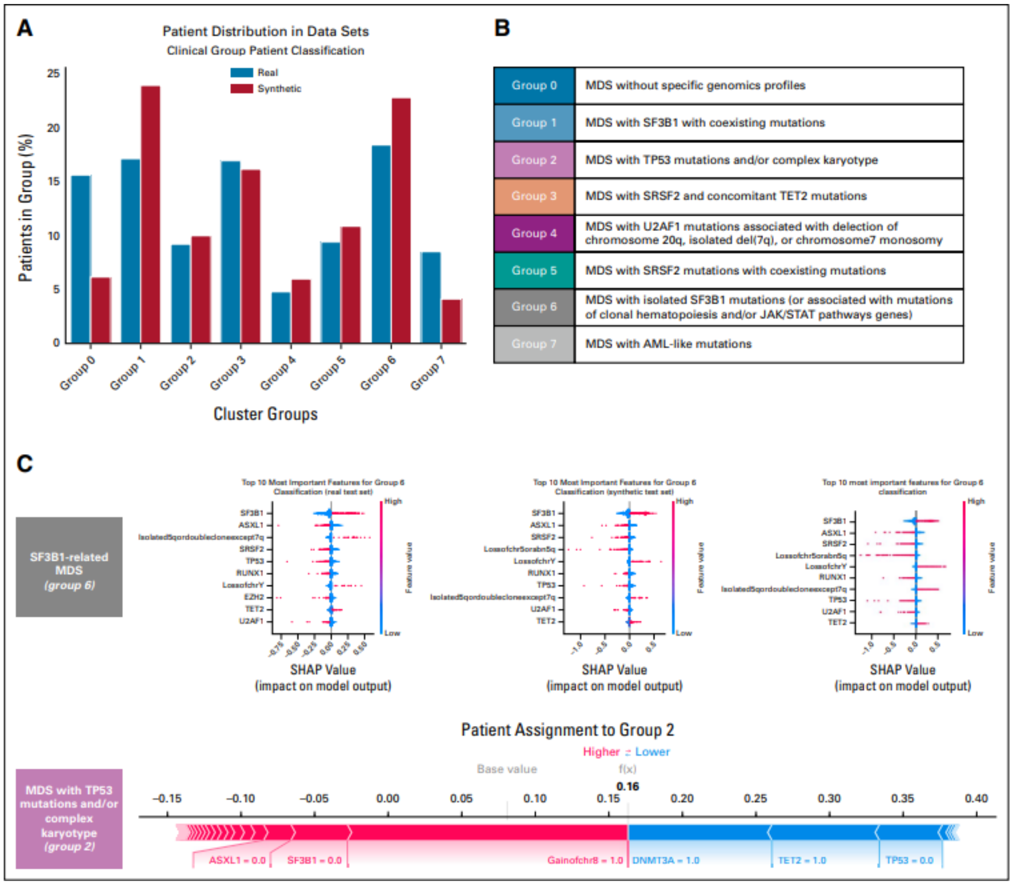
Importantly, a similar trend was noted by replicating all these experiments in a cohort of 1,002 synthetic patients with AML generated form an equal number of real subjects (CSF > 90%; GSF ¿ 88% in all cases), thus providing evidence for a generalizability of the generative model across different clinical settings.

In terms of privacy metrics, in all experiments on the three different synthetic patient populations, the IMS analysis showed that none of the real patients were copied in the synthetic data sets; moreover, we obtained similar good distance results in all experiments for NNDR (values from 0.60 to 0.71).

5.3.3 Accelerating Translational Research by Synthetic Data (Setting C)

Starting from a MDS cohort available in 2014 (N = 944) [109], we generated a 300% augmented synthetic data set of 2,832 patients. Fidelity and privacy performances were comparable with previous experiments (CSF = 92%; GSF = 89%; NNDR = 0.62). We aimed to recapitulate and anticipate in this cohort of synthetic patients the most relevant insights in the field of personalized medicine (ie, the definition of new molecular MDS classification provided on a cohort of 2,043 real patients20 and the definition of the Molecular International Prognostic Scoring Systems [IPSS-M], defined on a cohort 2,897 real patients [44]).

First, Dirichlet processes were applied to synthetic data to define genomic-based clinical entities, resulting in the identification of the same eight disease categories described in a real cohort of 2,043 patients in 2022. Patients' classification into clinical groups followed a similar distribution as the real cohort, and explainability analysis (SHAP) also showed that similar features drive the patients' classification in both data sets. This is visualized in Figure 5.4.



Group 0	MDS without specific genomics profiles
Group 1	MDS with SF3B1 with coexisting mutations
Group 2	MDS with TP53 mutations and/or complex karyotype
Group 3	MDS with SRSF2 and concomitant TET2 mutations
Group 4	MDS with U2AF1 mutations associated with delection of chromosome 20q, isolated del(7q), or chromosome7 monosomy
Group 5	MDS with SRSF2 mutations with coexisting mutations
Group 6	MDS with isolated SF3B1 mutations (or associated with mutations of clonal hematopoiesis and/or JAK/STAT pathways genes)
Group 7	MDS with AML-like mutations

Figure 5.4: Definition of a molecular classification on augmented synthetic MDS cohort starting from 944 patients available in 2014, as performed in setting C. (A) Evaluation of the real (blue) and synthetic (red) patients' distribution considering genomic groups classification. (B) Genomic group definition according to Bersanelli et al [105]. (C) SHAP summary plot analysis on the top 10 most important features for a real test set, a synthetic test set, and a complete augmented synthetic data set for the genomic group 6. Below is the force plot showing the importance of the most relevant features in assigning a synthetic patient to genomic group 2. MDS, myelodysplastic syndromes; SHAP, Shapley Additive Explanations.

As a second experiment, we applied a L1-penalized Cox regression model to the synthetic data set of 2,832 patients to generate a molecular prognostic score (synthetic IPSS-M). After feature selection, we developed a prognostic tool on the synthetic cohort and compared it with IPSS-M developed on real patients. The comparison of the two scores reveals the same feature extraction and the identification of six risk categories with comparable probability of overall survival and leukemia-free survival (see Figure 5.5).

5.3.4 Accelerating Clinical Research and Conduction of Clinical Trials by Using Synthetic Data (Setting D)

We investigated the possibility to use a synthetic data set as a comparison group in a clinical trial. We therefore aimed to replicate a real patient cohort from a multicenter study including 187 patients with MDS who were treated with luspatercept [110].

Eligible patients were age 18 years or older and had an MDS with ring sideroblasts; were receiving regular red blood cells transfusions; and were refractory to erythropoiesis-stimulating agent therapy. Primary end point was transfusion independence (TI) for $\geq$8 weeks during weeks 1-24; key secondary end point was TI for $\geq$12 weeks during both weeks 1-24 and 1-48.

We generated a synthetic cohort (n = 187) from the patients included in the study using all data for training, and we compared the synthetic end points with the original study results. All the characteristics and metrics of the synthetic cohort were comparable with respect to the original data set, as shown in Figure 5.6, with high efficient coefficient of privacy preservability (NNDR = 0.71).

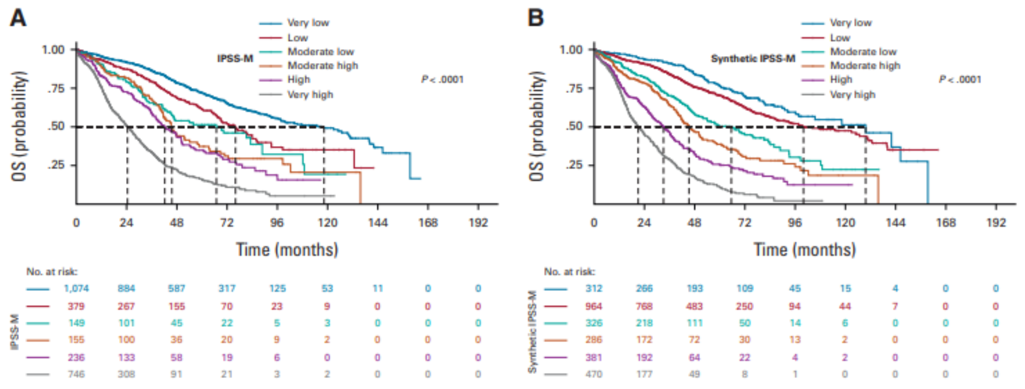
5.3.5 Generator of Synthetic Data

To help clinicians to be familiar with generative AI to build synthetic data, we have created a prototype web portal [111] that allows to generate synthetic patients starting from 2,957 real MDS of IWG-PM cohort [44]. This portal allows to generate synthetic cohorts with different sizes, to verify the performance of the newly generated data (fidelity and privacy preservability), and to download the synthetic data set for research use.

5.4 Discussion

In this study, we showed that synthetic data may (1) efficiently recapitulate statistical properties and complex interactions between clinical and genomic features in hematologic malignancies; (2) replicate reliable estimates of survival and effectiveness of specific treatments; (3) overcome lack/imbalance of information of real data; and (4) allow effective data augmentation.

The implementation of this technology may allow to increase the scientific use and value of real data, and it is expected to accelerate precision medicine

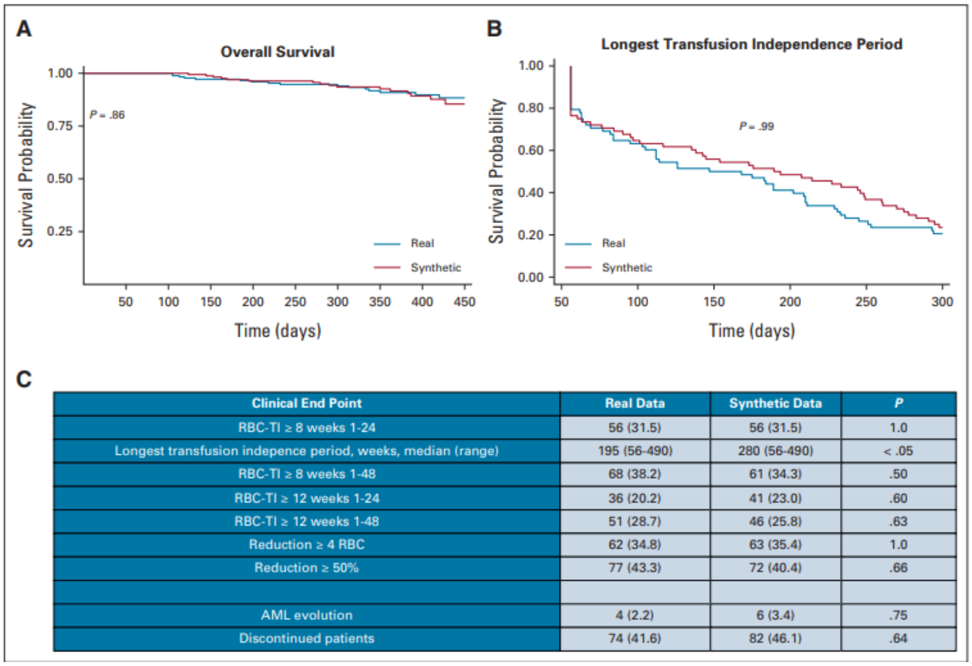


C

Measurement	Synthetic IPSS-M Risk Category					
	Very Low	Low	Moderate Low	Moderate High	High	Very High
Patients, No. (%)	312 (11)	964 (35)	326 (12)	286 (10)	381 (14)	470 (17)
Hazard ratio (95% CI) – LFS	0.5 (0.26-0.82)	1.0 Reference	2.0 (1.37-2.97)	2.5 (1.7-3.77)	5.4 (3.9-7.5)	9.9 (7.4-13.3)
Median LFS (months)	-	-	-	-	66.9	49.7
Hazard ratio (95% CI) – OS	0.77 (0.6-0.97)	1.0 Reference	1.8 (1.5-2.22)	2.5 (2.03-3.03)	3.9 (3.3-4.6)	6.4 (5.4-7.5)
Median OS (months)	129.7	100	65.5	45.3	33.1	21

	IPSS-M Risk Category					
	Very Low	Low	Moderate Low	Moderate High	High	Very High
Patients, No. (%)	1,074 (39)	379 (14)	149 (5.4)	155 (5.6)	236 (9)	746 (27)
Hazard ratio (95% CI) – LFS	0.5 (0.3-0.7)	1.0 Reference	1.6 (0.9-2.7)	1.3 (0.7-2.3)	2.4(1.6-3.8)	5.3 (3.7-7.6)
Median LFS (months)	-	-	-	-	-	55.9
Hazard ratio (95% CI) – OS	0.66 (0.5-0.8)	1.0 Reference	1.37 (1.0-1.8)	1.7 (1.3-2.2)	2 (1.6-2.6)	3.9 (3.2-4.7)
Median OS (months)	118.3	76	66.9	45.6	42.2	24.4

Figure 5.5: Survival analysis on synthetic molecular prognostic score generated (synthetic IPSS-M) performed in setting C. (A) Kaplan-Meier probability estimates of OS for synthetic patients with MDS are represented and stratified by IPSS-M risk categories as defined by Bernard et al.21 P value is from log-rank test. (B) Kaplan-Meier probability estimates of OS for synthetic patients with MDS are represented and stratified by synthetic IPSS-M risk categories. P value is from log-rank test. (C) Percentage of patients in each IPSS-M risk category (both synthetic and original) with the HRs for each outcome, and the median survival for each patient class, where values could be calculated. HR, hazard ratio; IPSS-M, Molecular International Prognostic Scoring System; LFS, leukemia-free survival; MDS, myelodysplastic syndromes; OS, overall survival.



Clinical End Point	Real Data	Synthetic Data	*P*
RBC-TI ≥ 8 weeks 1-24	56 (31.5)	56 (31.5)	1.0
Longest transfusion indepence period, weeks, median (range)	195 (56-490)	280 (56-490)	< .05
RBC-TI ≥ 8 weeks 1-48	68 (38.2)	61 (34.3)	.50
RBC-TI ≥ 12 weeks 1-24	36 (20.2)	41 (23.0)	.60
RBC-TI ≥ 12 weeks 1-48	51 (28.7)	46 (25.8)	.63
Reduction ≥ 4 RBC	62 (34.8)	63 (35.4)	1.0
Reduction ≥ 50%	77 (43.3)	72 (40.4)	.66
AML evolution	4 (2.2)	6 (3.4)	.75
Discontinued patients	74 (41.6)	82 (46.1)	.64

Figure 5.6: Comparison of clinical trial end points between real and synthetic patients, as performed in setting D. (A) Kaplan-Meier survival probability curves compared for real and synthetic patients' overall survival. (B) Kaplan-Meier curves of longest transfusion independence period for real and synthetic patients. The P values of the log-rank test are calculated, confirming the hypothesis of no difference in survival probabilities between real and synthetic cohorts. (C) Study end point comparison between real and synthetic cohorts. RBC-TI, rate of red blood cell transfusion independence.

in hematology and the conduction of clinical trials.

To help clinicians to be familiar with this new technology, we created a prototype web portal that allows to generate synthetic data from a real data set of patients with clinical and genomic information, and that provides a report of the quality of the newly generated synthetic patients.

The implementability of synthetic data in translational and clinical research depends on two main properties: (1) fidelity, ie, the newly generated data should be plausible and preserve structural properties of the real data; (2) privacy, that is, it should be possible to precisely quantify how much information about the original data is revealed through the releasing of the synthetic sample [112, 113].

The use of generative AI rapidly increased the implementation of synthetic data in life sciences in past years [96, 97, 98]. As an example, SyntheticMass hosts over one million synthetic patient records from the state of Massachusetts [114]. In Europe, synthetic data sets that mimic a part of the Netherlands Cancer Registry and Public Health England's Cancer Registration are now available for research purposes [115, 116]. The creation of a synthetic data bank makes the information accessible while also streamlining the data sets that medical research teams have to work with. But, there are limitations: the more complex the data query, the more approximate the results; in particular, the generation of high-fidelity synthetic patients with comprehensive clinical and genomic information reproducing complex interactions among different data layers is still a challenge [96, 97, 98].

In this study, we used an optimized method (conditional GAN) [102, 103, 104] to recapitulate clinical and genomic properties of real patients with myeloid neoplasms, which are rare diseases characterized by large clinical and biological heterogeneity [46, 117]. The methodologic advantage of conditional GAN allowed us to face specific challenges in research on rare diseases (such as lack/imbalance of data) and we provided evidence for a high generalizability of the performances of the model across different clinical settings.

Synthetic data require an extensive validation of their reliability in recapitulating properties of real patients [96, 97, 98, 112, 113, 114]. We therefore created a SVF to perform a clear fidelity analysis of clinical, survival, and genomic information and that may represent a solid basis to define the quality of a newly generated synthetic data set. Moreover, we implemented a comprehensive approach for data explainability [108], thus facilitating the clinical interpretation of the results of deep learning analysis on synthetic data.

Sharing data has the potential to improve decision making and accelerate research and innovation [90, 91, 92, 118]. At the same time, many data are highly sensitive and sharing them may violate fundamental rights guarded by modern privacy regulations [95, 118]. Anonymization (where potentially identifiable variables are removed) is one way to make data available; however, intensive anonymization can degrade the data to the extent that they are no longer fit for purpose. Moreover, several reidentification attempts on anonymized data have been successful and have harmed public and regulators' trust in such methods [119, 120]. We showed that generative AI can guarantee a high privacy preservability of newly generated synthetic data. We focused on analyzing the distance between the real and synthetic patients and we showed that there was enough distance between the real and synthetic patients to avoid the risk of revealing sensitive information from the training data and not too far away to maintain correlations of the source real population [107].

We provided evidence that synthetic data can accelerate translational research in hematology. Since the first publication on clinical relevance of gene mutations in MDS, it took several years to collect real large patient populations for defining a molecular classification and molecular prognostic score [44, 105]. By generating synthetic data from a relative small cohort of patients available in 2014 [109], we were able to recapitulate the definition of genomic-based subgroups and of a molecular prognostic score as described in real cohorts many years later [44, 105].

Finally, synthetic patients could be used in the future to improve the conduction of clinical trials. The use of synthetic control arms may reduce clinical trial costs and duration. Moreover, using a synthetic control arm may ensure that all participants receive the active treatment, thus eliminating patient concerns about treatment assignment [121].

Secondary analyses of data from clinical trials can provide new insights compared with the original publications [122]. In this context, our findings suggest that generative AI can create synthetic patients that efficiently reproduce clinical characteristics and efficacy end points of the original study and that can be promptly available for secondary analyses.

As a possible improvement of our approach, recently, GAN technology was optimized to generate synthetic patients with time-series records and longitudinal evaluation of treatment response (multilabel time-series GAN [MTGAN]) [**?**]. MTGAN can preserve temporal information by developing a temporally correlated generation process, thus finally increasing the generation quality of uncommon diseases and the performance of predictive models.

To maximize the impact of this technology in accelerating precision medicine in hematology, it will be relevant to develop regulatory frameworks involving synthetic data and to define standards for synthetic data quality and privacy preservability [96, 99].

Chapter 6

Conclusion and Future Research

6.1 Conclusion

In this thesis, we presented an all-encompassing approach to the use of *omics* data and artificial intelligence (AI) in hematology. In doing so, we first mapped the hematology data landscape, outlining AI and omics opportunities; data repositories in the European Union; and relevant laws, regulations, and ethical guidelines. We concluded that AI has considerable potential to lead to personalization in hematology prognostics, diagnostics, and treatment. Unfortunately, this potential is currently not realized due to data scarcity and ethics limitations. We also identified synthetic data generation and federated learning as potential solutions to these limitations.

Through the GenoMed4All partners, we obtained a hematology database. Leveraging this, we turned to the validation of existing (non-AI-based) hematology models. In particular, we validated the *Molecular International Prognostic Scoring System* (IPSS-M) with a dataset of 2,876 patients. We also compared the model's performance to that of the *Revised International Prognostic Scoring System* (IPSS-R). We found that IPSS-M improves Myelodysplastic syndromes (MDS) prognostication and might result in a more effective selection of candidates to hematopoietic stem cell transplantation (HSCT). Additional factors other than gene mutations can be involved in determining hypomethylating agents (HMA) sensitivity. The definition of a minimum set of relevant genes may facilitate the clinical implementation of the score.

Next, we focused on the use of AI and omics data for precision MDS prognostication, comparing novel AI-based methods to traditional models (e.g. IPSS). In particular, recurrently mutated genes and chromosomal abnormalities have been identified in myelodysplastic syndromes (MDS). We aim to integrate these genomic features into disease classification and prognostication. To do so, we retrospectively enrolled 2,043 patients. Using Bayesian networks and Dirichlet processes, we combined mutations in 47 genes with cytogenetic abnormalities to identify genetic associations and subgroups. Random-effects Cox proportional hazards multistate modeling was used for developing prognostic models. An independent validation on 318 cases was performed. We found that genomic landscape in MDS reveals distinct subgroups associated with specific clinical features and discrete patterns of evolution, providing a proof of concept for next-generation disease classification and prognosis.

Having shown that the use of omics data and AI models can refine and personalize prognostication in hematological diseases, we turned to data privacy and ethics. Collecting the information needed for AI in large patient populations is challenging and there are many issues concerning patient privacy that need to be accounted for. One approach that can circumvent these issues is the creation of synthetic data that captures the complexities of the original data set (distributions, non-linear relationships, and noise) without including any real patient information. We therefore aimed to: 1) Apply advanced synthetic data generation methods to real-world datasets of different hematological malignancies. 2) Develop a Synthetic Validation Framework to evaluate the quality of synthetic data and perform data augmentation. 3) Test the capability of synthetic data to accelerate translational research.

To achieve these aims, we implemented a Conditional Tabular Wasserstein Generative Adversarial Networks (GAN) architecture with Gradient Penalty to generate synthetic data. Use cases were different cohorts of patients with myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) with available clinical and molecular features. We created a Synthetic Validation Framework to evaluate the quality of generated synthetic data: Clinical Synthetic Fitness (CSF) and Genomic Synthetic Fitness (GSF) scores were calculated as the average of multiple metric tests adopted. Patients were stratified by Hierarchical Dirichlet (HD) clustering. Explainability analysis was carried out by SHapley Additive exPlanations approach (SHAP). Survival analyses were performed by Kaplan-Meier curves and CoxPH models.

We found that GAN-generated synthetic data recapitulate statistical properties and complexity of clinical and genomic features in different hematological malignancies, replicate reliable survival estimates and allow effective data augmentation. The implementation of this technology seems to accelerate precision medicine research in hematology.

Combined, these findings show: 1) the excess demand for omics-based AI in hematology; 2) where and how the required data is stored; 3) the practical and ethical limitations of wider AI adoption in the field; 4) that already gathered data can be integrated into existing prognostic frameworks (IPSS); 5) That omics-based AI outperforms such existing frameworks; and 6) that synthetic data facilitates safe, privacy-respecting use of the information required for next-level hematology prognostics; diagnostics; and treatment analysis.

6.2 Future Research

More research is required to integrate and standardize existing hematology databases. Currently, the hematology data landscape is fragmented, with many organizations leveraging small databases. This translates to a consistent need for advanced data (pre)processing when studies are conducted.

The fields of data ethics and privacy protection are constantly evolving. Research should continuously address the practical impact of recent developments. This allows practitioners to operate in an ethical manner, leveraging the latest concepts, findings, and privacy protecting technologies.

We have shown that omics-based AI has tremendous potential for precision medicine in hematology. More research is required to solidify this finding. Future research should leverage larger volumes of omics-data, integrating more dynamic data types. More advanced AI frameworks should be developed to find genomic markers for a wider range of hematological diseases.

We have shown that synthetic data reconciles data privacy with analytic utility. As the involved technology is highly novel, more research is needed to corroborate these findings. In particular, existing studies using real data can be replicated with synthetic data. Possible limitations should also be explored. Additionally, other emerging PETs should be investigated, such as federated learning.

About the Author

Victor Savevski holds a Medical Engineering degree, a Master of Health Management from SDA Bocconi and a Digital Strategy from Harvard University. He is Chief Innovation Officer and AI Center Director at Humanitas. He is also Adjunct Professor at Humanitas University in Artificial Intelligence and Digital Health Systems in Medicine.

As Chief Digital Officer in Humanitas, he has launched over 50 digital health products, health media platforms, websites, apps and digital therapeutics. Combined, these are currently used by over 90 millions users, aiding them in starting up and managing a large team of developers, web managers, marketers and project managers to digitize and grow the digital presence of the hospitals, university and research centers.

Currently, Victor leads the Humanitas AI Center, focused on research and development on Clinical Applications based on Artificial Intelligence & Machine Learning. In the *GenoMed4All* project, Victor supports the secure and efficient pooling of genomics, clinical and other ”-omics” data through innovative privacy technologies. This makes advanced artificial intelligence innovation possible, even in haematology, a field characterized by its sophistication, confidentiality, and data sparsity.

Besides Genomics4All, Victor has been involved in projects on public health preparedness for Covid-19 (CovidX); pediatric cancer genomic data inventory (EU4CHILD); integrated decision support tools for urban environments, tailored to the needs of European citizens and public stakeholders in domains of health, prosperity, security and overall well-being to address the detrimental impact of Climate Change (HARMONIA). He is a co-founder of *HEALTH 2.0*, the leading market intelligence on new health technology companies; an advisory board member of *NINA Capital*; and a contributor to *Wired*.

Bibliography

[1] The European Hematology Association. EHAweb.org, 2023. URL https://ehaweb.org/.

[2] Chayakrit Krittanawong, Kipp W. Johnson, Edward Choi, Scott Kaplin, Eric Verner, Mullai Murugan, Zhen Wang, Benjamin S. Glicksber, Christopher I. Amos, Michael C. Schatz, and W.H. Wilson Tang. Artificial intelligence and machine learning in endocrinology and metabolism: The dawn of a new era. *Life*, 12(2), 2022. doi: 10.3390/life12020279. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8875522/.

[3] Sriram Gubbi, Pavel Hamet, Johanne Tremblay, Christian A. Koch, and Fady Hannah-Shmouni. Artificial intelligence and machine learning in endocrinology and metabolism: The dawn of a new era. *Frontiers in Endocrinology*, 10, 2019. ISSN 1664-2392. doi: 10.3389/fendo.2019.00185. URL https://www.frontiersin.org/articles/10.3389/fendo.2019.00185.

[4] David M. Kurtz, Mohammad S. Esfahani, Floris Scherer, Joanne Soo, Michael C. Jin, Chih Long Liu, Aaron M. Newman, Ulrich Dürsen, Andreas Hüttman, Olivier Casasnovas, Jason R. Westin, Matthais Ritgen, Sebastian Böttcher, Anton W. Langerak, M. Roschewski, Wyndham H. Wilson, Davide Gaidano, Gianluca ans Rossi, Jasmin Bahlo, Michael Hallek, Robert Tibshirani, Maximilian Diehn, and Ash A. Alizadeh. Dynamic risk profiling using serial tumor biomarkers for personalized outcome prediction. *Cell*, 178(3):699–713, 11 2019. ISSN 0006-4971. doi: 10.1016/j.cell.2019.06.011. URL https://pubmed.ncbi.nlm.nih.gov/31280963/.

[5] Zodwa Dlamini, Amanda Skepu, Namkug Kim, Mahlori Mkhabele, Richard Khanyile, Thulo Molefi, Sikhumbuzo Mbatha, Botle Setlai, Thanyani Mulaudzi, Mzubanzi Mabongo, Meshack Bida, Minah Kgoebane-Maseko, Kgomotso Mathabe, Zarina Lockhat, Mahlatse Kgokolo, Nkhensani Chauke-Malinga, Serwalo Ramagaga, and Rodney Hull. Ai and precision oncology in clinical cancer genomics: From prevention to targeted cancer therapies-an outcomes based patient care. *Informatics in Medicine Unlocked*, 31:100965, 2022. ISSN 2352-9148.

doi: https://doi.org/10.1016/j.imu.2022.100965. URL https://www.sciencedirect.com/science/article/pii/S2352914822001113.

[6] Moritz Gerstung, Andrea Pellagatti, Luca Malcovati, Aristoteles Giagounidis, Matteo G. Della Porta, Martin Jädersten, Hamid Dolatshad, Amit Verma, Nicholas C.P. Cross, Paresh Vyas, Sally Killick, Eva Hellström-Lindberg, Mario Cazzola, Elli Papaemmanuil, Peter J. Campbell, and Jacqueline Boultwood. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nature Communications*, 5901(6), 11 2015. doi: https://doi.org/10.1038/ncomms6901. URL https://www.nature.com/articles/ncomms6901#citeas.

[7] Mario Cazzola, Matteo G Della Porta, and Luca Malcovati. The genetic basis of myelodysplasia and its clinical relevance. *Blood, The Journal of the American Society of Hematology*, 122(25):4021–4034, 2013.

[8] Mario Cazzola, Marianna Rossi, and on behalf of the Associazione Italiana per la Ricerca sul Cancro Gruppo Italiano Malattie Mieloproliferative Malcovati, Luca. Biologic and clinical significance of somatic mutations of SF3B1 in myeloid and lymphoid neoplasms. *Blood*, 121(2):260–269, 01 2013. ISSN 0006-4971. doi: 10.1182/blood-2012-09-399725. URL https://doi.org/10.1182/blood-2012-09-399725.

[9] Siddhartha Jaiswal and Benjamin L. Ebert. Clonal hematopoiesis in human aging and disease. *Science*, 366(6465):eaan4673, 2019. doi: 10.1126/science.aan4673. URL https://www.science.org/doi/abs/10.1126/science.aan4673.

[10] Jul 2023. URL https://genomed4all.eu/.

[11] Stefan Stanojevic, Yijun Li, Aleksandar Ristivojevic, and Lana X Garmire. Computational methods for single-cell multi-omics integration and alignment. *Genomics, Proteomics & Bioinformatics*, 20(5):836–849, 2022.

[12] Jacek Lorkowski, Oliwia Kolaszyńska, and Mieczysław Pokorski. Artificial intelligence and precision medicine: A perspective. In *Integrative Clinical Research*, pages 1–11. Springer, 2021.

[13] Xiujing He, Xiaowei Liu, Fengli Zuo, Hubing Shi, and Jing Jing. Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. *Seminars in Cancer Biology*, 88:187–200, 2023. ISSN 1044-579X. doi: https://doi.org/10.1016/j.semcancer.2022.12.009. URL https://www.sciencedirect.com/science/article/pii/S1044579X22002632.

[14] Babak Arjmand, Shayesteh Kokabi Hamidpour, Akram Tayanloo-Beik, Parisa Goodarzi, Hamid Aghayan, Hossein Adibi, and Bagher Larijani. Machine learning: A new prospect in multi-omics data analysis of cancer. *Frontiers in Genetics*, 13, 01 2022. doi: 10.3389/fgene.2022.824451.

[15] Roni Shouval, Joshua Fein, Bipin Savani, Mohamad Mohty, and Hanan Galski. Machine learning and artificial intelligence in haematology. *British Journal of Haematology*, 192, 06 2020. doi: 10.1111/bjh.16915.

[16] Torsten Haferlach and Wencke Walter. Challenging gold standard hematology diagnostics through the introduction of whole genome sequencing and artificial intelligence. *International Journal of Laboratory Hematology*, 45, 02 2023. doi: 10.1111/ijlh.14033.

[17] Yousra El Alaoui, Adel Elomri, Marwa Qaraqe, Regina Padmanabhan, Ruba Yasin Taha, Halima El Omri, Abdelfatteh El Omri, and Omar Aboumarzouk. A review of artificial intelligence applications in hematology management: Current practices and future prospects. *Journal of Medical Internet Research*, 24, 07 2022. doi: 10.2196/36490.

[18] Elisa Lin, Franklin Fuda, Hung S Luu, Andrew M. Cox, Fengqi Fang, Junlin Feng, and Mingyi Chen. Digital pathology and artificial intelligence as the next chapter in diagnostic hematopathology. *Seminars in Diagnostic Pathology*, 40(2):88–94, 2023. ISSN 0740-2570. doi: https://doi.org/10.1053/j.semdp.2023.02.001. URL https://www.sciencedirect.com/science/article/pii/S0740257023000126. Artificial Intelligence (AI) , machine learning ML) and digital pathology integration are the next major chapter in our diagnostic pathology and laboratory medicine arena.

[19] EuroBloodNet. EuroBloodNet.eu, 2023. URL https://eurobloodnet.eu/.

[20] Samuel D. Warren II and Louis Brandeis. The right to privacy. *Harvard Law Review*, 1890.

[21] Daniel J. Solove. *Understanding Privacy*. Harvard University Press, 2008.

[22] Daniel J. Solove. Conceptualizing privacy. *California Law Review*, 2002.

[23] Luciano Floridi. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press UK, 2014.

[24] Maurizio Mensi and Pietro Falletta. *Il diritto del web. Casi e materiali*. CEDAM, 2015.

[25] William L. Prosser. Privacy. *California Law Review*, 1960.

[26] European Commission. General Data Protection Regulation (GDPR), 2016. URL https://gdpr-info.eu/.

[27] Center for Medicare and Medicaid Services. Health Insurance Portability and Accountability Act (HIPAA), 2022. URL https://www.cms.gov/regulations-and-guidance/administrative-simplification/hipaa-aca.

[28] Legal Information Institute. 45 CFR § 160.103, 2022.

[29] Article 29 Data Protection Working Party (WP29). Opinion 05/2014 on anonymisation techniques, 2014.

[30] Szu-Chuang Li, Bo-Chen Tai, and Yennun Huang. Evaluating variational autoencoder as a private data release mechanism for tabular data. pages 198–1988, 12 2019. doi: 10.1109/PRDC47002.2019.00050.

[31] Supreme Court of California. RICHARD SANDER et al., Plaintiffs and Appellants, v. STATE BAR OF CALIFORNIA et al., Defendants and Respondents., 2013.

[32] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how?, 2022.

[33] European Commission. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (ARTIFICIAL INTELLIGENCE ACT) and Amending Certain Union Legislative Acts.* European Commission: Brussels, 28 April 2021, 2021.

[34] Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective, 2018.

[35] Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission. Ethics guidelines for trustworthy AI, 2019.

[36] Elisabetta Sauta, Marie Robin, Matteo Bersanelli, Erica Travaglino, Manja Meggendorfer, Lin-Pierre Zhao, Juan Berrocal, Claudia Sala, Giulia Maggioni, Massimo Bernardi, Carmen Grazia, Luca Vago, Giulia Rivoli, Lorenza Borin, Saverio D'Amico, Cristina Tentori, Marta Ubezio, Alessia Campagna, Antonio Russo, and Matteo Porta. Real-world validation of molecular international prognostic scoring system for myelodysplastic syndromes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 41:JCO2201784, 03 2023. doi: 10.1200/JCO.22.01784.

[37] Luca Malcovati, Eva Hellström-Lindberg, David Bowen, Lionel Ades, Jaroslav Cermak, Consuelo Del Cañizo, Matteo G Della Porta, Pierre Fenaux, Norbert Gattermann, Ulrich Germing, et al. Diagnosis and treatment of primary myelodysplastic syndromes in adults: recommendations from the european leukemianet. *Blood, The Journal of the American Society of Hematology*, 122(17):2943–2964, 2013.

[38] Peter L Greenberg, Heinz Tuechler, Julie Schanz, Guillermo Sanz, Guillermo Garcia-Manero, Francesc Solé, John M Bennett, David Bowen,

Pierre Fenaux, Francois Dreyfus, et al. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood, The Journal of the American Society of Hematology*, 120(12):2454–2465, 2012.

[39] Matteo Giovanni Della Porta, Heinz Tuechler, Luca Malcovati, J Schanz, G Sanz, Guillermo Garcia-Manero, Francesco Sole, John M Bennett, D Bowen, Pierre Fenaux, et al. Validation of who classification-based prognostic scoring system (wpss) for myelodysplastic syndromes and comparison with the revised international prognostic scoring system (ipss-r). a study of the international working group for prognosis in myelodysplasia (iwg-pm). *Leukemia*, 29(7):1502–1513, 2015.

[40] Matteo Bersanelli, Erica Travaglino, Manja Meggendorfer, Tommaso Matteuzzi, Claudia Sala, Ettore Mosca, Chiara Chiereghin, Noemi Di Nanni, Matteo Gnocchi, Matteo Zampini, et al. Classification and personalized prognostic assessment on the basis of clinical and genomic features in myelodysplastic syndromes. *Journal of Clinical Oncology*, 39(11):1223, 2021.

[41] Elli Papaemmanuil, Moritz Gerstung, Luca Malcovati, Sudhir Tauro, Gunes Gundem, Peter Van Loo, Chris J Yoon, Peter Ellis, David C Wedge, Andrea Pellagatti, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood, The Journal of the American Society of Hematology*, 122(22):3616–3627, 2013.

[42] Elsa Bernard, Yasuhito Nannya, Robert P Hasserjian, Sean M Devlin, Heinz Tuechler, Juan S Medina-Martinez, Tetsuichi Yoshizato, Yusuke Shiozawa, Ryunosuke Saiki, Luca Malcovati, et al. Implications of tp53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes. *Nature medicine*, 26(10):1549–1556, 2020.

[43] Aziz Nazha, Rami Komrokji, Manja Meggendorfer, Xuefei Jia, Nathan Radakovich, Jacob Shreve, C Beau Hilton, Yasunubo Nagata, Betty K Hamilton, Sudipto Mukherjee, et al. Personalized prediction model to risk stratify patients with myelodysplastic syndromes. *Journal of clinical oncology*, 39(33):3737–3746, 2021.

[44] E Bernard, H Tuechler, PL Greenberg, RP Hasserjian, JEA Ossa, Y Nannya, SM Devlin, M Creignou, P Pinel, L Monnier, et al. Molecular international prognostic scoring system for myelodysplastic syndromes. nejm evidence, 1 (7), 2022.

[45] GenoMed4All. GenoMed4All, 2023. URL `https://GenoMed4All.eu/`.

[46] Daniel A Arber, Attilio Orazi, Robert Hasserjian, Jürgen Thiele, Michael J Borowitz, Michelle M Le Beau, Clara D Bloomfield, Mario Cazzola, and James W Vardiman. The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood, The Journal of the American Society of Hematology*, 127(20):2391–2405, 2016.

[47] Joseph D Khoury, Eric Solary, Oussama Abla, Yassmine Akkari, Rita Alaggio, Jane F Apperley, Rafael Bejar, Emilio Berti, Lambert Busque, John KC Chan, et al. The 5th edition of the world health organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. *Leukemia*, 36(7):1703–1719, 2022.

[48] Daniel A Arber, Attilio Orazi, Robert P Hasserjian, Michael J Borowitz, Katherine R Calvo, Hans-Michael Kvasnicka, Sa A Wang, Adam Bagg, Tiziano Barbui, Susan Branford, et al. International consensus classification of myeloid neoplasms and acute leukemias: integrating morphologic, clinical, and genomic data. *Blood, The Journal of the American Society of Hematology*, 140(11):1200–1228, 2022.

[49] Bingshu E Chen, Joan L Kramer, Mark H Greene, and Philip S Rosenberg. Competing risks analysis of correlated failure time data. *Biometrics*, 64 (1):172–179, 2008.

[50] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18): 2543–2546, 1982.

[51] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[52] Moritz Gerstung, Elli Papaemmanuil, Inigo Martincorena, Lars Bullinger, Verena Gaidzik, Peter Paschka, Michael Heuser, Felicitas Thol, Niccolò Bolli, Peter Ganly, Arnold Ganser, Ultan Mcdermott, Konstanze Döhner, Richard Schlenk, Hartmut Döhner, and Peter Campbell. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nature Genetics*, 49, 01 2017. doi: 10.1038/ng.3756.

[53] CoxHD Model, 2017. URL `https://github.com/mg14/CoxHD`.

[54] Theo De Witte, David Bowen, Marie Robin, Luca Malcovati, Dietger Niederwieser, Ibrahim Yakoub-Agha, Ghulam J Mufti, Pierre Fenaux, Guillermo Sanz, Rodrigo Martino, et al. Allogeneic hematopoietic stem cell transplantation for mds and cmml: recommendations from an international expert panel. *Blood, The Journal of the American Society of Hematology*, 129(13):1753–1762, 2017.

[55] Matteo G Della Porta, Anna Gallì, Andrea Bacigalupo, Silvia Zibellini, Massimo Bernardi, Ettore Rizzo, Bernardino Allione, Maria Teresa Van Lint, Pietro Pioltelli, Paola Marenco, et al. Clinical effects of driver somatic mutations on the outcomes of patients with myelodysplastic syndromes treated with allogeneic hematopoietic stem-cell transplantation. *Journal of clinical oncology*, 34(30):3627, 2016.

[56] R Coleman Lindsley, Wael Saber, Brenton G Mar, Robert Redd, Tao Wang, Michael D Haagenson, Peter V Grauman, Zhen-Huan Hu,

Stephen R Spellman, Stephanie J Lee, et al. Prognostic mutations in myelodysplastic syndrome after stem-cell transplantation. *New England Journal of Medicine*, 376(6):536–547, 2017.

[57] Corey S Cutler, Stephanie J Lee, Peter Greenberg, H Joachim Deeg, Waleska S Pérez, Claudio Anasetti, Brian J Bolwell, Mitchell S Cairo, Robert Peter Gale, John P Klein, et al. A decision analysis of allogeneic bone marrow transplantation for the myelodysplastic syndromes: delayed transplantation for low-risk myelodysplasia is associated with improved outcome. *Blood*, 104(2):579–585, 2004.

[58] Matteo G Della Porta, Christopher H Jackson, Emilio P Alessandrino, Marianna Rossi, Andrea Bacigalupo, Maria Teresa van Lint, Massimo Bernardi, Bernardino Allione, Alberto Bosi, Stefano Guidi, et al. Decision analysis of allogeneic hematopoietic stem cell transplantation for patients with myelodysplastic syndrome stratified according to the revised international prognostic scoring system. *Leukemia*, 31(11):2449–2457, 2017.

[59] N Kroger, Simona Iacobelli, G Franke, Uwe Platzbecker, Ruzena Uddin, K Hubel, Christof Scheid, Thomas Weber, Marie Robin, Matthias Stelljes, et al. Dose-reduced versus standard conditioning followed by allogeneic stem-cell transplantation for patients with myelodysplastic syndrome: a prospective randomized phase iii study of the ebmt (ricmac trial). *Journal of Clinical Oncology*, 35(19):2157–2164, 2017.

[60] Pierre Fenaux, Ghulam J Mufti, Eva Hellstrom-Lindberg, Valeria Santini, Carlo Finelli, Aristoteles Giagounidis, Robert Schoch, Norbert Gattermann, Guillermo Sanz, Alan List, et al. Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase iii study. *The lancet oncology*, 10(3):223–232, 2009.

[61] AM Zeidan, MA Sekeres, G Garcia-Manero, DP Steensma, K Zell, J Barnard, NA Ali, C Zimmerman, G Roboz, Amy DeZern, et al. Comparison of risk stratification tools in predicting outcomes of patients with higher-risk myelodysplastic syndromes treated with azanucleosides. *Leukemia*, 30(3):649–657, 2016.

[62] Andrea Kuendgen, Catharina Müller-Thomas, Michael Lauseker, Torsten Haferlach, Petra Urbaniak, Thomas Schroeder, Carolin Brings, Michael Wulfert, Manja Meggendorfer, Barbara Hildebrandt, et al. Efficacy of azacitidine is independent of molecular and clinical characteristics-an analysis of 128 patients with myelodysplastic syndromes or acute myeloid leukemia and a review of the literature. *Oncotarget*, 9(45):27882, 2018.

[63] John S Welch, Allegra A Petti, Christopher A Miller, Catrina C Fronick, Michelle O'Laughlin, Robert S Fulton, Richard K Wilson, Jack D Baty, Eric J Duncavage, Bevan Tandon, et al. Tp53 and decitabine in acute

myeloid leukemia and myelodysplastic syndromes. *New England Journal of Medicine*, 375(21):2023–2036, 2016.

[64] Matilde Y Follo, Carlo Finelli, Sara Mongiorgi, Cristina Clissa, Costanza Bosi, Nicoletta Testoni, Francesca Chiarini, Giulia Ramazzotti, Michele Baccarani, Alberto M Martelli, et al. Reduction of phosphoinositide-phospholipase c beta1 methylation predicts the responsiveness to azacitidine in high-risk mds. *Proceedings of the National Academy of Sciences*, 106(39):16811–16816, 2009.

[65] Yao-Chung Liu, Junsu Kwon, Emiliano Fabiani, Zhijian Xiao, Yanjing V Liu, Matilde Y Follo, Jinqin Liu, Huijun Huang, Chong Gao, Jun Liu, et al. Demethylation and up-regulation of an oncogene after hypomethylating therapy. *New England Journal of Medicine*, 386(21):1998–2010, 2022.

[66] Lionel Adès, Raphael Itzykson, and Pierre Fenaux. Myelodysplastic syndromes. *The Lancet*, 383(9936):2239–2252, 2014.

[67] MATTEO GIOVANNI Della Porta, E Travaglino, E Boveri, M Ponzoni, Luca Malcovati, E Papaemmanuil, GM Rigolin, C Pascutto, G Croci, U Gianelli, et al. Minimal morphological criteria for defining bone marrow dysplasia: a basis for clinical implementation of who classification of myelodysplastic syndromes. *Leukemia*, 29(1):66–75, 2015.

[68] Leonor Senent, Leonor Arenillas, Elisa Luño, Juan C Ruiz, Guillermo Sanz, and Lourdes Florensa. Reproducibility of the world health organization 2008 criteria for myelodysplastic syndromes. *haematologica*, 98(4): 568, 2013.

[69] Luca Malcovati, Matteo Giovanni Della Porta, Cristiana Pascutto, Rosangela Invernizzi, Marina Boni, Erica Travaglino, Francesco Passamonti, Luca Arcaini, Margherita Maffioli, Paolo Bernasconi, et al. Prognostic factors and life expectancy in myelodysplastic syndromes classified according to who criteria: a basis for clinical decision making. *Journal of Clinical Oncology*, 23(30):7594–7603, 2005.

[70] E Papaemmanuil, Mario Cazzola, J Boultwood, Luca Malcovati, P Vyas, D Bowen, A Pellagatti, JS Wainscoat, E Hellstrom-Lindberg, C Gambacorti-Passerini, et al. Somatic sf3b1 mutation in myelodysplasia with ring sideroblasts. *New England Journal of Medicine*, 365(15): 1384–1395, 2011.

[71] Kenichi Yoshida, Masashi Sanada, Yuichi Shiraishi, Daniel Nowak, Yasunobu Nagata, Ryo Yamamoto, Yusuke Sato, Aiko Sato-Otsubo, Ayana Kon, Masao Nagasaki, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, 478(7367):64–69, 2011.

[72] Elli Papaemmanuil, Moritz Gerstung, Lars Bullinger, Verena I Gaidzik, Peter Paschka, Nicola D Roberts, Nicola E Potter, Michael Heuser, Felicitas Thol, Niccolo Bolli, et al. Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine*, 374(23): 2209–2221, 2016.

[73] Jacob Grinfeld, Jyoti Nangalia, E Joanna Baxter, David C Wedge, Nicos Angelopoulos, Robert Cantrill, Anna L Godfrey, Elli Papaemmanuil, Gunes Gundem, Cathy MacLean, et al. Classification and personalized prognosis in myeloproliferative neoplasms. *New England Journal of Medicine*, 379(15):1416–1430, 2018.

[74] J. Grinfeld, J. Nangalia, and et al. Baxter, E.J. R package for hierarchical dirichlet process, . URL `https://github.com/nicolaroberts/hdp`.

[75] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[76] Aris Perperoglou. Cox models with dynamic ridge penalties on time-varying effects of the covariates. *Statistics in medicine*, 33(1):170–180, 2014.

[77] J. Grinfeld, J. Nangalia, and et al. Baxter, E.J. Euromds project: Personalized prediction of clinical outcome in patients with myelodysplastic syndrome according to genomic and clinical features, . URL `https://mds.itb.cnr.it/#/mds`.

[78] Andrea Pellagatti, Richard N Armstrong, Violetta Steeples, Eshita Sharma, Emmanouela Repapi, Shalini Singh, Andrea Sanchi, Aleksandar Radujkovic, Patrick Horn, Hamid Dolatshad, et al. Impact of spliceosome mutations on rna splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. *Blood, The Journal of the American Society of Hematology*, 132(12):1225–1240, 2018.

[79] Luca Malcovati, Elli Papaemmanuil, David T Bowen, Jacqueline Boultwood, Matteo G Della Porta, Cristiana Pascutto, Erica Travaglino, Michael J Groves, Anna L Godfrey, Ilaria Ambaglio, et al. Clinical significance of sf3b1 mutations in myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology*, 118(24):6239–6246, 2011.

[80] Thorsten Klampfl, Heinz Gisslinger, Ashot S Harutyunyan, Harini Nivarthi, Elisa Rumi, Jelena D Milosevic, Nicole CC Them, Tiina Berg, Bettina Gisslinger, Daniela Pietra, et al. Somatic mutations of calreticulin in myeloproliferative neoplasms. *New England Journal of Medicine*, 369(25):2379–2390, 2013.

[81] Brian Reilly, Tiffany N Tanaka, Dinh Diep, Huwate Yeerna, Pablo Tamayo, Kun Zhang, and Rafael Bejar. Dna methylation identifies genetically and prognostically distinct subtypes of myelodysplastic syndromes. *Blood advances*, 3(19):2845–2858, 2019.

[82] So Masaki, Shun Ikeda, Asuka Hata, Yusuke Shiozawa, Ayana Kon, Seishi Ogawa, Kenji Suzuki, Fumihiko Hakuno, Shin-Ichiro Takahashi, and Naoyuki Kataoka. Myelodysplastic syndrome-associated srsf2 mutations cause splicing changes by altering binding motif sequences. *Frontiers in genetics*, 10:338, 2019.

[83] Yang Liang, Toma Tebaldi, Kai Rejeski, Poorval Joshi, Giovanni Stefani, Ashley Taylor, Yuanbin Song, Radovan Vasic, Jamie Maziarz, Kunthavai Balasubramanian, et al. Srsf2 mutations drive oncogenesis by activating a global program of aberrant alternative splicing in hematopoietic cells. *Leukemia*, 32(12):2659–2671, 2018.

[84] T Haferlach, Y Nagata, V Grossmann, Y Okuno, U Bacher, G Nagae, S Schnittger, M Sanada, A Kon, T Alpermann, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*, 28(2): 241–247, 2014.

[85] Ulrike Bacher, Torsten Haferlach, Susanne Schnittger, Melanie Zenger, Manja Meggendorfer, Sabine Jeromin, Andreas Roller, Vera Grossmann, Maria-Theresa Krauth, Tamara Alpermann, et al. Investigation of 305 patients with myelodysplastic syndromes and 20q deletion for associated cytogenetic and molecular genetic lesions and their prognostic impact. *British journal of haematology*, 164(6):822–833, 2014.

[86] Detlef Haase, Kristen E Stevenson, Donna Neuberg, Jaroslaw P Maciejewski, Aziz Nazha, Mikkael A Sekeres, Benjamin L Ebert, Guillermo Garcia-Manero, Claudia Haferlach, Torsten Haferlach, et al. Tp53 mutation status divides myelodysplastic syndromes with complex karyotypes into distinct prognostic subgroups. *Leukemia*, 33(7):1747–1758, 2019.

[87] Austin G Kulasekararaj, Jie Jiang, Alexander E Smith, Azim M Mohamedali, Syed Mian, Shreyans Gandhi, Joop Gaken, Barbara Czepulkowski, Judith CW Marsh, and Ghulam J Mufti. Somatic mutations identify a subgroup of aplastic anemia patients who progress to myelodysplastic syndrome. *Blood, The Journal of the American Society of Hematology*, 124(17):2698–2704, 2014.

[88] Saverio D'Amico, Elisabetta Sauta, Matteo Bersanelli, Daniele Dall'Olio, Claudia Sala, Lorenzo Dall'Olio, Pierandrea Morandini, Tobia Tommasini, Marilena Bicchieri, Matteo Zampini, Victor Savevski, Iñigo Prada-Luengo, Anders Krogh, Uwe Platzbecker, Maria Diez-Campelo, Valeria Santini, Pierre Fenaux, Torsten Haferlach, Castellani Gastone, and Matteo G. Della Porta. Synthetic data generation by artificial intelligence to accelerate translational research and precision medicine in hematological

malignancies. *Blood*, 140(Supplement 1):9744–9746, 11 2022. ISSN 0006-4971. doi: 10.1182/blood-2022-168646. URL `https://doi.org/10.1182/blood-2022-168646`.

[89] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[90] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[91] Michael J Pencina, Benjamin A Goldstein, and Ralph B D'Agostino. Prediction models-development, evaluation, and clinical application. *The New England journal of medicine*, 382(17):1583–1586, 2020.

[92] Bhavneet Bhinder, Coryandar Gilvary, Neel S Madhukar, and Olivier Elemento. Artificial intelligence in cancer research and precision medicine. *Cancer discovery*, 11(4):900–915, 2021.

[93] Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

[94] Andrew Daniel Trister. The tipping point for deep learning in oncology. *JAMA oncology*, 5(10):1429–1430, 2019.

[95] Sharona Hoffman. Privacy and security—protecting patients' health information. *New England Journal of Medicine*, 387(21):1913–1916, 2022.

[96] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[97] Jean-Francois Rajotte, Robert Bergen, David L Buckeridge, Khaled El Emam, Raymond Ng, and Elissa Strome. Synthetic data as an enabler for machine learning applications in medicine. *Iscience*, 25(11), 2022.

[98] Sergey I Nikolenko. Synthetic data for deep learning. *arXiv preprint arXiv:1909.11512*, 2019.

[99] S Castellanos. Fake it to make it: Companies beef up ai models with synthetic data. *WSJ Pro*, 2021.

[100] World Health Organization et al. Ethics and governance of artificial intelligence for health: Who guidance. 2021.

[101] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[102] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[103] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[104] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[105] Matteo Bersanelli, Erica Travaglino, Manja Meggendorfer, Tommaso Matteuzzi, Claudia Sala, Ettore Mosca, Chiara Chiereghin, Noemi Di Nanni, Matteo Gnocchi, Matteo Zampini, et al. Classification and personalized prognostic assessment on the basis of clinical and genomic features in myelodysplastic syndromes. *Journal of Clinical Oncology*, 39(11):1223, 2021.

[106] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18): 2543–2546, 1982.

[107] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctabgan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.

[108] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[109] T Haferlach, Y Nagata, V Grossmann, Y Okuno, U Bacher, G Nagae, S Schnittger, M Sanada, A Kon, T Alpermann, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*, 28(2): 241–247, 2014.

[110] Luca Lanino, Prassede Salutari, Alessandra Perego, Bruno Fattizzo, Marta Riva, Marta Ubezio, Pellegrino Musto, Daniela Cilloni, Esther Natalie Oliva, Maria Teresa Voso, et al. Efficacy and safety of luspatercept in adult patients with transfusion-dependent anemia due to very low, low and intermediate risk myelodysplastic syndromes (mds) with ring sideroblasts, who had an unsatisfactory response to or are ineligible for erythropoietin-based therapy: A retrospective multicenter study by fondazione italiana sindromi mielodisplastiche (fisim ets). *Blood*, 140(Supplement 1):6945–6948, 2022.

[111] Saverio D'amico, Daniele Dall'Olio, Claudia Sala, Lorenzo Dall'Olio, Elisabetta Sauta, Matteo Zampini, Gianluca Asti, Luca Lanino, Giulia Maggioni, Alessia Campagna, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. *JCO Clinical Cancer Informatics*, 7:e2300021, 2023.

[112] Zahra Azizi, Chaoyi Zheng, Lucy Mosquera, Louise Pilote, and Khaled El Emam. Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ open*, 11(4):e043497, 2021.

[113] Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR medical informatics*, 10(4):e35734, 2022.

[114] Junqiao Chen, David Chun, Milesh Patel, Epson Chiang, and Jesse James. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (synthea) using clinical quality measures. *BMC medical informatics and decision making*, 19(1):1–9, 2019.

[115] Integraal Kankercentrum Nederland IKNL. Netherlands comprehensive cancer organisation.

[116] Pia Horvat, Christen M Gray, Alexandrina Lambova, Jennifer B Christian, Laura Lasiter, Mark Stewart, Jeff Allen, Paul Clarke, Cong Chen, and Adam Reich. Comparing findings from a friends of cancer research exploratory analysis of real-world end points with the cancer analysis system in england. *JCO Clinical Cancer Informatics*, 5:1155–1168, 2021.

[117] Daniel A Arber, Attilio Orazi, Robert P Hasserjian, Michael J Borowitz, Katherine R Calvo, Hans-Michael Kvasnicka, Sa A Wang, Adam Bagg, Tiziano Barbui, Susan Branford, et al. International consensus classification of myeloid neoplasms and acute leukemias: integrating morphologic, clinical, and genomic data. *Blood, The Journal of the American Society of Hematology*, 140(11):1200–1228, 2022.

[118] Heidi Beate Bentzen, Rosa Castro, Robin Fears, George Griffin, Volker Ter Meulen, and Giske Ursin. Remove obstacles to sharing health data with researchers outside of the european union. *Nature Medicine*, 27(8): 1329–1333, 2021.

[119] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 510–526. Springer, 2019.

[120] Saira Ghafur, Jackie Van Dael, Melanie Leis, Ara Darzi, and Aziz Sheikh. Public perceptions on data sharing: key insights from the uk and the usa. *The Lancet Digital Health*, 2(9):e444–e446, 2020.

[121] Damien G Finniss, Ted J Kaptchuk, Franklin Miller, and Fabrizio Benedetti. Biological, clinical, and ethical advances of placebo effects. *The Lancet*, 375(9715):686–695, 2010.

[122] Toby Wilkinson, Siddharth Sinha, Niels Peek, and Nophar Geifman. Clinical trial data reuse–overcoming complexities in trial design and data sharing. *Trials*, 20:1–4, 2019.