Article:
von Asmuth, E.G.J., Halkes, C.J.M., Versluis, J. et al. (2026) An extraction pipeline for analysis of hematopoietic stem cell transplantation data. Bone Marrow Transplantation. ISSN: 0268-3369

https://doi.org/10.1038/s41409-026-02818-z

Cytogenetics pipeline

For free-text cytogenetics, the following pipeline was used:

1. Transform the text to lower case, transliterate non-Latin letters to Latin equivalent.
2. Separate the text into tokens by splitting on spaces, around typography, around numbers, and after content which might be a sex chromosome.
3. Attempt to annotate tokens using International System for Human Cytogenomic Nomenclature (ISCN), combining tokens as needed while being permissive of common omissions and typographical errors.
4. Use heuristics to assign remaining tokens as noise words (common words used in cytogenetic notation that have no special meaning), danger words (negations, indicating some abnormality is absent instead of present), words indicating specific abnormalities, words indicating a complex karyotype, typographical characters, and possible unannotated chromosomes with or without the arm and location specified.
5. If an unannotated possible chromosome with or without an arm is preceded by a specific abnormality, then typography, associate that abnormality with the unannotated chromosome as well (e.g., “+4, 5” is parsed to trisomy of chromosome 4 and 5).
6. If a possible chromosome with or without an arm is associated with a word indicating a specific abnormality (separated only using noise words, other chromosomes and/or typography), assign that abnormality to that chromosome. Prefer words preceding the chromosome indicating an abnormality, except when the chromosome is followed by a dash, then prefer words following it (e.g., trisomy of chromosome 3, 5q- and 7q-deletion is parsed as a trisomy of chromosome 3 and a deletion of 5q and 7q).
7. If the free text is entered next to a checkbox indicating a specific abnormality (e.g., 4, 5 entered next to the checkbox “Trisomy”), assign that abnormality to all unannotated chromosomes.
8. If a danger word is detected in step 4 (e.g., no translocation of chromosome 4), discard all annotations for this free-text entry and process it as unknown.
9. Combine all checkbox and free-text entries for that timepoint and patient.

For step 3, annotation using ISCN, we developed a strategy based on CyDAS[1], an open-source program that can parse ISCN, but adjusted it to be much more permissive of typographical and data entry errors, as long as those still allowed for unambiguous identification of the chromosomal abnormality. An example of the pipeline on ISCN free-text data is provided in figure 1 (simplified, step 1 omitted). An example of the pipeline on non-ISCN free-text data is provided in figure 2.

In step 4, special words were identified through a review of the first 2000 records, and through a search on potential special words and word fragments. Special words in English, German and French were identified.

Figure 1: example of pipeline output on ISCN data, step 1 omitted, output at step 2, at the end of the pipeline, and risk categories.

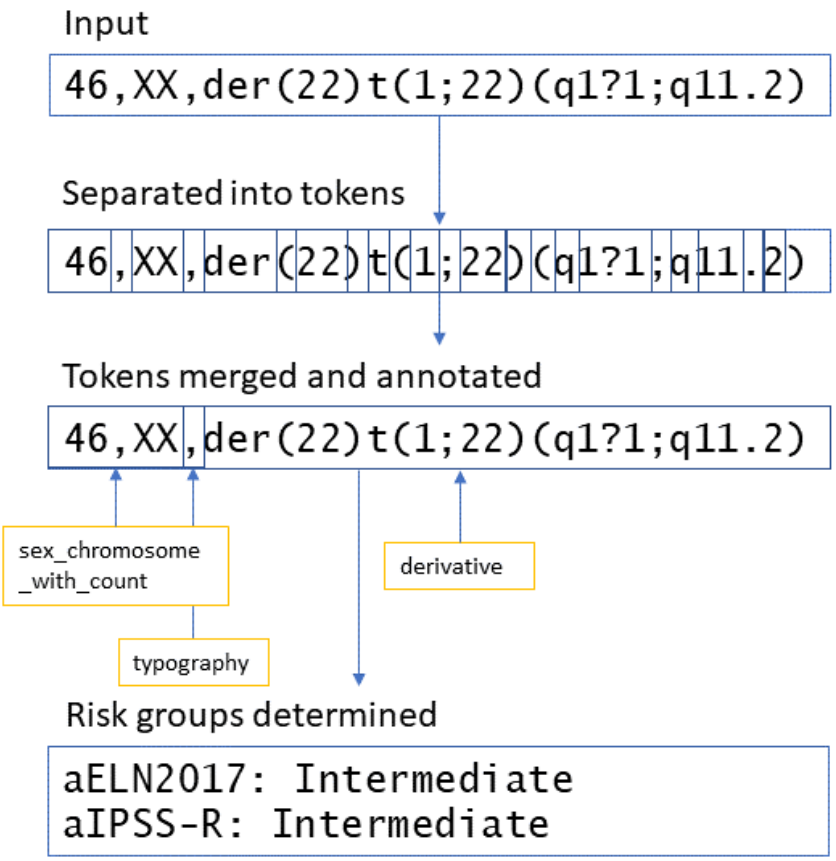
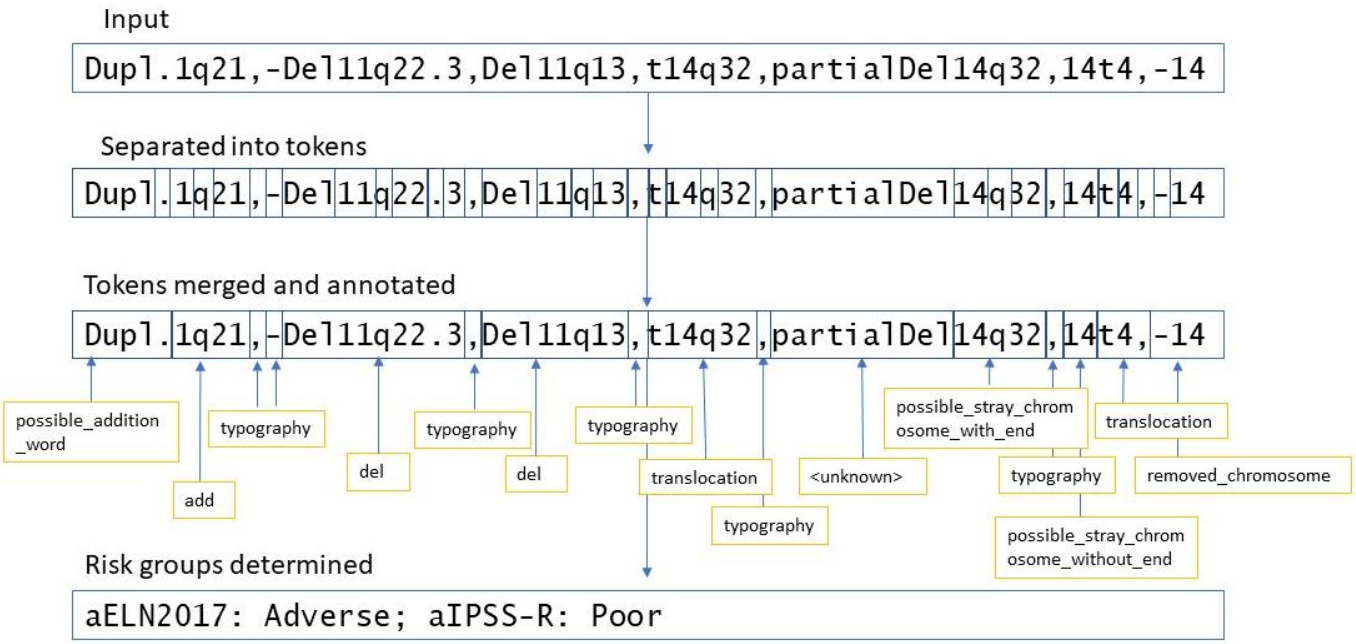


Figure 2: example of pipeline output on non-ISCN free-text data, step 1 omitted, output at step 2, at the end of the pipeline, and risk categories.



The annotated tokens were used to assign the IPSS-R and ELN2017 cytogenetic risk categories for patients with MDS and AML respectively[2,3]. If no token was annotated to be a specific abnormality,

the free-text entry was considered to be uninterpretable. Specific abnormalities were considered present if a derivative chromosome containing that abnormality was present. The karyotype was considered accurate even if an abnormality, or the chromosomes or bands involved in an abnormality, were indicated to be uncertain. If multiple cell lines with different abnormalities were present, this was considered equal to one cell line with those abnormalities combined. If an abnormality was present but it was unknown which chromosome or chromosomes gave rise to that abnormality, it was not assumed to be any specific abnormality, but included to count towards a complex karyotype. Marker chromosomes were not included to count towards a complex karyotype.

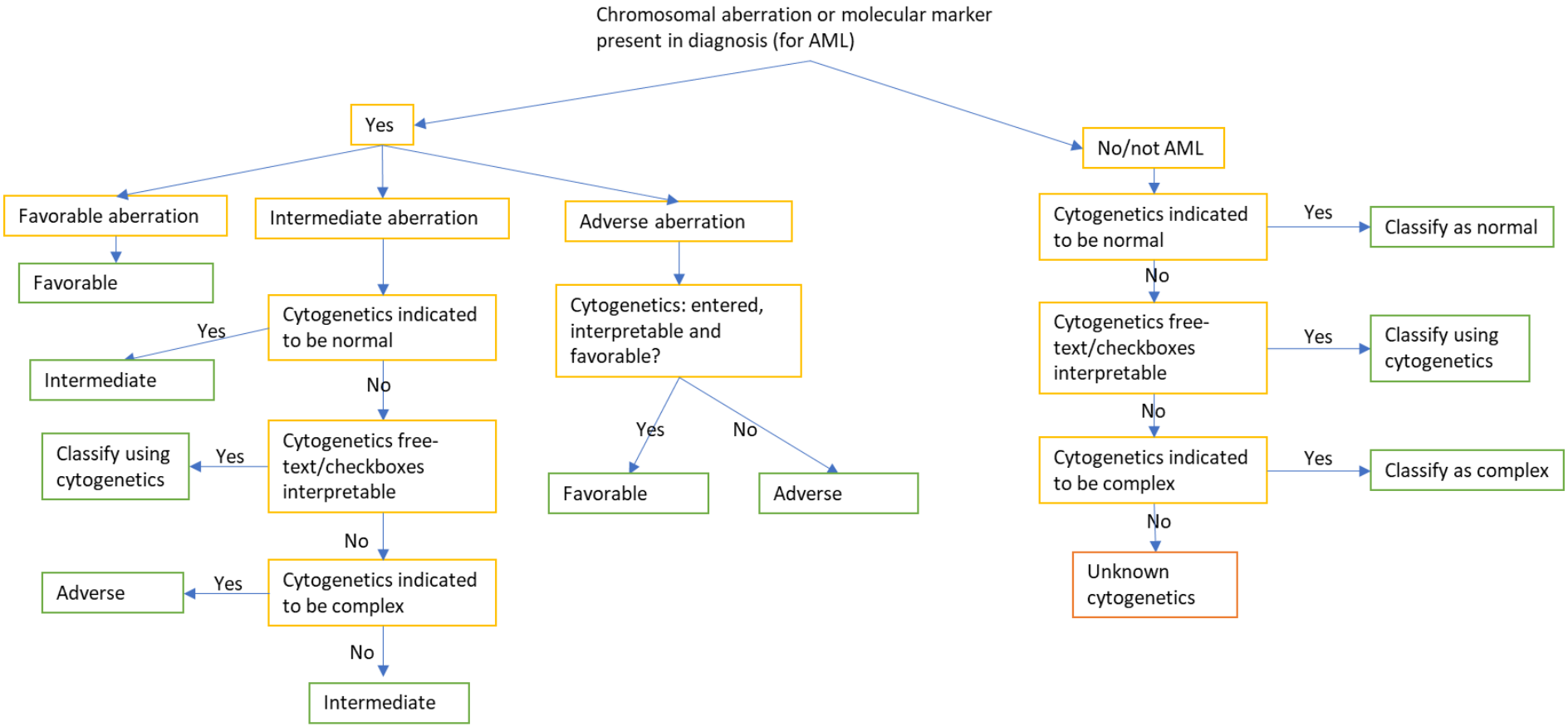
To be compatible with the checkboxes within the EBMT registry data collection forms, which lack band positions for some abnormalities, band position of the abnormality was only considered when available (e.g. t(8;21) was considered to be a *RUNX1-RUNX1T1* translocation for the purposes of ELN2017 risk score assignment if the location was not specified).

A karyotype was assigned to be complex using the definition as entered in the EBMT registry (any 3 or more chromosomal abnormalities), instead of using definitions from risk scores. For the IPSS-R, the “Very poor” risk group was not assigned (defined as 4 or more chromosomal abnormalities), as for some transplants the “Complex karyotype” checkbox was checked but no further details on the number or nature of abnormalities were entered, and assigned the “Poor” risk group (exactly 3 chromosomal abnormalities) even when the number of chromosomal abnormalities was known to exceed 3.

For the ELN2017 risk score, we used the adjusted version as used in the Disease Risk Stratification Score (DRSS)[4], which analyzes molecular markers separately instead of incorporating them into the risk score, does not assign patients to the “Adverse” risk category based on a monosomal karyotype, and assigns patients with t(15;17) to the “Favorable” risk category instead of considering it an entirely separate entity not incorporated into the risk stratification.

Finally, the risk score based on specific cytogenetic abnormalities was integrated with cytogenetics present in the WHO diagnosis code for AML, and with checkboxes indicating if the patient had a normal karyotype or a complex karyotype, using the schema seen in figure 3.

Figure 3: integration of different sources of cytogenetic information



1. Hiller B, Bradtke J, Balz H, Rieder H. CyDAS: a cytogenetic data analysis system. *Bioinformatics* 2005; **21**(7): 1282-3.
2. Dohner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* 2017; **129**(4): 424-47.
3. Greenberg PL, Tuechler H, Schanz J, et al. Revised International Prognostic Scoring System for Myelodysplastic Syndromes. *Blood* 2012; **120**(12): 2454-65.
4. Shouval R, Fein JA, Labopin M, et al. Development and validation of a disease risk stratification system for patients with haematological malignancies: a retrospective cohort study of the European Society for Blood and Marrow Transplantation registry. *Lancet Haematol* 2021; **8**(3): e205-e15.