GENOMED4ALL

D6.1
Literature mining and preprocessing

GENOMED4ALL receives funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 101017549

GENOMED4ALL

Genomics for Next Generation Healthcare

D6.1
Literature mining and preprocessing

Revision **v1.1**

Work package	WP6
Task	6.1
Due date	31-08-2021
Submission date	31-08-2021
Deliverable lead	UCPH
Version	V1.1
Authors	Tiziana Sanavia (UNITO), Lorenzo Dall'Olio (UNIBO), Iñigo Prada-Luengo (UCPH), Anders Krogh (UCPH) & Gastone Castellani (UNIBO)
Reviewers	Davide Piscia (CRG) and Maurizio Ortali (Cineca)

Abstract

This deliverable contains the first literature mining and the first version of the AI software release. It is prepared three months into WP6 (started M6) and therefore presents the preliminary work.

The literature mining, which is the first part, consists of a review of the most important AI methods and has as an appendix a paper from consortium partners, which also represents a large literature mining effort.

The AI software release consists of 3 parts:

1. Software written in Python for analysis of MDS data, including functions for preprocessing and clustering.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

2. Software written in R for survival analysis.
3. Third party software used in the project, which is described in the text and in the gitlab repository.

The software is available for the entire consortium in the GenoMed4all GitLab repository and is also included in appendices. It will be made publicly available at a later stage in the project.

Keywords

AI literature mining, AI software

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

Document revision history

Version	Date	Description of change	Contributor(s)
V0.1	23-08-2021	1st version of deliverable template	Tiziana Sanavia (UNITO), Lorenzo Dall'Olio (UNIBO), Iñigo Prada-Luengo (UCPH), Anders Krogh (UCPH) & Gastone Castellani (UNIBO)
V1.0	31-08-2021	Final version after internal review	Tiziana Sanavia (UNITO), Lorenzo Dall'Olio (UNIBO), Iñigo Prada-Luengo (UCPH), Anders Krogh (UCPH) & Gastone Castellani (UNIBO)
V1.1	01-09-2021	Minor changes after internal review	Tommaso Folgia, Tiziana Sanavia (UNITO), Lorenzo Dall'Olio (UNIBO), Iñigo Prada-Luengo (UCPH), Anders Krogh (UCPH) & Gastone Castellani (UNIBO)

Disclaimer

Copyright notice

Project co-funded by the European Commission in the H2020 Programme

Nature of the deliverable		**R+OTHER**
Dissemination level		
PU	Public, fully open. e.g., website	✓
CL	Classified information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to GENOMED4ALL project and Commission Services	

*** Deliverable types:**

R: document, report (excluding periodic and final reports).
DEM: demonstrator, pilot, prototype, plan designs.
DEC: websites, patent filings, press and media actions, videos, etc.
OTHER: software, technical diagrams, etc.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

Table of contents

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

1 Preprocessing pipelines for clinical, genomic and imaging data in GenoMed4all case studies

For all GenoMed4All case studies (Myelodysplastic Syndrome, Multiple Myeloma and Sickle Cell Disease), two main preliminary pipelines have been developed to apply data cleaning and feature extraction on all clinical, genomic and imaging data. These pipelines are structured as parallelizable modules in order to help the federated implementation of the pipeline. All the code will be open source and it will be released in Python language.

A. Pipeline for clinical and genomic data

Two alternative pipelines are being explored:

1) After data collection, one imputation (MiceRanger https://cran.r-project.org/web/packages/miceRanger/index.html, with variable percentage) step is performed, in order to mitigate the data missingness(*1*). After the imputation step, data clustering is used to identify a putative patient stratification. Finally, the obtained patient stratification significance is tested by looking at some clinical outcomes: Overall Survival, Disease Free Survival, Relapse, Response to Therapy etc. (**Figure 1, upper panel**).
2) After data collection, no imputation step is applied. In this case there is only a data sparsity reduction, and, after this step, dimensionality reduction and clustering are performed, in order to feed the prediction models (**Figure 1, lower panel**).

B. Pipeline for radiomic data

After data collection, feature extraction on the Region Of Interest (ROI) is performed by using well known and widely used in the Radiomics community software such as FreeSurfer, 3DSlicer and PyRadiomics, or by implementing ad-hoc extraction methods such as those based on the co-occurrence matrix (e.g. Gray-level co-occurrence matrix) to extract Haralick features, which are used to quantify an image based on the texture. These features are then clustered in order to stratify patients through a predictive model, testing their importance for each considered clinical outcome, e.g. Overall Survival, Disease Free Survival, Relapse, Response to Therapy etc. As a complementary step we also consider the possibility to take further input from the pipeline designed for the genomic and clinical data, using available information such as age, gender, morphometric data, Body Mass Index, Comorbidities, Previous therapies etc. (**Figure 2**).

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

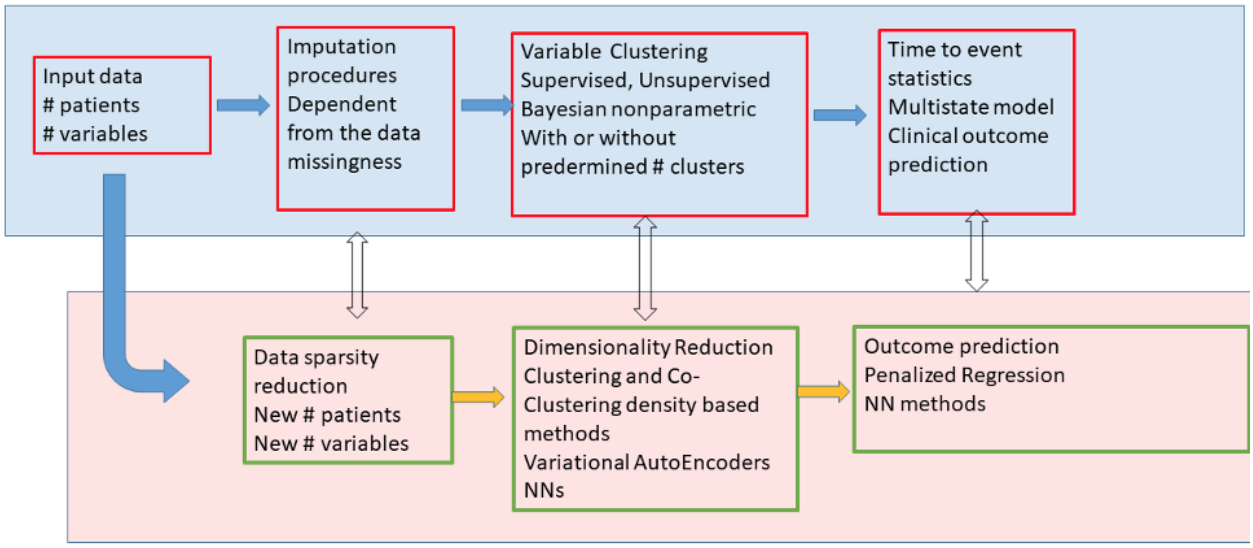


Figure 1. Pipeline for clinical and genomic data. Two possible alternatives are displayed, where the application of the imputation method is dependent from the mechanisms of missingness(*1*) After the imputation step, dimensionality reduction and/or data clustering are used to identify a putative patient stratification.

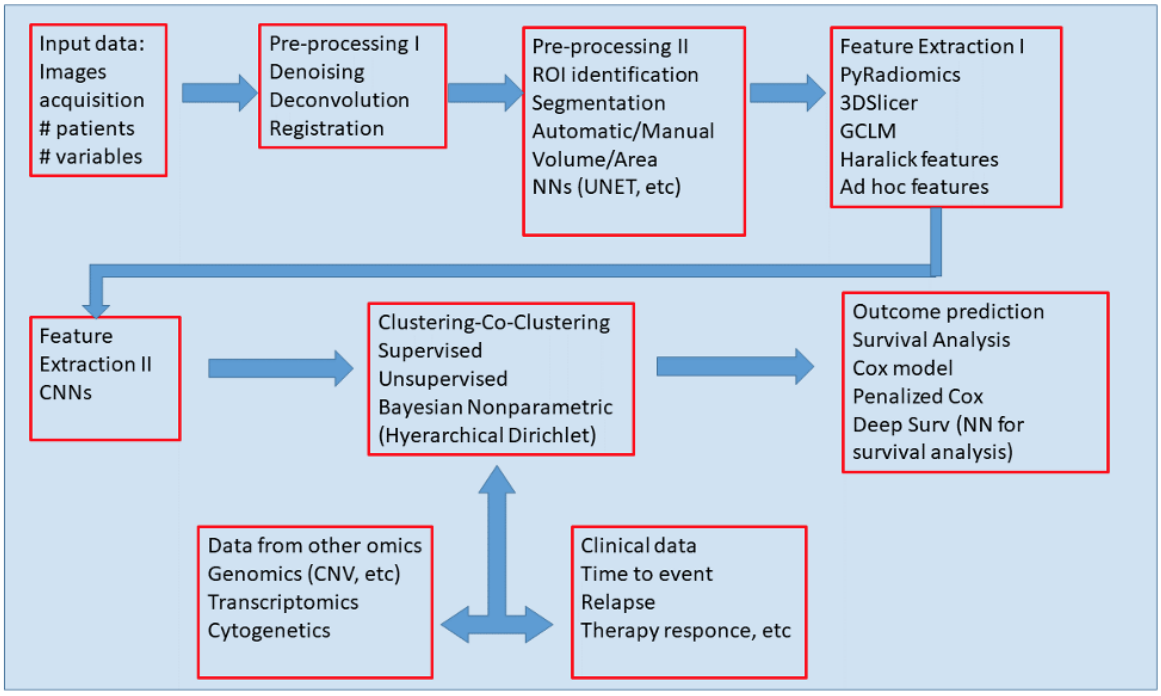


Figure 2. Pipeline for radiomic data. Main steps for integrating imaging with clinical and genomic data.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

1.1 Literature mining

Apart from the standard literature review done routinely by all researchers in the project, we have set up systematic literature mining to ensure that new literature comes to our attention. In literature mining we have searched broader than preprocessing methods, because the types of preprocessing used for specific data types are almost always published together with the machine learning applications they are used with. Therefore, our searches focus on machine learning together with other terms. Not all “machine learning” approaches use the term “machine learning” directly. We have found that terms like “model” or “prediction” are too unspecific to give useful results. The terms “artificial intelligence” and “bayesian” are sufficiently narrow for this field and we have found that having either of these or “machine learning” in the title, abstract or keywords returns the most useful results. The literature mining reported here has been used together with disease specific terms as reported for each disease.

The results of the literature mining have been saved in a database of literature in the standard bibtex literature database format and a few scripts were developed to refine the files downloaded from Scopus and a semi-automated setup for producing the pdf overview of all the literature that use Latex. The bibtex files and the related files are shared in the GitLab repository. The search terms are also available and easy to modify using other or additional keywords or a different time interval. We used Scopus https://www.scopus.com/, because of the possibility to construct sophisticated queries and the ability to save these in several different formats (including bibtex). Unless otherwise stated, only literature from the last four years (since 2017) with search words occurring in the title, abstract or keywords was considered to limit the number of entries.

The literature mining was performed independently for the three different diseases covered here and the lists contain a few papers multiple times. These are typically review-type papers covering several haematological diseases and therefore found in two or three of the searches. The results of the searches are attached in Appendix 2.

Abstracts were also mined for words and words pairs of special interest for the project. For this the same searches were performed except that the limitation of publication year was removed. The abstracts were downloaded in plain text and processed the following way:

1. All non-letters characters were replaced by a space and all uppercase letters were changed to lowercase
2. Words shorter than three characters were deleted
3. For each word the first five letters were used as a key
4. The number of all the keys were counted and a list of single keys occuring more than two times and less than 0.5*N times were selected, where N is the number of abstracts.
5. For each selected key, also the number of times it occurs followed by another key was recorded

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

The list of words and word pairs are quite long. By manual inspection, we found some relevant terms telling what sort of machine learning or analysis methods are used and what data types are considered. These results are displayed in the table below.

Some interesting information can be extracted. For instance, it is evident that neural networks and support vector machines are popular for all three diseases and that mRNA and miRNA analyses are mainly used for MDS. However, this analysis should mainly be used as an inspiration for future work.

	Key	**MDS**	**MM**	**SCD**
Abstracts		116	177	73
Single words		1124	1429	829
Word pairs		801	1427	409
Single words				
clustering	clust	8	17	0
SNP	snp	0	10	8
mutation	mutat	29	16	11
Bayesian	bayes	56	82	16
AUC	auc	7	26	0
pathway	pathw	3	34	6
mRNA	mrna	21	4	0
	mrnas	3	4	0
miRNA	mirna	18	7	0
Word pairs				
logistic regression	logis-regre	3	7	3
neural network	neura-netwo	8	22	8
support vector machine	suppo-vecto	3	9	13
random forrest	rando-forre	0	0	0
feature extraction	featu-extra	0	0	3
segmentation algorithm	segme-algor	3	0	4
gradient boosting	gradi-boost	0	3	0
decision tree	decis-tree	0	6	0
component analysis	compo-analy	3	4	0

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

hazard ratio	hazar-ratio	0	19	0
decision support	decis-suppo	4	6	0
single cell	singl-cell	0	7	0

Table 1. Results of the automatic literature mining procedure for single and word pairs.

2 Use case Myelodysplastic Syndrome

Myelodysplastic syndromes (MDS) are clonal hematopoietic disorders characterized by peripheral blood cytopenia and increased risk of evolution into acute myeloid leukemia (AML). The natural history of MDS is heterogeneous ranging from conditions with a near-normal life expectancy to forms close to AML, with a short life expectancy. Disease-related risk is currently assessed by International Prognostic Scoring System (IPSS/IPSS-R) based on percentage of bone marrow blasts, number of peripheral blood cytopenias and presence of specific clonal cytogenetic abnormalities. While IPSS/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. The development of MDS is driven by mutations on genes involved in RNA splicing, DNA methylation, chromatin modification, transcriptional regulation, and signal transduction. Somatic mutations occur in the genomes of hematopoietic stem cells (HSC) at a low but detectable frequency during normal DNA replication, inducing genomic instability with increased risk of acquiring additional mutations. A number of genomic studies (2) found that the most frequent mutations were in three chromatin-related genes: DNMT3A, TET2 and ASXL1 and in genes encoding for RNA splicing factors (SF3B1, SRSF2). All these mutation frequencies increase with age. Other gene mutations are in DNA methylation (DNMT3A, TET2, IDH1/2), chromatin modification (EZH2, ASXL1), transcriptional regulation (RUNX1), signal transduction (KRAS, CBL). Gene mutations have been reported to influence survival and risk of disease progression in MDS, and the evaluation of the mutation status may add significant information to currently used prognostic scores. For instance, we found (2) that SF3B1 mutations were independent predictors of favorable prognosis, while driver mutations of ASXL1, SRSF2, RUNX1, TP53 and EZH2 genes were associated with a reduced probability of survival. Chromosomal abnormalities also contribute to MDS pathophysiology and cytogenetic testing. For example, presence of duplications, deletions and/or insertions are also important in MDS diagnosis. In myeloid malignancies, classifications on the basis of clinical and morphological criteria could be complemented by introducing genomic features that are closer to the disease biology and capture better clinical-pathological entities. Here, we aim at:

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

1. improving disease classification, using genomic patterns associated with clinical features/outcomes. Classification is important because, for each MDS subtype, the risk of evolution into AML and the risk of death is significantly different.

2. defining novel prognostic models to predict the overall survival and risk of leukemic evolution in order to drive the treatment at individual patient level, e.g. patients with a poor prognosis can be treated with more aggressive treatments, while patients with long life expectancy can even be left without treatment.

2.1 Dataset description

A preliminary dataset of 2,043 patients is already available and fully described in (*2*) and in the supplementary files. The dataset includes comprehensive clinical data at diagnosis, mutational screening by targeted NGS at diagnosis and information on overall survival and leukemic evolution. The data has been anonymized and provides both clinical and genomic data. Specifically, data are structured according to the following main tables:

- **Main dataframe**: it contains one row per patient, with the following information
  - Demographic (patient label, sex, age at data collection)
  - Clinical (prognostic scores, hematochemical variables, bone marrow and comorbidity scores)
  - Cytogenetics (presence/absence of chromosomal alterations)
  - Genomic (presence/absence of targeted mutations across 47 target genes)
  - Outcome (longitudinal data on leukemia-free, overall and AML-adjusted overall survival)

- **DataVariants**: it reports, for each row, information gathered by sequencing at each single mutation at site level for a subset (n=1500) of patients. Specifically:
  - Patient level
  - Variant Allele Frequency (VAF)
  - Tumor depth (number of aligned sequenced reads carrying the variant)
  - Total depth (number of sequenced reads aligned to the reference genome)
  - Gene name
  - Chromosomal position
  - Copy number variation at the genetic locus (e.g. loss, gain)
  - Karyotype

In addition to this preliminary dataset, further multi-omics data will be considered in the next months, i.e. whole exome(WES)/genome(WGS) and RNA-sequencing data, using also public repositories (e.g. GEO, MILE databases).

2.2 Specific preprocessing for MDS

A. Pre-processing of genomic data

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

Common workflows already established in the literature will be used to preprocess sequencing data. **Figure 3** shows the main steps that will be considered for preprocessing the data, considering both what was already done for the pre-processed data and the upcoming DNA and RNA sequencing data. Quality control for the raw reads is applied to analyze sequence quality, sequence length, GC content, the presence of adaptors, ambiguous bases, overrepresented *k*-mers and duplicated reads in order to detect sequencing errors, PCR artefacts or contaminations. Possible adapters are either removed or trimmed. The remaining reads are then aligned to a reference genome. Different algorithms are used for DNA and RNA sequences. Popular examples are BWA-mem for DNA-seq, which is based on the compression algorithm Burrows Wheeler transform, and STAR (Spliced Transcript Alignment to a Reference) for RNA-seq, which uses a seed searching algorithm and allows truncation of reads while mapping. After alignment, for DNA-seq there are further preprocessing steps of quality filtering. Due to PCR amplification during sample preparation and also because of bridge amplification in Illumina sequencing it can happen that a duplicate of reads can be sequenced. Such duplicates can result in erroneous variant calls when detecting SNPs and are therefore nearly always discarded from analysis, using tools like Picard. Reads then can be re-aligned and undergo base-quality recalibration, which adjust for over-/under- estimations of the read quality scores. These steps are implemented as best practices in the Genome Analysis Toolkit (GATK). This tool is also generally used to perform variant calling, which identifies single nucleotide variants and small insertions/deletions in the genome. With a list of called variants it can be daunting to come to grips with what effect, if any, they could have on the individuals harboring them. Tools like ANNOVAR help to retrieve annotations related to the variants, like known variants from 1000 genome project and dbSNP database or scores of the potential effects of genetic variants from Variant Effect Prediction (VEP) or Combined Annotation Dependent Depletion (CADD). More complex variants, like copy number alterations and other structured variants (e.g. translocations, inversions) can be extracted through ad-hoc callers like Sequenza, Delly, Lumpy etc. Further filtering and functional interpretation of the variants are then applied (according to potential pathogenicity, missense mutations, frameshifts, etc.) and these steps will be decided together with the clinicians. On the other hand, counts of the aligned reads from RNA-sequencing data will be extracted through known quantification approaches (e.g. HTseq tool), which allows a normalized number of counts for each gene, taking into account the sequencing depth in each sample. Once the matrix of gene expression is available, with genes by row and samples by column, it is possible to perform common analysis like dimension reduction to check possible outliers or further batch effects or compare expression between groups of samples and extract genes showing statistically significant differential expression. Finally, through biological annotations from publicly available databases like Gene Ontology or KEGG, it is possible to group

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

relevant genes according to their biological functions or specific pathways by performing enrichment analysis.

Considering the available dataset described above, the strategy was to perform targeted multiplexed amplicon-based sequencing (Illumina, San Diego, CA, USA) starting from genomic DNA; the resulting libraries were sequenced on Illumina platforms (NextSeq500) in paired-end mode. Median sequencing depth was 2107x. Variants with a VAF lower than 0.01 and/or variants with a coverage <200x were filtered out. Functionally annotated variants were then also excluded based on the information retrieved from public databases (dbSNP, gnomAD) and the expected germline allele frequency. Single nucleotide polymorphisms (SNP) were annotated according to the NCBI dbSNP (http://www.ncbi.nlm.nih.gov/snp; Build 137) and gnomAD (http://gnomad.broadinstitute.org; gnomAD r2.0.1) databases. The remaining variants were considered as possible somatic mutations and their pathogenic value was evaluated in order to differentiate known and putative pathogenic mutations from variants of unclear significance. Details are reported in (*2*).

B. Data cleaning and homogenization of clinical and post-processed genomic features

Considering the main dataset described above, before performing classification/ clustering and survival analysis, the reported features undergo simple steps for homogenizing and cleaning the dataset. Specifically, we used two different pre-processing steps:

- *Pre-processing for clustering*. Starting from the main dataset, the pipeline extracts 2 dataframes: 1) control_db, containing the NGS-related columns and optionally also the cytogenetic-related columns, 2) target, containing columns associated with the outcome (Leukemia-Free, Overall, and Event-Free survival). This preprocessing avoids imputation, therefore the pipeline removes the NGS columns containing missing values, and, if present, the rows (patients) with missing values in their cytogenetic columns. The importance of optionally dealing with cytogenetic information is given by the fact that the acquisition of this kind of data can fail, and it is very invasive for the patient. Therefore, the pipeline performs two identical, parallel analyses: one including cytogenetic features and excluding patients with failed acquisition, and the other excluding cytogenetic features and including all patients in the analysis. Dimension reduction is also performed through Uniform Manifold Approximation and Projection (UMAP)(*3*) in order to improve and stabilize clustering performance. UMAP is a manifold learning algorithm that builds a lower dimensional embedding of input data based on the nearest neighbors' topology. The main advantages of UMAP are the preservation of the global structure with respect to similar algorithms

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

(e.g. t-SNE), the superior running time performance and no computational restriction on embedding dimension.

- *Pre-processing for survival analysis*. The related script performs simple data cleaning operations: exclusion of some of the clinical variables not used for predictions due to high number of missing values, data homogenization and imputation, using either mean or median values. No dimension reduction is applied.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

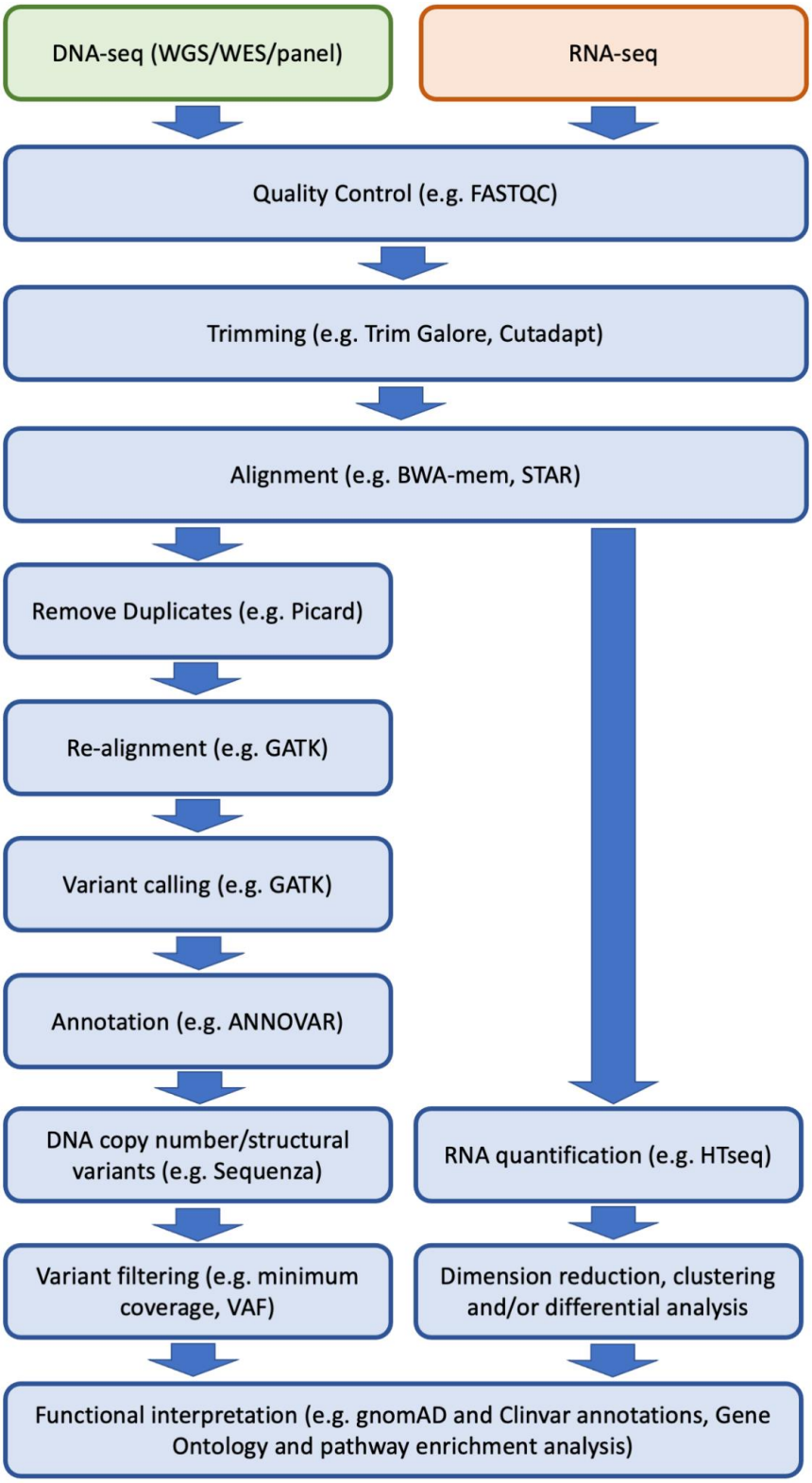


Figure 3. Preliminary pipeline to be used to pre-process sequencing data.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

2.3 Literature mining

Literature related to this section was mined as described earlier, using the additional terms "Myelodysplastic syndrome" and MDS. The abbreviation MDS can mean other things, such as the common dimensionality reduction method “multidimensional scaling” and “Movement Disorder Society”, so the abbreviation is only allowed in the title and we have excluded literature mentioning Parkinson’s disease. The search string is as follows (exact match is indicated with curly braces):

(TITLE-ABS-KEY ("machine learning") OR TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY ({AI}) OR TITLE-ABS-KEY (bayesian)) AND (TITLE-ABS-KEY ("Myelodysplastic syndrome”) OR TITLE ({MDS}) AND NOT TITLE-ABS-KEY (parkinson)) AND PUBYEAR > 2016

The search returned 55 matches and are included as an appendix (and available in the GitLab repository). Most of the papers are highly relevant for the project.

2.4 Software

The software developed on this use case is described in the Gitlab repository of the consortium under the following directory (https://gitlab.com/genomed4all/wp6/preprocessing/LiteratureMining). To improve the readability, we also add it as an appendix.

3 Use case Multiple Myeloma

Multiple Myeloma (MM) is a plasma cell neoplasm which is characterized by the proliferation of monoclonal plasma cells, mainly in the bone marrow. These mutant plasma cells gradually replace the normal plasma cells in the Bone Marrow (they leave less space to the normal plasma cells), and they produce an abnormal antibody called M protein, that may cause anemia, excessive bleeding, decreased ability to fight infection (by impaired immune function), tumors, kidney damage, bone destruction (bone pain and osteolytic lesions, with consequent increase risk of fractures) and also to hypercalcemia.

The exact cause of MM has not yet been identified and the mutational profile is extremely heterogeneous and it varies from person to person. From a pathophysiological point of view, high-risk cytogenetic damages include deletions and translocations on specific chromosomes (chr 1 and 11). There are some specific mutations that have been identified as genetic risk factors, (e.g. genes such as ATM, BRAF, CCND1, DIS3, FAM46C, KRAS, NRAS, TP53 and others) but MM is not thought to be a hereditary disease. Copy Number Alterations (CNA) are other relevant genomic measurements that can be used for the prediction of disease trajectories

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

3.1 Dataset description

Dataset from Bologna University (S.Orsola Hospital)

A cohort of 80 subjects (numbered from 1 to 80) affected by Multiple Myeloma is already available. For each patient, 2,409 measures of Copy Number (CN) are provided through SNP arrays (see section *3.2 Specific preprocessing for MM* below). **Figure 4** displays some examples from the resulting dataset. Each CN refers to one, or more than one, gene (e. g. CN-4 corresponds to GABRD and PRKCZ, while CN-6 to SKI).

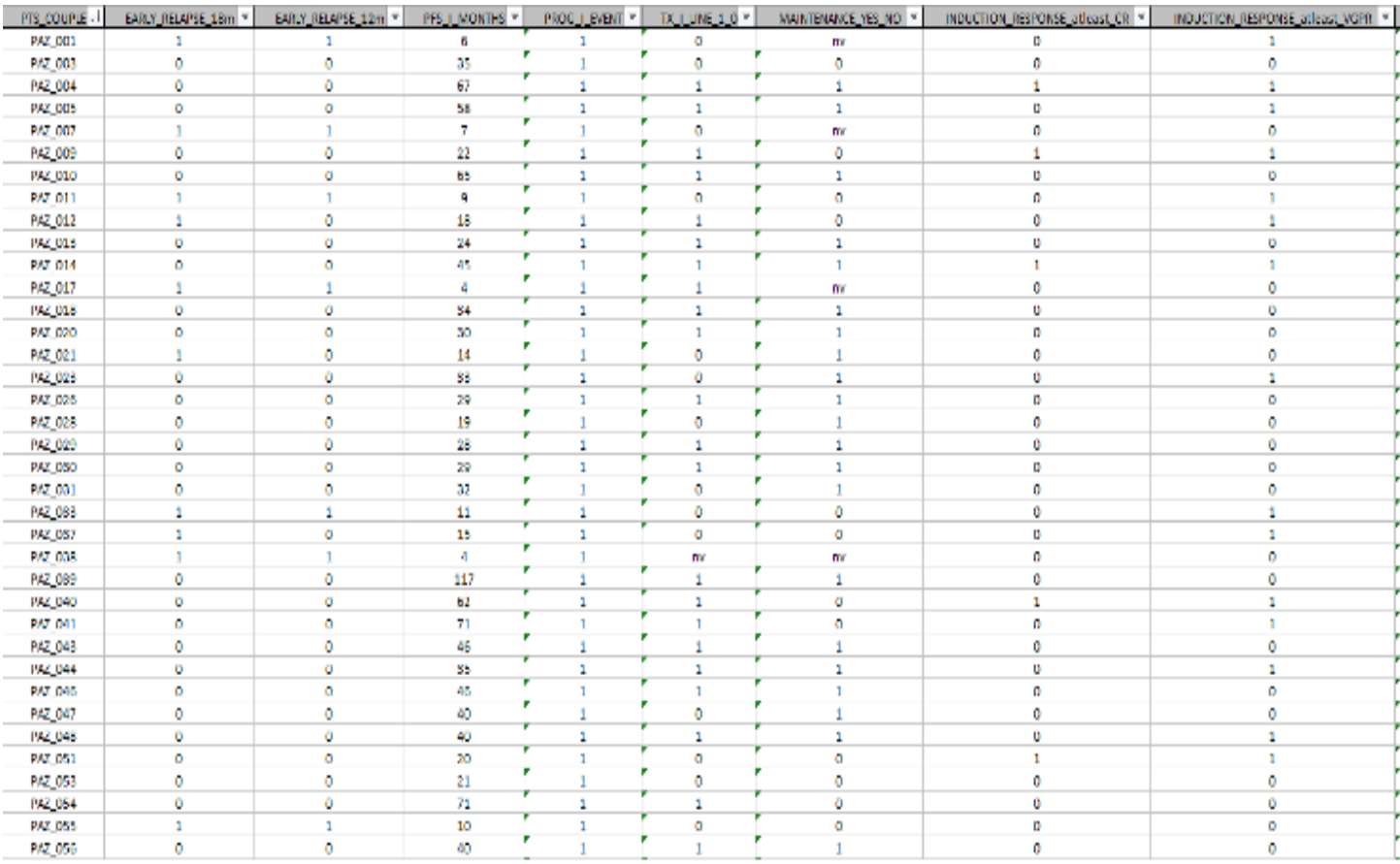
Clinical data are also associated with the copy number measurements (**Figure 5**). 45 patients were treated with transplantation while 32 not. Moreover, 37 patients followed a maintenance therapy. All underwent relapse, with 20 patients in early relapse, i.e. within 18 months (8 of them within 12 months). Patients differ according to therapy response: 40 patients resulted in at least Very Good Partial Remission (VGPR) for induction therapy and 12 of them at least Complete Remission (CR). The aim of this study was to identify subjects with a common copy-number evolution between these two stages of the disease.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

X2395	X2396	X2397	X2398	X2399	X2400	X2401	X2402	X2403	X2404	X2405	X2406	X2407
1,96184	1,96184	1,96184	1,96184	1,96184	1,96184	1,96184	1,96184	1,96184	1,96184	1,96184	1,96184	1,96184
1,992981	1,992981	1,992981	1,992981	1,992981	1,992981	1,992981	1,992981	1,992981	1,992981	1,992981	1,992981	1,992981
1,881721	1,881721	1,881721	1,881721	1,881721	1,881721	1,881721	1,881721	1,881721	1,881721	1,881721	1,881721	1,881721
1,994733	1,994733	1,994733	1,994733	1,994733	1,994733	1,994733	1,994733	1,994733	1,994733	1,994733	1,994733	1,994733
1,914773	1,914773	1,914773	1,914773	1,914773	1,914773	1,914773	1,914773	1,914773	1,914773	1,914773	1,914773	1,914773
2,003852	2,003852	2,003852	2,003852	2,003852	2,003852	2,003852	2,003852	2,003852	2,003852	2,003852	2,003852	2,003852
1,718897	1,718897	1,718897	1,718897	1,718897	1,718897	1,718897	1,718897	1,718897	1,718897	2,070638	2,070638	1,885053
2,000745	2,000745	2,000745	2,000745	2,000745	2,000745	2,000745	2,000745	2,000745	2,000745	2,000745	2,000745	2,000745
2,87911	2,87911	2,87911	2,87911	2,665102	2,665102	2,665102	2,665102	2,665102	2,665102	2,665102	2,665102	2,665102
0,903529	0,903529	0,903529	0,903529	0,903529	0,903529	0,903529	0,903529	0,903529	0,903529	0,903529	0,903529	0,903529
1,993268	1,993268	1,993268	1,993268	1,993268	1,993268	1,993268	1,993268	1,993268	1,993268	1,993268	1,993268	1,993268
1,918325	1,918325	1,918325	1,918325	1,918325	1,918325	1,918325	1,918325	1,918325	1,918325	2,030476	2,030476	2,030476
2,004684	2,004684	2,004684	2,004684	2,004684	2,004684	2,004684	2,004684	2,004684	2,004684	2,004684	2,004684	2,004684
1,18032	1,18032	1,18032	1,18032	1,18032	1,18032	1,18032	1,18032	1,18032	1,18032	1,18032	1,18032	1,18032
1,981853	1,981853	1,981853	1,981853	1,981853	1,981853	1,981853	1,981853	1,981853	1,981853	1,981853	1,981853	1,981853
1,949649	1,949649	1,949649	1,949649	1,949649	1,949649	1,949649	1,949649	1,949649	1,949649	1,949649	1,949649	1,949649
1,849598	1,849598	1,849598	1,849598	1,849598	1,849598	1,849598	1,849598	1,849598	1,849598	1,849598	1,849598	1,849598
1,030962	1,030962	1,030962	1,030962	1,030962	1,030962	1,030962	1,030962	1,030962	1,030962	1,030962	1,030962	1,030962
1,970194	1,970194	1,970194	1,970194	1,970194	1,970194	1,970194	1,970194	1,970194	1,970194	1,970194	1,970194	1,970194
1,038249	1,038249	1,038249	1,038249	1,038249	1,038249	1,038249	1,038249	1,038249	1,038249	1,038249	1,038249	1,038249
0,907895	0,907895	0,907895	0,907895	0,907895	0,907895	0,907895	0,907895	0,907895	0,907895	0,907895	0,907895	0,907895
0,920847	0,920847	0,920847	0,920847	0,920847	0,920847	0,920847	0,920847	0,920847	0,920847	0,920847	0,920847	0,920847
2,05102	2,05102	2,05102	2,05102	2,05102	2,05102	2,05102	2,05102	2,05102	2,05102	2,05102	2,05102	2,05102
2,032688	2,032688	2,032688	2,032688	2,032688	2,032688	1,463948	1,463948	1,463948	1,463948	1,463948	1,463948	1,463948
1,916876	1,916876	1,916876	1,916876	1,916876	1,916876	1,916876	1,916876	1,916876	1,916876	1,916876	1,916876	1,916876
1,999675	1,999675	1,999675	1,999675	1,999675	1,999675	1,999675	1,999675	0,962701	0,962701	0,962701	0,962701	0,962701
1,96007	1,96007	1,96007	1,96007	1,96007	1,96007	1,96007	1,96007	1,96007	1,96007	1,96007	1,96007	1,96007
1,962789	1,962789	1,962789	1,962789	1,962789	1,962789	1,962789	1,962789	1,962789	1,962789	1,962789	1,962789	1,962789
2,084198	2,084198	2,084198	2,084198	2,084198	2,084198	2,084198	2,084198	2,084198	2,084198	2,084198	2,084198	2,084198

Figure 4: Screenshot from the Copy Number dataset. Each row is a patient, while each column is the Copy Number levels of some portion of the genome.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**



PTS_COUPLE_ID	EARLY_RELAPSE_18m	EARLY_RELAPSE_12m	PFS_I_MONTHS	PROG_I_EVENT	TX_I_LINE_1_0	MAINTENANCE_YES_NO	INDUCTION_RESPONSE_atleast_CR	INDUCTION_RESPONSE_atleast_VGPR
PAZ_001	1	1	6	1	0	nv	0	1
PAZ_003	0	0	35	1	0	0	0	0
PAZ_004	0	0	67	1	1	1	1	1
PAZ_005	0	0	58	1	1	1	0	1
PAZ_007	1	1	7	1	0	nv	0	0
PAZ_009	0	0	22	1	1	0	1	1
PAZ_010	0	0	65	1	1	1	0	0
PAZ_011	1	1	9	1	0	0	0	1
PAZ_012	1	0	18	1	1	0	0	1
PAZ_015	0	0	24	1	1	1	0	0
PAZ_014	0	0	41	1	1	1	1	1
PAZ_017	1	1	4	1	1	nv	0	0
PAZ_018	0	0	84	1	1	1	0	0
PAZ_020	0	0	30	1	1	1	0	0
PAZ_021	1	0	14	1	0	1	0	0
PAZ_025	0	0	88	1	0	1	0	1
PAZ_026	0	0	29	1	1	1	0	0
PAZ_028	0	0	19	1	0	1	0	0
PAZ_029	0	0	28	1	1	1	0	0
PAZ_030	0	0	29	1	1	1	0	0
PAZ_031	0	0	32	1	0	1	0	0
PAZ_033	1	1	11	1	0	0	0	1
PAZ_037	1	0	15	1	0	0	0	1
PAZ_038	1	1	4	1	nv	nv	0	0
PAZ_039	0	0	117	1	1	1	0	0
PAZ_040	0	0	62	1	1	0	1	1
PAZ_041	0	0	71	1	1	0	0	1
PAZ_043	0	0	45	1	1	1	0	0
PAZ_044	0	0	55	1	1	1	0	1
PAZ_045	0	0	45	1	1	1	0	0
PAZ_047	0	0	40	1	0	1	0	0
PAZ_048	0	0	40	1	1	1	0	1
PAZ_051	0	0	20	1	0	0	1	1
PAZ_053	0	0	21	1	0	0	0	0
PAZ_054	0	0	71	1	1	0	0	0
PAZ_055	1	1	10	1	0	0	0	0
PAZ_056	0	0	40	1	1	1	0	0

Figure 5: Screenshot from the Clinical dataset. Each row is a patient, while each column represents clinical information about the treatment and the response of that patient.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

For each patient, CN were measured at two different time points: at the diagnosis of the disease and after the relapse, and they were named Copy Number Diagnosis (CND) and Copy Number Relapse (CNR), respectively. The plots of CND vs CNR (the so-called evolutionary trajectories) are shown in **Figure 6**. By visual inspection of these plots, we can identify different trends. For example, a situation where CND and CNR are the same, will produce the bisector of the first and third quadrant. A situation where CNR is on average greater than CND, will produce a line with an angle greater than 45 degrees, etc. Hence it is possible to group patients according to their differences between CND and CNR. The figure 6 reports some of these trajectories, and it is possible to notice that patients can be grouped according to the evolutionary trajectories (we do not report all the 80 trajectories, for the sake of brevity and conciseness). There are patients with a stable trajectory (S), as in **Figure 6a**, while others with linear trajectory (L), **Figure 6b**. Also, patients with branched (B) and drifted (D) trajectories are present, **Figure 6c** and **6d**, respectively. The hematologists, who provided the data, established two different empirical classifications, namely High-Risk (H-R, which considers only a subset of genes) and Genomic (G, that takes into account all the genes for which a CN measure is available), to patients' profiles. A classification consists in assigning a class (S, L, B or D) to each subject, as shown in **Figure 4**. Hence, the organization of CN data in these "evolutionary trajectories" is a preprocessing process that generates a 2409×2 matrix for each patient.

These matrices can be used as input for a modified Dirichlet process clustering algorithm. Briefly speaking, the Dirichlet process is the conjugate prior of an infinite nonparametric discrete distribution as the Dirichlet distribution is the conjugate prior of the categorical multivariate distribution. The modified Dirichlet process clustering method we are using is an update of the classical Dirichlet process clustering algorithm, corrected for using continuous measurements, as those generated by CN. The details of the method are reported in (*4*). Basically, by using this method, we surmise to cluster patients by using a 2D mixture of Gaussian distributions.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

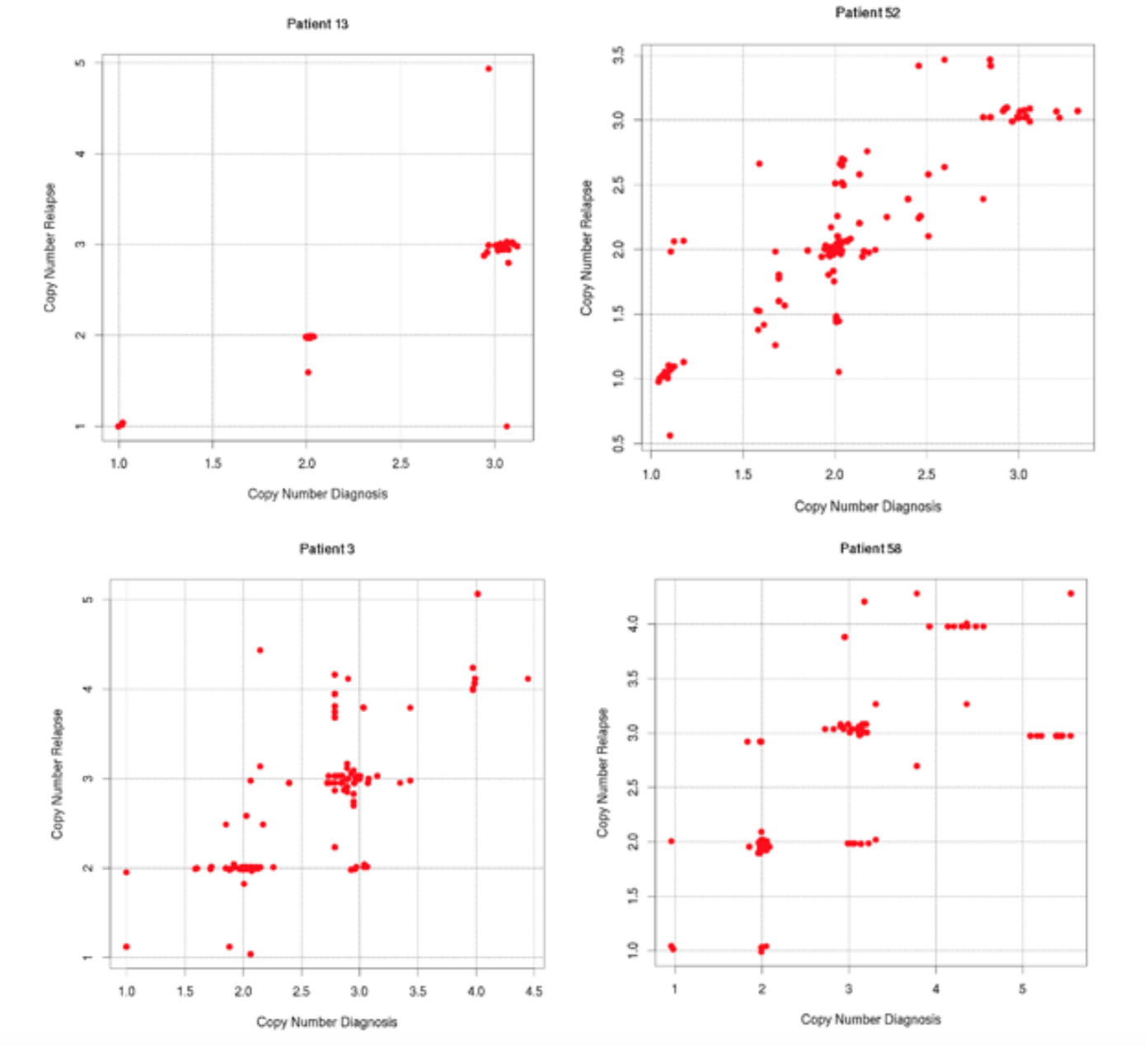


Figure 6 Examples of patients' evolutionary trajectories. (a) Stable trajectory, patient 13; (b) Linear trajectory, patient 52; (c) Branched trajectory, patient 3; (d) Drifted trajectory, patient 58

A further preprocessing was performed by rounding all copy number values to the first decimal place and then represented in a 2D space by applying UMAP(*3*), a manifold learning algorithm for dimensionality reduction, which has been tested using several values of the input parameters in order to verify the robustness of the results.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

Subsequently, groups of patients with similar copy-number profiles at diagnosis – i.e. points that were close in the 2D space obtained through UMAP – were identified via DBSCAN(*5*), a non-parametric density-based clustering algorithm.

In addition to genomic, cytogenetic and other Molecular Biology based measurements, the progression of MM will be measured by Medical Imaging techniques, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) combined with Computed Tomography (CT), with the aim of building "hybrid" prediction models (i.e. using genomic and imaging features) pursuing a "radiomic approach".

Dataset from MMRF CoMMpass (Relating Clinical Outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile)

Study Description

The primary objective of this observational study is to identify the molecular profiles and clinical characteristics that define subsets of myeloma patients during the course of the disease. Understanding the molecular basis of cancer is a critical step toward devising the most effective treatment of the patient as an individual. The promise of molecular targeted therapeutics and personalized cancer care has been demonstrated in breast and lung cancer and chronic myeloid leukemia. However, similar examples of success in multiple myeloma have not been achieved despite extensive basic research as well as clinical advances. What is well understood is that myeloma is a heterogeneous disease with great genetic and epigenetic complexity. Therefore, there remains a critical need to understand myeloma patient biology in the context of current patient care. The objective of this longitudinal study is to identify patient subgroups and phenotypes defined by molecular profiling and clinical features. These profiles will enable a better understanding of mechanisms of disease, drug response and patient relapse. Ultimately the study is intended to drive successful drug development and patient care in multiple myeloma.

Study Design

Study Type: Observational prospective study with a cohort of 1154 participants

The official title is: **A Prospective, Longitudinal, Observational Study in Newly Diagnosed Multiple Myeloma (MM) Patients to Assess the Relationship Between Patient Outcomes, Treatment Regimens and Molecular Profiles**. The study started in July 2011 and will be completed (in term of patient enrollment and measurements) in September 2023. This is a prospective observational study in patients with symptomatic multiple myeloma who have not yet initiated therapy for their disease. The Outcome Measures are divided in primary and secondary.

Primary Outcome Measures:

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

1. Molecular profiles and clinical characteristics that define subsets of myeloma patients at initial diagnosis and at relapse of disease. [Time Frame: Baseline to 8 years.]
2. Standard clinical and laboratory assessments. Genomic tests (DNA and RNA sequencing, etc.) on bone marrow aspirates obtained at baseline, suspected complete response, and relapse/progression.

Secondary Outcome Measures:

1. Response rates [Time Frame: Up to one year after baseline.] IMWG criteria: stringent complete response, complete response, very good partial response, partial response, no response.
2. Survival rates [Time Frame: Five to eight years after baseline] Progression-free survival and overall survival
3. Bone disease assessed radiographically [Time Frame: Baseline and during five to eight years of follow-up]
4. Health-related quality of life [Time Frame: Baseline and during five to eight years of follow-up] EORTC QLQ-C30 and QLQ-MY20
5. Resource utilization [Time Frame: Baseline and during five to eight years of follow-up] Hospitalizations and ER visits
6. Severe adverse events [Time Frame: Five to eight years] Severe/CTCAE grade 3-4 adverse events (checklist)

The Eligibility Criteria are:

1. Ages Eligible for Study: 18 Years and older (Adult, Older Adult)
2. Sexes Eligible for Study: All
3. Accepts Healthy Volunteers:No
4. Sampling Method: Non-Probability Sample
5. Study Population: Newly diagnosed, symptomatic, multiple myeloma, candidates for systemic treatment

Inclusion Criteria:

1. Patient is at least 18 years old.
2. Patient has been diagnosed with symptomatic MM with measurable disease that includes at least one of the following:
3. Serum M protein ≥ 1g/dl Urine M protein ≥ 200 mg/24 hrs Involved free light chain level ≥ 10 mg/dl and an abnormal serum free light chain ratio (<0.26 or >1.65).
4. The patient is a candidate for systemic therapy that includes an IMiD® (e.g., lenalidomide, pomalidomide, thalidomide) and/or proteasome inhibitor (e.g., bortezomib, carfilzomib) as part of the initial regimen.
5. No more than 30 days from baseline bone marrow evaluation as per this protocol to initiation of first-line therapy.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

6. Patient has read, understood and signed informed consent.

Exclusion Criteria:

1. Patient is already receiving systemic therapy for MM (a single dose of bisphosphonates and up to 100 mg total dose of dexamethasone or equivalent corticosteroids are permitted prior to registration on study).
2. Patient had another malignancy within the last 5 years (except for basal or squamous cell carcinoma, or in situ cancer of the cervix).
3. Patient is enrolled in a blinded clinical trial for the first-line treatment of multiple myeloma. Patients may be enrolled in subsequent clinical trials as long as continued access to data and tissue, as per this protocol, is not prohibited.

3.2 Specific preprocessing for MM

A. SNP arrays and sequencing data

Considering the data described above, for each patient, SNPs array raw data (CEL files) were analysed with a bioinformatic pipeline including Rawcopy, ASCAT and GISTIC 2.0 tools, aimed at calling and mapping clonal and subclonal CNAs (Copy Number Alterations) across the whole genome. Specifically, raw CEL files processing was performed using Rawcopy v1.1 R Package (*6*). The significance threshold used for segmentation was set at 10−7. Rawcopy analysis was used in order to normalize Affymetrix arrays, extract quality metrics and obtain raw logR and BAF signals for each SNP array probe. Quality metrics were also produced by using Chromosome Analysis Suite (ChAS) v3.3. We kept only samples that pass all quality thresholds defined as: RawCopy MAPD < 0.23, ChAS MAPD < 0.25 and ChAS QC < 10.00. The raw logR and BAF tracks of all samples that passed the quality checks were used as the input for ASCAT v2.5.2 (*7*). This analysis produced a genomic copy number track per patient, adjusted for its relative computed normal cell contamination level. This step removes the effect of imperfect enrichment of tumor cells, enabling the detection and quantification of subclonal CNAs tumor fractions. ASCAT samples with ploidy > 3.5, reflecting an ambiguous possible whole genome duplication event, were refitted to match a diploid state for simplicity of analysis. In addition, Broad Institute GISTIC v2.0 tool (*8*) was employed to identify a set of focal genomic regions (i.e., covering <25% of chromosome arm), with a non-random confluence of highly frequent, small CNAs, covering well-known tumour suppressor genes and oncogenes, widely regarded as relevant in MM biology (e.g. TP53, RB1, MYC, CKS1B). A complete callset for each sample and each chromosome arm was built, keeping in consideration both broad arm-level CNAs calls plus any CNA detected in a focal region.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

B. Pre-processing for the MMRF-COMMPASS data

From a general point of view, the pre-processing of these data is organized as shown in Fig.3. A more detailed description of all the used pipelines is reported in an appendix document (MMRF_CoMMpass_IA13_Methods.pdf).

C. Pre-processing of MRI and PET imaging data

The preprocessing for the PET and MRI imaging data regarding the MM will be performed by using the pipeline in Fig.2. 8F-Fluorodeoxyglucose (FDG) positron emission tomography (PET)/computed tomography (CT) is currently the standard technique to define minimal residual disease (MRD) status outside the bone marrow (BM) in patients with multiple myeloma (MM)(*9*). More specifically for the PET measurements a semi-automated pipeline for the image segmentation can be used, in order to identify the Region Of Interest (ROI). To extract the radiomic features, the *Pyradiomics* Python library(*10*), based on 3D Slicer(*11*), can be used. In addition to these features, also the GLCM-derived features, such as the famous Haralick features(*12*). After the feature extraction, the feature selection will be performed with different techniques, such as LASSO regression, Genetic Algorithm and other methods based on penalized regression and Neural Networks. The MRI imaging data will be pre-processed by using the same methodology, based on 3DSlices, FreeSurfer, Pyradiomics and Feature selection.

3.3 Literature mining

The abbreviation MM for multiple myeloma is not specific enough and returns too many irrelevant matches. We therefore only consider matches with the full name of the disease. In this case the search is

(TITLE-ABS-KEY ("machine learning") OR TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY ({AI}) OR TITLE-ABS-KEY (bayesian)) AND TITLE-ABS-KEY ("Multiple Myeloma") AND PUBYEAR > 2016

It returned 115 matches, but one irrelevant one was removed manually, because the list of authors was a full page, leaving 114 entries in the list. The large number of matches is unfortunately partly (but not only) due to a higher fraction of low-relevance matches.

3.4 Software

The information about the third software packages used in MM use case is described in the Gitlab repository of the consortium under the following directory:
https://gitlab.com/genomed4all/wp6/preprocessing/mm/-/blob/main/README.md

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

4 Use case Sickle Cell Disease

Sickle cell disease (SCD) is a monogenic disease caused by a point mutation in the β-globin gene (HBB). This mutation causes hemoglobin to have an increased affinity for itself. Under normal conditions, high oxygen cells look ok and the patient only has a slight anemia, but without oxygen red blood cells become deformed and stiff, losing the ability to carry oxygen. In high-income countries, SCD is typically diagnosed at birth.

Sickle cell disease is an inherited red blood cell disorder in which a mutation in the oxygen-carrying protein hemoglobin causes a rigid sickle-shape of red blood cells. This leads to a decrease in the number of healthy red blood cells to carry oxygen throughout the body. Moreover, the normal shaped red blood cells are flexible and can move easily through blood vessels, while the sickle-shaped red blood cells can occlude the small blood vessels and cause a decrease or a block of blood flow and oxygen to parts of the body.

This disease is heterogeneous in its presentation, with differences in the rate and severity of complications even within a single genotype. Even patients with the most severe genotype may vary in their clinical presentation from being continuously admitted for the management of acute complications to rarely requiring medical care. The symptoms of SCD can include Anemia, Pain, Swelling of hands and feet, Vision problems, Stroke and Silent infarcts. In particular, silent infarcts (SI) are defined by an MRI signal abnormality of at least 3 mm in one direction and visible on two views on FLAIR T2-weighted images in a patient with normal neurological examination.

4.1 Data Description

A prospective study is going to be conducted, including the following clinical information and measurements:

- Measurements from Oxygenscan: a device where a laser points on red blood cells to see how they are stretched out allowing measurements for the elongation index of the cells and red blood cells stiffness. This instrument is used to find the point of sickling (POS), which is a highly reproducible measure and it is defined as the elbow point on the curve “elongation index” vs. “oxygen tension”. The larger the POS, the worse the outcome. For patients with SCD, POS values are typically in the range [10,100].
- Measurements of C-Reactive Protein (CRP): this protein is used as a proxy for quantifying the inflammation, which is a main modifier of SCD pathophysiology and it is correlated with higher mortality. In addition, several SNPs involved with CRP measurements are already known in the literature(*13*).

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

- Relevant clinical outcomes for SCD like microalbuminuria, retinopathy, osteonecrosis, ischemic stroke, acute chest syndrome, priapism, vaso-occlusive crisis, hemorrhagic stroke, heart failure and liver disease.

The unmet needs on SCD are:

1. SCD is a disease with high phenotypical heterogeneity
2. There are multiple outcomes that we can consider. Among them we focus, with radiomics, on the risk of developing SCI
3. The final aim is to define a personalized predictive model based on integration of genomics/radiomics and clinical information as a basis of therapeutic decision making in these disorders (*from Proposal*)

A definition of clinical aims to be performed is:

1. Localization of lesions and “Feature Extraction”
2. Correlation between SCI and clinical/Hematological parameters
3. Prediction of risk of developing SCI in the future

Some useful features are:

1. Site of lesions (map of the brain)
2. Site of lesions in relation with vascular territories
3. Total volume of lesions (mm^3)
4. Number of lesions
5. Asymptomatic lacunar infarctions

A paradigmatic example is the **Definition of SCI (Silent Cerebral Infarction) used in published studies coming from Padova center.** Silent infarcts (SI) were defined as an MRI signal abnormality of at least 3 mm in one direction and visible on two views on FLAIR T2-weighted images in a patient with normal neurological examination. Volume of ischemic lesions was calculated after manually drawing the signal abnormalities on FLAIR images [Σ Area lesions x (slice thickness + interslice gap)] using a dedicated software (MedStation®)(*14*, *15*).

DEFINITION OF SCI (Silent Cerebral Infarction)

These are defined as the presence of abnormalities on a magnetic resonance imaging (MRI) scan consistent with cerebral infarction (T-2 weighted and FLAIR imaging) without a clinical history or abnormalities on physical examination that are consistent with a previous stroke (*16*). MRI lesions have to be at least 3 mm in diameter in children (*17*),

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

whereas in adults a more restrictive definition is sometimes used which includes a lesion measuring at least 5 mm on MRI.

For these upcoming data, the planned pre-processing pipeline to be used will be similar to those described for the general Radiomics (section 1) case. For imaging data, a set of MRI images from 109 patients affected by SCD is already available from Padova University. Each of these patients undergoes Magnetic Resonance exams over time to monitor the course of the disease. The distribution of the number of exams per patient is 1/2/15 (min/median/max). The exam consists of different MRI sequences: T1 weighted and FLAIR, as in **Figure 7**. In all, the dataset includes 296 MR exams; an expert neurologist has reported lesions in 146 cases. The age distribution at the time of the exam is 1/10/25 years (min/median/max), as reported in Figure 8. The aim of these exams is the identification of white matter (WM) lesions like silent infarcts. This kind of lesion looks like a hyperintense WM area as in **Figure 7**. We will use the FLAIR sequence (axial direction with 5mm of slice spacing) to identify the lesion. The high-resolution T1 Weighted image is acquired to perform skull stripping and atlas registration.

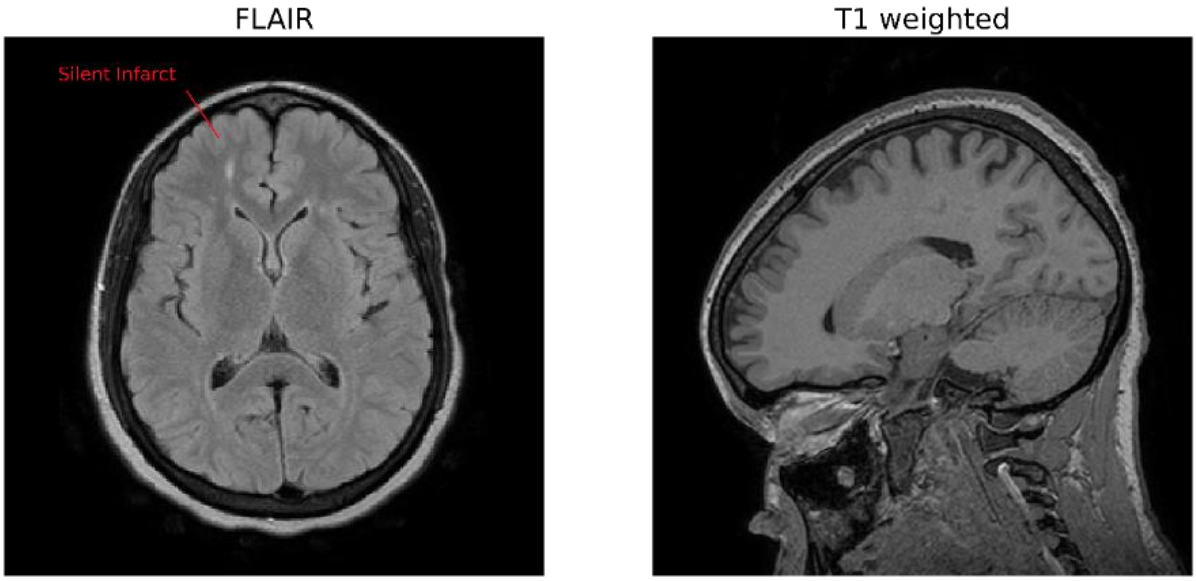


Figure 7. Example of MRI data from SCD patients affected by silent infarct. Left: FLAIR image (axial view) with the presence of a silent infarct. Right: T1 weighted image in Sagittal view.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

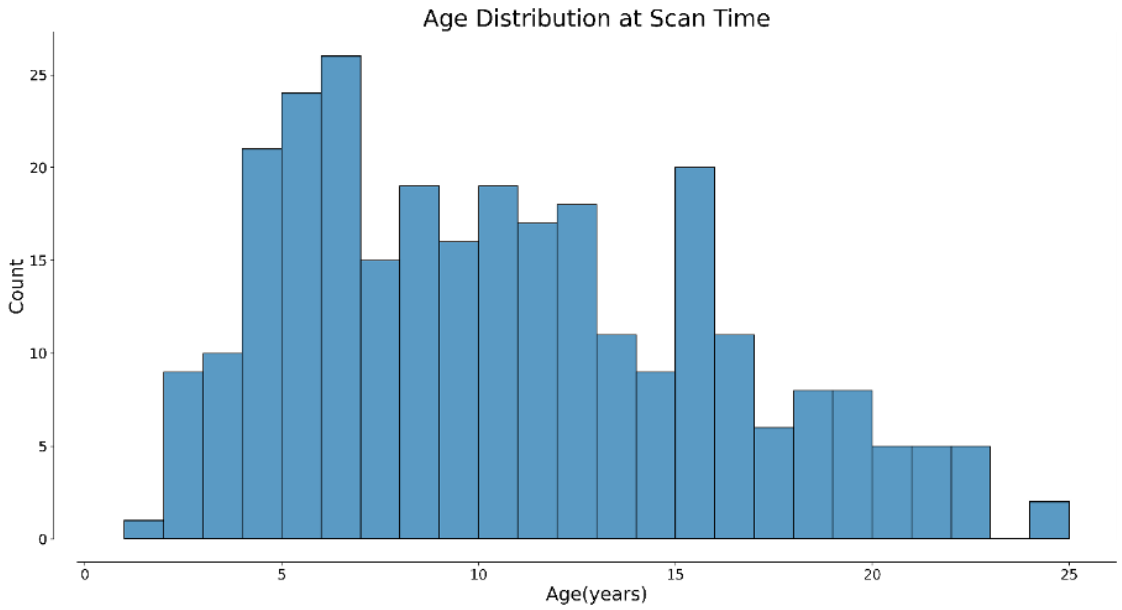


Figure 8. Histogram of the distribution of patient age at the exam time.

To process these data and extract relevant features from the identified areas (i.e. number of lesions, dimension and position), an automated pipeline for the identification of the WM lesions has been developed (Figure 9). In particular, the pre-processing step includes the skull stripping, the image registration to an atlas and the bias field correction, performed using FreeSurfer software. For the segmentation step, we plan to use a combination of 3DSlicer (*18*) and ITK (*19*). In the end, for the feature extraction, we will use PyRadiomics Software (*10*). This approach will ensure the robustness and reproducibility of the feature extraction step. An expert neurologist will check the segmentation quality.

Figure 9. MRI Pipeline flow chart. High-resolution T1 weighted images are provided as input, which undergo skull stripping and atlas registration as preprocessing steps. The pre-processed FLAIR images are then used for the lesion segmentation and the extraction of main features like lesion dimension and position.

4.2 Literature mining

This is challenged by other meanings of the abbreviation SCD, such as “sudden cardiac death”, “stromal cellular density”, and “subjective cognitive decline”, so we only used the full term “sickle cell disease”. The full search query was

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

(TITLE-ABS-KEY ("machine learning") OR TITLE-ABS-KEY ("artificial intelligence") OR TITLE-ABS-KEY ({AI}) OR TITLE-ABS-KEY (bayesian)) AND TITLE-ABS-KEY ("sickle cell disease") AND PUBYEAR > 2016

The search returned 47 articles included in the Appendix.

4.3 Software

The information about the third software packages used in SCD use case is described in the Gitlab repository of the consortium under the following directory: https://gitlab.com/genomed4all/wp6/preprocessing/scd/-/blob/main/README.md

5 Appendices

The deliverable from this work package contains four appendices:

1. Appendix 1: The code used for the literature mining, added as a pdf with the name "A1_literature_mining_code.pdf".
2. Appendix 2: The results (list of papers) obtained from the literature mining search. It is added as a pdf named "A2_literature_mining_description.pdf"
3. Appendix 3. The methods used COMPASS study. It is added as a pdf named "A3_COMPASS_methods.pdf".
4. Appendix 4. Code used to preprocess the MDS use-case. It is a pdf named "A4_MDS_code.pdf"

6 References

1. H. Kang, The prevention and handling of the missing data. *Korean J. Anesthesiol.* **64**, 402–406 (2013).

2. M. Bersanelli, E. Travaglino, M. Meggendorfer, T. Matteuzzi, C. Sala, E. Mosca, C. Chiereghin, N. Di Nanni, M. Gnocchi, M. Zampini, M. Rossi, G. Maggioni, A. Termanini, E. Angelucci, M. Bernardi, L. Borin, B. Bruno, F. Bonifazi, V. Santini, A. Bacigalupo, M. T. Voso, E. Oliva, M. Riva, M. Ubezio, L. Morabito, A. Campagna, C. Saitta, V. Savevski, E. Giampieri, D. Remondini, F. Passamonti, F. Ciceri, N. Bolli, A. Rambaldi, W. Kern, S. Kordasti, F. Sole, L. Palomo, G. Sanz, A. Santoro, U. Platzbecker, P. Fenaux, L. Milanesi, T. Haferlach, G. Castellani, M. G. Della Porta, Classification and Personalized Prognostic Assessment on the Basis of Clinical and Genomic Features in Myelodysplastic Syndromes. *J. Clin. Oncol.* **39**, 1223–1233 (2021).

3. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018), (available at http://arxiv.org/abs/1802.03426).

4. G. J. Ross, D. Markwick, Dirichletprocess: An R package for fitting complex Bayesian nonparametric models, (available at

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

https://repo.bppt.go.id/cran/web/packages/dirichletprocess/vignettes/dirichletprocess.pdf).

5. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, Others, in *kdd* (1996), vol. 96, pp. 226–231.

6. M. Mayrhofer, B. Viklund, A. Isaksson, Rawcopy: Improved copy number analysis with Affymetrix arrays. *Sci. Rep.* **6**, 36158 (2016).

7. P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Børresen-Dale, V. N. Kristensen, Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).

8. C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, G. Getz, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

9. B. Jamet, C. Bailly, T. Carlier, C. Touzeau, C. Nanni, E. Zamagni, L. Barré, A.-V. Michaud, M. Chérel, P. Moreau, C. Bodet-Milin, F. Kraeber-Bodéré, Interest of Pet Imaging in Multiple Myeloma. *Front. Med.* **6**, 69 (2019).

10. J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. W. L. Aerts, Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **77**, e104–e107 (2017).

11. R. Kikinis, S. D. Pieper, K. G. Vosburgh, in *Intraoperative Imaging and Image-Guided Therapy*, F. A. Jolesz, Ed. (Springer New York, New York, NY, 2014), pp. 277–289.

12. R. M. Haralick, K. Shanmugam, I. 'hak Dinstein, Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).

13. S. Ligthart, A. Vaez, U. Võsa, M. G. Stathopoulou, P. S. de Vries, B. P. Prins, P. J. Van der Most, T. Tanaka, E. Naderi, L. M. Rose, Y. Wu, R. Karlsson, M. Barbalic, H. Lin, R. Pool, G. Zhu, A. Macé, C. Sidore, S. Trompet, M. Mangino, M. Sabater-Lleal, J. P. Kemp, A. Abbasi, T. Kacprowski, N. Verweij, A. V. Smith, T. Huang, C. Marzi, M. F. Feitosa, K. K. Lohman, M. E. Kleber, Y. Milaneschi, C. Mueller, M. Huq, E. Vlachopoulou, L.-P. Lyytikäinen, C. Oldmeadow, J. Deelen, M. Perola, J. H. Zhao, B. Feenstra, LifeLines Cohort Study, M. Amini, CHARGE Inflammation Working Group, J. Lahti, K. E. Schraut, M. Fornage, B. Suktitipat, W.-M. Chen, X. Li, T. Nutile, G. Malerba, J. 'an Luan, T. Bak, N. Schork, F. Del Greco M, E. Thiering, A. Mahajan, R. E. Marioni, E. Mihailov, J. Eriksson, A. B. Ozel, W. Zhang, M. Nethander, Y.-C. Cheng, S. Aslibekyan, W. Ang, I. Gandin, L. Yengo, L. Portas, C. Kooperberg, E. Hofer, K. B. Rajan, C. Schurmann, W. den Hollander, T. S. Ahluwalia, J. Zhao, H. H. M. Draisma, I. Ford, N. Timpson, A. Teumer, H. Huang, S. Wahl, Y. Liu, J. Huang, H.-W. Uh, F. Geller, P. K. Joshi, L. R. Yanek, E. Trabetti, B. Lehne, D. Vozzi, M. Verbanck, G. Biino, Y. Saba, I. Meulenbelt, J. R. O'Connell, M. Laakso, F. Giulianini, P. K. E. Magnusson, C. M. Ballantyne, J. J. Hottenga, G. W. Montgomery, F. Rivadineira, R. Rueedi, M. Steri, K.-H. Herzig, D. J. Stott, C. Menni, M. Frånberg, B. St Pourcain, S. B. Felix, T. H. Pers, S. J. L. Bakker, P. Kraft, A. Peters, D. Vaidya, G. Delgado, J. H. Smit, V. Großmann, J. Sinisalo, I. Seppälä, S. R. Williams, E. G. Holliday, M. Moed, C. Langenberg, K. Räikkönen, J. Ding, H. Campbell, M. M. Sale, Y.-D. I. Chen, A. L. James, D. Ruggiero, N. Soranzo, C. A. Hartman, E. N. Smith, G. S. Berenson, C. Fuchsberger, D. Hernandez, C. M. T. Tiesler, V. Giedraitis, D. Liewald, K. Fischer, D. Mellström, A. Larsson, Y. Wang,

W. R. Scott, M. Lorentzon, J. Beilby, K. A. Ryan, C. E. Pennell, D. Vuckovic, B. Balkau, M. P. Concas, R. Schmidt, C. F. Mendes de Leon, E. P. Bottinger, M. Kloppenburg, L. Paternoster, M. Boehnke, A. W. Musk, G. Willemsen, D. M. Evans, P. A. F. Madden, M. Kähönen, Z. Kutalik, M. Zoledziewska, V. Karhunen, S. B. Kritchevsky, N. Sattar, G. Lachance, R. Clarke, T. B. Harris, O. T. Raitakari, J. R. Attia, D. van Heemst, E. Kajantie, R. Sorice, G. Gambaro, R. A. Scott, A. A. Hicks, L. Ferrucci, M. Standl, C. M. Lindgren, J. M. Starr, M. Karlsson, L. Lind, J. Z. Li, J. C. Chambers, T. A. Mori, E. J. C. N. de Geus, A. C. Heath, N. G. Martin, J. Auvinen, B. M. Buckley, A. J. M. de Craen, M. Waldenberger, K. Strauch, T. Meitinger, R. J. Scott, M. McEvoy, M. Beekman, C. Bombieri, P. M. Ridker, K. L. Mohlke, N. L. Pedersen, A. C. Morrison, D. I. Boomsma, J. B. Whitfield, D. P. Strachan, A. Hofman, P. Vollenweider, F. Cucca, M.-R. Jarvelin, J. W. Jukema, T. D. Spector, A. Hamsten, T. Zeller, A. G. Uitterlinden, M. Nauck, V. Gudnason, L. Qi, H. Grallert, I. B. Borecki, J. I. Rotter, W. März, P. S. Wild, M.-L. Lokki, M. Boyle, V. Salomaa, M. Melbye, J. G. Eriksson, J. F. Wilson, B. W. J. H. Penninx, D. M. Becker, B. B. Worrall, G. Gibson, R. M. Krauss, M. Ciullo, G. Zaza, N. J. Wareham, A. J. Oldehinkel, L. J. Palmer, S. S. Murray, P. P. Pramstaller, S. Bandinelli, J. Heinrich, E. Ingelsson, I. J. Deary, R. Mägi, L. Vandenput, P. van der Harst, K. C. Desch, J. S. Kooner, C. Ohlsson, C. Hayward, T. Lehtimäki, A. R. Shuldiner, D. K. Arnett, L. J. Beilin, A. Robino, P. Froguel, M. Pirastu, T. Jess, W. Koenig, R. J. F. Loos, D. A. Evans, H. Schmidt, G. D. Smith, P. E. Slagboom, G. Eiriksdottir, A. P. Morris, B. M. Psaty, R. P. Tracy, I. M. Nolte, E. Boerwinkle, S. Visvikis-Siest, A. P. Reiner, M. Gross, J. C. Bis, L. Franke, O. H. Franco, E. J. Benjamin, D. I. Chasman, J. Dupuis, H. Snieder, A. Dehghan, B. Z. Alizadeh, Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. *Am. J. Hum. Genet.* **103**, 691–706 (2018).

14. R. Colombatti, E. De Bon, A. Bertomoro, A. Casonato, E. Pontara, E. Omenetto, G. Saggiorato, A. Steffan, T. Damian, G. Cella, S. Teso, R. Manara, P. Rampazzo, G. Meneghetti, G. Basso, M. T. Sartori, L. Sainati, Coagulation activation in children with sickle cell disease is associated with cerebral small vessel vasculopathy. *PLoS One*. **8**, e78801 (2013).

15. M. Montanaro, R. Colombatti, M. Pugliese, C. Migliozzi, F. Zani, M. E. Guerzoni, S. Manoli, R. Manara, G. Meneghetti, P. Rampazzo, F. Cavalleri, M. Giordan, P. Paolucci, G. Basso, G. Palazzi, L. Sainati, Intellectual function evaluation of first generation immigrant children with sickle cell disease: the role of language and sociodemographic factors. *Ital. J. Pediatr.* **39**, 36 (2013).

16. M. R. DeBaun, S. A. Sarnaik, M. J. Rodeghier, C. P. Minniti, T. H. Howard, R. V. Iyer, B. Inusa, P. T. Telfer, M. Kirby-Allen, C. T. Quinn, F. Bernaudin, G. Airewele, G. M. Woods, J. A. Panepinto, B. Fuh, J. K. Kwiatkowski, A. A. King, M. M. Rhodes, A. A. Thompson, M. E. Heiny, R. C. Redding-Lallinger, F. J. Kirkham, H. Sabio, C. E. Gonzalez, S. L. Saccente, K. A. Kalinyak, J. J. Strouse, J. M. Fixler, M. O. Gordon, J. P. Miller, M. J. Noetzel, R. N. Ichord, J. F. Casella, Associated risk factors for silent cerebral infarcts in sickle cell anemia: low baseline hemoglobin, sex, and relative high systolic blood pressure. *Blood*. **119**, 3684–3690 (2012).

17. J. F. Casella, A. A. King, B. Barton, D. A. White, M. J. Noetzel, R. N. Ichord, C. Terrill, D. Hirtz, R. C. McKinstry, J. J. Strouse, T. H. Howard, T. D. Coates, C. P. Minniti, A. D. Campbell, B. A. Vendt, H. Lehmann, M. R. Debaun, Design of the silent cerebral infarct transfusion (SIT) trial. *Pediatr. Hematol. Oncol.* **27**, 69–89 (2010).

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

18. A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, R. Kikinis, 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging*. **30**, 1323–1341 (2012).

19. M. McCormick, X. Liu, J. Jomier, C. Marion, L. Ibanez, ITK: enabling reproducible research and open science. *Front. Neuroinform.* **8**, 13 (2014).

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**