



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/239337/>

Version: Accepted Version

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, 61 (4). pp. 400-407. ISSN: 0268-3369

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

© 2026 The Authors. Except as otherwise noted, this author-accepted version of a journal article published in Bone Marrow Transplant is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

An extraction pipeline for analysis of hematopoietic stem cell transplantation data

Erik GJ von Asmuth¹, Constantijn JM Halkes², Jurjen Versluis³, Dirk-Jan A Eikema⁴, Emanuele Angelucci⁵, Ali Bazarbachi⁶, Fabio Ciceri⁷, Raffaella Greco⁷, Mette Hazenberg⁸, Krzysztof Katwak⁹, Donal P McLornan¹⁰, Bénédicte Neven¹¹, Antonio M Risitano¹², Mirjam Steinbuch¹³, Anna Sureda¹⁴, John Snowden¹⁵, Arjan C Lankester¹, Hein Putter¹⁶, Liesbeth C de Wreede¹⁶

¹ Willem Alexander Children's Hospital, Leiden University Medical Centre, Leiden, The Netherlands

² Department of Hematology, Leiden University Medical Centre, Leiden, The Netherlands

³ Department of Hematology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands

⁴ EBMT Leiden Study Unit, Leiden, The Netherlands

⁵ Hematology and Cellular Therapy Center, IRCCS Ospedale Policlinico San Martino, Genova, Italy

⁶ Bone Marrow Transplantation Program, Department of Internal Medicine, American University of Beirut Medical Center, Beirut, Lebanon

⁷ Unit of Hematology and Bone Marrow Transplantation, IRCCS San Raffaele Hospital, Milan, Italy

⁸ Department of Hematology, Amsterdam University Medical Centres, Amsterdam, The Netherlands

⁹ Clinical Department of Pediatric BMT, Hematology and Oncology, Wroclaw Medical University, Wroclaw, Poland

¹⁰ Department of Haematology and Stem Cell Transplantation, University College Hospital, London, United Kingdom

¹¹ Pediatric Hematology and Immunology Unit, Necker Hospital for Sick Children, Assistance Publique–Hôpitaux de Paris, Paris, France

¹² Hematology and Hematopoietic Transplant Unit, Azienda Ospedaliera di Rilievo Nazionale "San Giuseppe Moscati" (A.O.R.N. Giuseppe Moscati), Avellino, Italy

¹³ EBMT Executive Office, Utrecht, The Netherlands

¹⁴ Clinical Hematology Department, Institut Català d'Oncologia–L'Hospitalet, Spain

¹⁵ Division of Clinical Medicine, School of Medicine and Population Health, The University of Sheffield, UK

¹⁶ Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands

Corresponding author:

Erik GJ von Asmuth - e.g.j.von_asmuth@lumc.nl

Conflict of Interest: all authors declare to have no relevant conflicts of interest

Abstract

Many clinical studies are based on registry analyses, but exact approaches of data extraction and pre-processing are rarely included, while this is critical for reliability and reproducibility of results. We aimed to develop an open-source data extraction pipeline which generates a ready-to-analyze dataset focused on relevant determinants of outcomes after hematopoietic stem cell transplantation (HSCT). This pipeline was developed using EBMT registry data, including 54,457 allogeneic and 63,651 autologous HSCT procedures. The pipeline determines HLA matching from molecular data, assesses cytogenetic risk for acute myeloid leukemia and myelodysplastic syndrome, processes

molecular markers, assigns the hematopoietic cell transplantation comorbidity index (HCT-CI) based on comorbidities, and maps disease states to simplified categories. We prospectively assessed the recently developed disease risk stratification system (DRSS), showing that the pipeline produces consistent results with previous studies. The hazard ratio correlation between our cohort and the original DRSS derivation cohort was 0.92 with a 2-year AUC of 0.616, indicating similar effects and predictive performance. We aim to establish a new standard by promoting transparent, standardized and uniform extraction of registry data, enhancing reproducibility in registry studies.

Introduction

Hematopoietic stem cell transplantation (HSCT) is a potentially curative high-risk procedure for a variety of malignant and non-malignant indications¹. HSCT can be allogeneic, when stem cells are harvested from a donor, or autologous, for which the patient's own stem cells are used. More than 1.5 million HSCT have been performed globally, with the global transplant rate steadily increasing². Within EBMT centers, about 50,000 transplants are performed yearly, with 42% being allogeneic and 58% being autologous³.

Given the relatively low number of transplants performed in individual centers or within countries, international collaboration and registries are essential for efficiently investigating prognostic factors in current HSCT practice. The EBMT registry employs an electronic database to systematically capture detailed transplantation-related information, including patient data (e.g., comorbidities), disease characteristics (e.g., cytogenetic or molecular diagnostics), transplant-specific data (e.g., conditioning), and donor characteristics (e.g., Human Leukocyte Antigen (HLA) typing)⁴. Since standards in HLA matching, cytogenetic classification, and molecular diagnostics have changed over time⁵⁻⁷, it is crucial to record raw data to allow for the development of updated classification systems and interpretations of these evolving data⁸.

However, data extraction and preparation are significant challenges in any research effort, as data may be available in both interpreted and raw format, and inconsistencies may occur when reinterpreting the raw data. These inconsistencies need to be considered and handled during the extraction process. Data extraction methods are often not clearly explained in published manuscripts, but information about the extraction and preparation process is crucial for reproducibility.

This manuscript describes a new data extraction pipeline for HSCT data in the EBMT aimed at prognostic modelling. The pipeline includes handling of cytogenetic information, including the handling of free-text data in non-standardized formats, determining the number of HLA mismatches including handling of intermediate-resolution HLA typing data and handling patients with multiple donors or multiple diagnoses.

This pipeline was created in R, version 4.4.0, using dplyr and duckdb for data manipulation. For the interpretation of HLA data, the current IPD IMGT/HLA database is used⁹, which is retrieved automatically by code.

Our motivating example is EBMT benchmarking, where survival after first HSCT is compared across centers, while adjusting for differences in case mix. Case mix has previously been modeled using patient, donor and disease factors including the revised disease risk index (rDRI)¹⁰, HCT-specific comorbidity index (HCT-CI)¹¹, donor type, amongst others¹². This causes particular challenges during extraction, as all diseases need to be handled, and no patients should be excluded based on missing or inconsistent data. For different research, different choices could be made, but the foundation laid by this pipeline may be used.

The data was extracted on May 12, 2022 and includes 54,457 allogeneic and 63,651 autologous HSCT procedures performed between 2017 until 2021. The in- and exclusion criteria of EBMT benchmarking were used, meaning first allogeneic and autologous transplants for specific diseases were included, with autologous transplants only including adult patients¹².

Which data is produced by the pipeline, including any modification of source data, is described in table 1. Below, we will discuss in detail how source data is interpreted to come to a readily analyzable data file for all fields where complex interpretation steps are needed.

Diagnosis

The diagnosis for which HSCT is performed is captured in the EBMT registry in a general category first, followed by more detailed data entry fields. Patients can have multiple diagnoses preceding HSCT, as HSCT can be performed for multiple diagnoses. For example, patients with Fanconi anemia have an increased risk of developing Acute Myeloid Leukemia (AML), while both Fanconi anemia and AML can be treated with allogeneic HSCT¹³. The relevance and values of certain prognostic factors, such as cytogenetics, molecular markers and whether a disease is of de novo origin or secondary to another disease or treatment depend on the diagnosis and can only be entered for specific diagnoses.

In our pipeline, for patients with multiple primary diagnoses, the diagnosis most closely preceding the HSCT is considered the diagnosis for which the HSCT was performed, and used to interpret further disease-specific data. De novo or secondary disease types were considered for Acute myeloid leukemia (AML), B-cell non-Hodgkin lymphoma, myelodysplastic syndrome (MDS), myeloproliferative neoplasm (MPN) and MDS/MPN overlap syndromes, consistent with the EBMT registry⁸. AML was defined as secondary if at least one of the following was noted: the diagnosis was indicated to be transformed or of secondary origin, the subcategory of AML was therapy-related myeloid neoplasm, or it was AML with myelodysplasia related changes and was indicated to be preceded by MDS or an MDS/MPN overlap syndrome.

For MDS, MPN, and MDS/MPN overlap syndromes, the World Health Organization (WHO) classification at transplant was available in addition to the disease remission status¹⁴. If the disease had transformed to AML at time of transplant, the indication diagnosis was changed to secondary AML, and processed as such for the purposes of cytogenetic risk classification and further disease stage interpretation.

Disease remission status

Disease remission status is an important prognostic factor in HSCT for malignant disease. In the EBMT registry, disease status is captured at each follow-up moment after HSCT, and the options depend on preceding diseases. For analysis, the number of disease statuses is often reduced, for example by mapping statuses like “stringent complete remission” to “complete remission”

Remission statuses were mapped depending on the diagnosis for which HSCT was performed (table 2). The number of previous remissions was only considered for patients in complete remission (CR). Some reported disease statuses that are inconsistent with the primary diagnosis, for example chronic phase for acute leukemia, were set to Unknown/other. A mapping for each disease and disease remission status is provided in supplemental file 1.

Patient and donor HLA matching

HLA matching between patient and donor is the most important factor in selecting a donor for allogeneic HSCT¹⁵. Molecular HLA typing can be performed with different resolutions. With intermediate-resolution molecular HLA typing, the result may be ambiguous, identifying a small number of candidate alleles¹⁶. Current EBMT donor

selection guidelines recommend matching on 10 alleles, namely the A, B, C, DRB1 and DQB1 HLA loci, whereas the other HLA loci are considered secondary¹⁵. For this project, we only considered matching on these 10 alleles.

Donor relation can provide insights into HLA matching, as siblings may inherit the same chromosome 6 from both parents, making them fully HLA identical, whereas syngeneic siblings are always HLA identical. Within the EBMT registry, donor relation is recorded, the total number of mismatches is recorded for related donors, and the number of HLA mismatches per locus can be recorded. Both related and unrelated donors could previously be specified as matched or mismatched, but currently this distinction is only made for related donors. In whether a donor is considered as matched or mismatched and in the latter case, in how many loci, there is room for interpretation which implies that the same donor-patient pair could be coded differently in different centers.

HLA type can be recorded in the EBMT registry, with ambiguous HLA being allowed. HLA can be entered using serological or molecular typing. Only molecular typing was processed in this project. Choices in interpretation of number of mismatches have to be made.

Specific HLA mismatches can be considered permissible based on studies showing they are not immunogenic and not associated with adverse outcomes¹⁷. In the case of homozygous mismatches, where both patient and donor are homozygous and mismatching, two alleles mismatch, but only one protein differs. In unidirectional mismatches, where either patient or donor are homozygous, a mismatching protein is only foreign to the donor or to the patient. Two unidirectional mismatches can occur in opposing directions, where both patient and donor can only recognize one mismatching protein, but these are different proteins.

In counting the number of HLA mismatches, we counted the number of mismatching HLA proteins. We considered the mismatches C*03:03 to C*03:04 and DRB1*14:01 to DRB1*14:54 permissible, and thus matching^{18,19}.

Unidirectional mismatches were considered to be a mismatch, irrespective of direction. Unidirectional mismatches in opposing directions were considered separate mismatches. In case of both patient and donor being homozygous and mismatching, this was considered a single mismatch, as there was only one mismatching HLA protein. Null alleles, where the HLA protein encoded will not be expressed on the cell wall, were considered equal to homozygosity of the non-null allele. Expression suffixes other than null alleles were not considered when assessing whether alleles matched.

In case of ambiguous HLA typing, any pair of HLA alleles that could be matched was considered matched, except when that required an ambiguous allele to become a null allele. The following formats that indicate ambiguous HLA were parsed: MAC groups (e.g., B*08:YETY), slash-delimited (e.g., B*08:01/B*08:19N/B*08:109), G-groups (e.g., A*01:01:01G) and P-groups (e.g., A*01:01P). XX codes (e.g., A*01:XX) were deemed too inaccurate and handled as missing data. For MAC codes, we retrieved the list of codes on August 21, 2023²⁰, for P- and G-groups we used version 3.52.0 of the IPD-IMGT HLA database²¹.

The number of HLA mismatches was considered known if either the donor was syngeneic or an identical sibling, molecular HLA data was available on all assessed loci, the number of molecularly assessed mismatches was entered for each locus, or the donor HLA was indicated to have been assessed molecularly and to be identical to that of the patient. When multiple variables for matching were available, the number of mismatches as determined on molecular HLA data was preferred over other methods to handle ambiguous typing and permissible mismatches uniformly (supplemental figure 1).

The donor relation was made consistent with the number of molecular mismatches by setting the number of mismatches to 0 in the case of a syngeneic or identical sibling donor. When HLA mismatches were present, the donor type was changed from matched family donor to mismatched family donor, and vice versa if no mismatches were found. The matched unrelated and mismatched unrelated donor categories were merged into unrelated as few patients fell into these categories with missing HLA data. Syngeneic donors were merged with identical sibling as there were few syngeneic donors (table 3).

Multiple donors

Allogeneic transplants can be performed with multiple donors to increase the total number of stem cells transplanted. Most commonly, this is done with cord blood HSCT, where multiple cord blood units of different donors are combined, though it is also sometimes done with different graft sources and donor relations²².

If both donors had identical characteristics, those characteristics were used. If multiple donors with different graft sources or donor relations were used in a single transplant, the level “multiple” was assigned, which was the case in 123 out of 429 transplants with multiple donors for graft source, and 109 out of 429 for donor relation. For donor age and HLA matching, the highest age and highest number of HLA mismatches were used. For sex and CMV

serostatus matching, the transplant was considered mismatched if either donor had a mismatching sex or mismatching CMV serostatus.

Comorbidities

In the EBMT registry, comorbidities are scored according to the HCT-CI, which was developed in 2005 to predict non-relapse mortality after allogeneic HSCT for hematological malignancies. The HCT-CI evaluates specific comorbidities, assigning them points according to their impact on non-relapse mortality¹¹. Patients can be entered as having no comorbidities, and in that case, no data on specific comorbidities are registered. For these patients, it is not entered which comorbidities have been assessed, and thus we need to assume all comorbidities were assessed sufficiently. Differences in comorbidity assessment may result in underestimated risk scores, such as only performing pulmonary function tests in patients with pulmonary complaints.

In our pipeline, when processing comorbidities, if a patient was entered as having no comorbidities, all specific comorbidities were assumed to be absent. If a patient was scored as having comorbidities, but no specific comorbidities were reported, the HCT-CI was set to unknown. Otherwise, the HCT-CI was calculated assuming each comorbidity that was entered as unknown was not present. The pipeline also includes information on specific comorbidities, where these are only set to not present for patients with no comorbidities and thus it is preserved which comorbidities were not reported for patients with comorbidities.

Survival and censoring

Follow-up for surviving patients is recorded at 100 days and yearly post HSCT. Since data exports may occur between these predefined follow-up moments, and mortality is reported continuously, ascertainment bias may arise. This bias occurs if death is reported after the last predefined moment, since the follow-up date would be the last predefined moment for survivors but may occur beyond that for patients who died. This causes an apparent excess in death after the last predefined follow-up moment, which affects survival estimates and models²³.

For that reason, we censored follow-up after the last predefined follow-up moment. For example, if patients are transplanted at the 1st of January 2021, they are censored at the 1st of January 2022 if they have follow-up data beyond that date, since their next follow-up moment (the 1st of January 2023) would lie beyond the data export

date. We assumed follow-up moments up to the 12th of April, 2022 could be recorded to account for delay in data entry and inexact dates of follow-up moments.

Cytogenetics

Cytogenetic abnormalities may occur in malignant and premalignant diseases. In previous prognostic scores for allogeneic HSCT, cytogenetic abnormalities were considered for AML and MDS^{10,24}. In the EBMT registry, karyotype analysis is recorded as not assessed, normal or abnormal, complex (defined as at least 3 cytogenetic abnormalities) or non-complex, and whether the karyotype was monosomal. Cytogenetic abnormalities can be specified by using checkboxes for common abnormalities with the possibility to enter free-text next to each checkbox. In addition to these checkboxes, the full karyotype can be provided in free text. For AML, common cytogenetic abnormalities are also captured in WHO diagnosis codes²⁵.

To analyze cytogenetic risk factors, we developed a custom pipeline to process all cytogenetic data entered in the EBMT registry, and to determine the European LeukemiaNet (ELN) 2017²⁶ and Revised International Prognostic Scoring System (IPSS-R)²⁷ risk scores for AML and MDS respectively. The ELN2017 risk score was adjusted to handle low-risk malignancies as done in the development of the DRSS²⁴. The IPSS-R risk score was adjusted by merging the *Poor* and *Very poor* risk groups, to allow for capturing patients where the exact cytogenetics were missing, but the karyotype was reported to be complex. The most recent cytogenetic assessment preceding transplant was used for risk group assignment.

This pipeline can annotate and classify cytogenetics data entry, both using text and checkboxes. Our work was loosely based on CyDAS²⁸, but adjusted to allow for imperfect data entry. All free text cytogenetic information was manually screened for coding standards and natural language used, and the pipeline was adjusted on the first 2000 cytogenetic entries by comparing researcher-assessed and pipeline-assessed interpretation of cytogenetics, and adjusting the pipeline on mismatches. In addition, the dataset was searched for rare cytogenetic abnormalities such as ring chromosomes and isomeric derivative chromosomes to assess proper handling of these abnormalities. Exact details on the cytogenetics pipeline are described in supplemental file 2.

Disease risk score assignment

To validate the pipeline, we prospectively assessed the DRSS in this dataset. The DRSS assigns risk groups based on disease, disease stage, cytogenetics and molecular markers. The DRSS was developed on previous EBMT registry data (not included in the present analysis) of patients with malignant disease transplanted between 2012 and 2016 and was associated with differences in survival between the risk groups.

To assign risk groups within AML and ALL patients, we extracted data for *FLT3-ITD*, *NPM1*, and *BCR-ABL*. To separate ALL into Philadelphia (Ph) positive, negative or unknown, we considered information from diagnosis code, molecular markers and cytogenetics.

The DRSS was evaluated on patients with similar inclusion criteria as those of the DRSS study (adults, allogeneic HSCT, disease not excluded in DRSS, no missing data in disease and disease status; leading to a cohort size of N=40,552, 74.5% of allogeneic HSCT patients), comparing the original risk group sizes and hazard ratio with those in our cohort. Hazard ratios measured in our cohort were highly consistent with the original derivation cohort (figure 1).

Discriminative performance for the DRSS risk categories, measured by 2-year area under the receiver operating characteristic (AUC) for overall survival, was 0.616 (95% confidence interval: 0.608-0.623) in our cohort, which was identical to 0.616 in derivation²⁴. When using the HRs from the derivation cohort instead of risk categories, performance increased to a 2-year AUC of 0.622 (0.614-0.629), indicating the DRSS is still associated with survival in this more recent cohort.

Cytogenetics were available in more often in our cohort compared to the original cohort (73.6% vs 44.5%). This resulted in differences in risk classification, with 26.4% of AML assigned to the low-risk category, while this was 14.6% in the original cohort (supplemental figure 2). This could be due to both improved data entry in recent years, as well as improved data processing.

Discussion

In this manuscript, we propose an extraction pipeline for HSCT data applicable to both allogeneic and autologous HSCT, where inaccurate, incomplete or inconsistent data is considered. Inaccuracy may be inherent to the measurement technique (e.g., intermediate-resolution HLA typing), to the data capture fields (cytogenetics

checkboxes only containing specific chromosomes), or to the entry of data itself. These inaccuracies are rarely elaborated on, but can affect data quality and subsequent analyses. Proper handling of all available information is crucial for analyses that should not exclude patients based on data availability, such as those performed in benchmarking.

When combining data from multiple registries, which is especially useful for studying rare diseases, it is important that both data collection and extraction are standardized to ensure a uniform definition of variables across datasets. A standardized extraction pipeline can aid such research. The EBMT registry has recently moved to OMOP-CDM, a standardized data format. This means future tools developed for the EBMT registry can also be applied to different registries that support OMOP-CDM, leading to quicker, higher-quality research.

Clinical studies in hematology often have missing data, and excluding cases with missing data leads to statistical inefficiency²⁹. Imputation can be applied to missing data, often under the assumption of missing at random. However, if data is missing due to interpretation steps made during extraction, e.g. by coding cases with multiple diagnoses as missing, this assumption may not hold³⁰. Both excluding and imputing these data could lead to invalid conclusions, thus even difficult-to-code cases should be considered or explicitly excluded. Applying this or similar methodology to clinical registry studies may thus be attractive.

This pipeline is publicly available ([link](#)), and integrated into the Yearly Outcome Review Tool³¹. We anticipate that this pipeline can be a practical tool to perform reproducible research on HSCT data.

Author contributions

EvA, CH, AL, HP and LdW designed the study. EvA created the pipeline. DJE provided the previous benchmarking pipeline as reference material, and gave input on the pipeline. EvA, CH, JV, AL, HP and LdW drafted the manuscript. AL, HP and LdW supervised the study. EA, AB, FC, RC, MH, KK, DM, BN, AR, MS, AS and JS gave input to and critically revised the manuscript.

Funding

This study was funded internally, using an LUMC institutional grant. No external or EBMT funding was used.

Data availability statement

This study used EBMT registry data. Data access requests should follow the procedure outlined on

<https://www.ebmt.org/registry/ebmt-data-sharing>. The research pipeline itself, including accompanying

classification tables, is publicly available on Zenodo at <https://doi.org/10.5281/zenodo.17614459>.

References

1. Copelan EA. Hematopoietic stem-cell transplantation. *N Engl J Med*. 2006;354(17):1813-1826.
2. Baldomero H, Neumann D, Hamad N, Atsuta Y, Sureda A, Iida M, et al. The role of registries in hematological disorders. *Best Pract Res Clin Haematol*. 2024;37(2):101556.
3. Passweg JR, Baldomero H, Ciceri F, Corbacioglu S, de la Camara R, Dolstra H, et al. Hematopoietic cell transplantation and cellular therapies in Europe 2021. The second year of the SARS-CoV-2 pandemic. A Report from the EBMT Activity Survey. *Bone Marrow Transplant*. 2023;58(6):647-658.
4. Aljurf M, Rizzo JD, Mohty M, Hussain F, Madrigal A, Pasquini MC, et al. Challenges and opportunities for HSCT outcome registries: perspective from international HSCT registries experts. *Bone Marrow Transplant*. 2014;49(8):1016-1021.
5. Dohner H, Wei AH, Appelbaum FR, Craddock C, DiNardo CD, Dombret H, et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood*. 2022;140(12):1345-1377.
6. Charrot S, Armes H, Rio-Machin A, Fitzgibbon J. AML through the prism of molecular genetics. *Br J Haematol*. 2020;188(1):49-62.
7. Mangum DS, Caywood E. A clinician's guide to HLA matching in allogeneic hematopoietic stem cell transplant. *Hum Immunol*. 2022;83(10):687-694.
8. The European Society for Blood and Marrow Transplantation. Data Collection; 2024. Available from: <https://www.ebmt.org/registry/ebmt-data-collection>.
9. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. *Nucleic Acids Res*. 2020;48(D1):D948-D955.
10. Armand P, Kim HT, Logan BR, Wang Z, Alyea EP, Kalaycio ME, et al. Validation and refinement of the Disease Risk Index for allogeneic stem cell transplantation. *Blood*. 2014;123(23):3664-3671.
11. Sorrow ML, Maris MB, Storb R, Baron F, Sandmaier BM, Maloney DG, Storer B. Hematopoietic cell transplantation (HCT)-specific comorbidity index: a new tool for risk assessment before allogeneic HCT. *Blood*. 2005;106(8):2912-2919.
12. Saccardi R, Putter H, Eikema DJ, Busto MP, McGrath E, Middelkoop B, et al. Benchmarking of survival outcomes following Haematopoietic Stem Cell Transplantation (HSCT): an update of the ongoing project of the European Society for Blood and Marrow Transplantation (EBMT) and Joint Accreditation Committee of ISCT and EBMT (JACIE). *Bone Marrow Transplant*. 2023;58(6):659-666.
13. Alter BP. Fanconi anemia and the development of leukemia. *Best Pract Res Clin Haematol*. 2014;27(3-4):214-221.
14. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-2405.
15. Carreras E, Dufour C, Mohty M, Kröger N. The EBMT Handbook: Hematopoietic Stem Cell Transplantation and Cellular Therapies: Springer International Publishing; 2019.
16. Little AM, Akbarzad-Yousefi A, Anand A, Diaz Burlinson N, Dunn PPJ, Evseeva I, et al. BSHI guideline: HLA matching and donor selection for haematopoietic progenitor cell transplantation. *Int J Immunogenet*. 2021;48(2):75-109.
17. Petersdorf EW. HLA structure and function in hematopoietic-cell transplantation. *Best Pract Res Clin Haematol*. 2024;37(3):101564.
18. Fernandez-Vina MA, Wang T, Lee SJ, Haagenson M, Aljurf M, Askar M, et al. Identification of a permissible HLA mismatch in hematopoietic stem cell transplantation. *Blood*. 2014;123(8):1270-1278.
19. Xiao Y, Lazaro AM, Masaberg C, Haagenson M, Vierra-Green C, Spellman S, et al. Evaluating the potential impact of mismatches outside the antigen recognition site in unrelated hematopoietic stem cell transplantation: HLA-DRB1*1454 and DRB1*140101. *Tissue Antigens*. 2009;73(6):595-598.

20. National Marrow Donor Program. Allele Code Lists; 2024. Available from: <https://bioinformatics.bethematchclinical.org/hla-resources/allele-codes/allele-code-lists/>.
21. Barker DJ, Maccari G, Georgiou X, Cooper MA, Flicek P, Robinson J, Marsh SGE. The IPD-IMGT/HLA Database. *Nucleic Acids Research*. 2023;51(D1):D1053-D1060.
22. Kekre N, Antin JH. Hematopoietic stem cell transplantation donor sources in the 21st century: choosing the ideal donor when a perfect match does not exist. *Blood*. 2014;124(3):334-343.
23. Motter JD, Segev DL, Massie AB. Don't stop thinking about tomorrow: Ascertainment bias in pre-post design transplant registry studies. *Am J Transplant*. 2022;22(6):1513-1514.
24. Shouval R, Fein JA, Labopin M, Cho C, Bazarbachi A, Baron F, 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-e215.
25. Jaffe R. WHO classification of tumours of haematopoietic and lymphoid tissues. *World Health Organization Classification of Tumours*. 2008:358-360.
26. Dohner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Buchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*. 2017;129(4):424-447.
27. Greenberg PL, Tuechler H, Schanz J, Sanz G, Garcia-Manero G, Solé F, et al. Revised International Prognostic Scoring System for Myelodysplastic Syndromes. *Blood*. 2012;120(12):2454-2465.
28. Hiller B, Bradtke J, Balz H, Rieder H. CyDAS: a cytogenetic data analysis system. *Bioinformatics*. 2005;21(7):1282-1283.
29. Bonneville EF, Schetelig J, Putter H, de Wreede LC. Handling missing covariate data in clinical studies in haematology. *Best Pract Res Clin Haematol*. 2023;36(2):101477.
30. Little RJ. Missing Data Assumptions. *Annual Review of Statistics and Its Application*. 2021;8(Volume 8, 2021):89-107.
31. von Asmuth EGJ, Putter H, Mohseny AB, Schilham MW, Snowden JA, Saccardi R, Lankester AC. Automating outcome analysis after stem cell transplantation: The YORT tool. *Bone Marrow Transplantation*. 2023;58(9):1017-1023.

Data category	Field description	Field name	Example values	Applicability	Processing in pipeline
Diagnosis	Diagnosis (main category)	DISMCLFD	Acute leukemia; Inherited disorders; Lymphoma	Always	Handling of multiple diagnoses; reassignment of transformed MDS, MPN and MDS/MPN overlap.
	Diagnosis (subcategory 1)	L1_code	MDS; Primary immune deficiency; ALL	Always	Data selected based on DISMCLFD
	Diagnosis (subcategory 2)	L2_code	Acute myelomonocytic leukemia; jMML; B lymphoblastic leukaemia/lymphoma, NOS	Multiple diseases	Data selected based on L1_code
	Diagnosis + stage as in DRSS	cat_name_drss	AML, Int cyto, Non(FLT3-ITD/NPM1wt), CR1; Congenital bone marrow failure, not in DRSS	Always	As in DRSS, but with excluded diagnoses handled similarly
Staging and disease-specific data	Remission status	VDISESTA	CR; PR; Upfront	Multiple diseases	Merging of categories, removal of inconsistent information
	Remission number	VNUMSTM_CR	1st; 2nd	Remission status is CR	None
	Molecular remission	VMOLECRE	Yes; No	Disease is acute leukemia; remission status is CR	None
	Disease of secondary origin	VSECORIG	Yes; No	Multiple diseases	Made consistent with L2_code and WHO stage for MDS, MPN and overlap syndromes
	MDS WHO stage	MDSSTAG	RA with excess of blasts-2 (RAEB-2); Childhood myelodysplastic syndrome	Disease is MDS	None
	MDS/MPN overlap WHO stage	MDMPSTAG	Same as at diagnosis (CMML, jMML, etc.); CMML Type II	Disease is MDS/MPN overlap	None
	MPN WHO stage	VMPS	Primary myelofibrosis ; Chronic neutrophilic leukaemia	Disease is MPN	None
	Disease was refractory in the past	REFRPAST	Yes; No	Disease is lymphoma	None
	International prognostic index for NHL	IPIOINDEX_diag	Low risk (0-1 score points); High risk (4 or 5)	Multiple forms of NHL	None
	Prognostic score (MIPI)	PROSCORE_diag	Low risk; Intermediate risk	Disease is mantle cell lymphoma	None
	Total number of lines of therapy	TOTNTHER	2; 3; >4	Disease is lymphoma	None
	International stratification score for multiple myeloma	ISS	I; II; III	Disease is multiple myeloma	None
	Light chain type	VPLCEDS2	IgG; IgA; IgD	Disease is multiple myeloma with light chain secretion	None
	Heavy chain type	VPLCEDS4	Kappa; Lambda	Disease is multiple myeloma with heavy chain secretion	None
Patient data	Age at SCT	age_sct	1.02; 65.4	Always	Calculated
	Performance score	KARNOFSK	100; 90	Always	None
	Patient sex	PATSEX	Male; female	Always	None

Donor and matching data	Age at diagnosis	AGEDIAG	1; 65	Always	Handling of multiple diagnoses
	Time between diagnosis and HSCT	time_diag_hsct	0.03; 10	Always	Calculated
	Previous autologous transplant	previous_auto	Yes; No	Allogeneic SCT	Calculated
	Donor age	donor_age	0; 15; 35	Allogeneic SCT	Set to 0 for cord blood; set to oldest if multiple donors are entered
	Sex match	sex_match	P: Female/D: Male; P: Male/D: Male	Allogeneic SCT	Calculated based on patient and donor sex; if both a male and female donor are used, sex match is mismatched
	CMV match	cmv_match	P: Positive/D: Positive; P: Negative/D: Negative	Allogeneic SCT	Calculated based on patient and donor CMV status; if both a CMV positive and negative donor are used, CMV match is mismatched
	Graft source	STEMCEDO1	CB; BM/PB/Other	Allogeneic SCT	All non-CB grafts merged since choice between BM and PB may be a physician's decision
	Multiple donors used	VMULTDON	Yes; No	Allogeneic SCT	Set to yes if multiple donors were entered
	Donor type	corrected_donor_type	Identical sibling; Unrelated; Multiple	Allogeneic SCT	Made consistent with HLA match; if multiple donors with differing types are used, donor type is set to "Multiple"
Comorbidities	Number of HLA mismatches (/10, molecular matching)	Matching_consensus_10	0; 1; 5	Allogeneic SCT	Calculated; see manuscript
	Number of HLA mismatches (categorical, any resolution)	Matching_consensus_cat	Fully matched ; 1 HLA locus mismatch; >=2 HLA loci mismatch	Allogeneic SCT	Made consistent with donor type and high resolution matching
	HCT-CI	hctci	0; 1	Always	Calculated; see manuscript
Cytogenetics	All HCT-CI components	Multiple	Depending on component	Always	Set to "no comorbidity" level if patient had no comorbidities
	ELN2017 classification, modified according to DRSS	drss_eln_classification	Favorable; Intermediate; Adverse	Diagnosis is AML; Cytogenetic information is available	Calculated; see manuscript
	IPSS-R classification, adjusted for compatibility with EBMT registry	ipssr_classification	Good; Intermediate; Poor	Diagnosis is MDS; Cytogenetic information is available	Calculated; see manuscript
Survival information	Time between SCT and assessment of survival data	years_since_sct	0.2; 1	Always	Calculated; see manuscript
	Survival status	is_dead	True; False	Always	Calculated; see manuscript

Table 1: fields produced by the pipeline, including any modifications done to source data in the pipeline, and the applicability of different fields.

Disease	Coded status	N allogeneic (%)	N autologous (%)
Acute leukemia	CR	18603 (83.2%)	646 (95.6%)
	Not in CR	3113 (13.9%)	15 (2.2%)
	Unknown/other	635 (2.8%)	15 (2.2%)
CML	Chronic phase	545 (65.5%)	
	Accelerated phase	95 (11.4%)	
	Blast crisis	154 (18.5%)	
	Unknown/other	38 (4.6%)	
Other chronic leukemia	CR	226 (41.7%)	17 (58.6%)
	PR	192 (35.4%)	12 (41.4%)
	Relapse/progression	102 (18.8%)	0 (0%)
	Unknown/other	22 (4.1%)	0 (0%)
Lymphoma	CR	1062 (48.4%)	10331 (60.3%)
	PR	691 (31.5%)	5279 (30.8%)
	Relapse/progression	365 (16.6%)	997 (5.8%)
	Unknown/other	76 (3.5%)	535 (3.1%)
MDS/MPN	CR	1429 (18.7%)	
	PR	1523 (20.0%)	
	Nonresponsive	1985 (26.0%)	
	Untreated	2047 (26.8%)	
	Unknown/other	650 (8.5%)	
Plasma cell disorders	CR	16 (21.9%)	5426 (20.3%)
	VGPR	Not applicable*	11392 (42.6%)
	PR	34 (46.6%)	8025 (30.0%)
	Relapse/progression	20 (27.4%)	1219 (4.6%)
	Unknown/other	3 (4.1%)	673 (2.5%)
Bone marrow failure	Not applicable	1804 (100%)	
Hemoglobinopathies	Not applicable	1072 (100%)	
Histiocytic disorders	Not applicable	204 (100%)	
Inherited disorders	Not applicable	1397 (100%)	

Table 2: acceptable disease states per main disease category (percentages calculated per disease category). *For patients with plasma cell disorders undergoing an allogeneic SCT, VGPR was merged with PR due to small numbers.

Donor type	0 mismatches	At least 1 mismatch	Missing data on at least 1 locus	Assumed truth in case of inconsistency
Identical sibling	11421 (71.8%)	168 (1.1%)	4320 (27.2%)	Donor type
Syngeneic	72 (66.7%)	4 (3.7%)	32 (29.6%)	Donor type
Matched other relative	528 (46.0%)	159 (13.9%)	460 (40.1%)	HLA data
Mismatched relative	248 (2.5%)	5267 (52.4%)	4543 (45.2%)	HLA data
Matched unrelated	646 (95.8%)	13 (1.9%)	15 (2.2%)	Always changed to unrelated
Mismatched unrelated	52 (20.4%)	189 (74.1%)	14 (5.5%)	Always changed to unrelated
Unrelated	18001 (56.4%)	5499 (17.2%)	8410 (26.4%)	Not applicable
Missing donor type	11 (7.1%)	6 (3.9%)	137 (89.0%)	Not applicable

Table 3: association between HLA matching, donor type and assumed truth

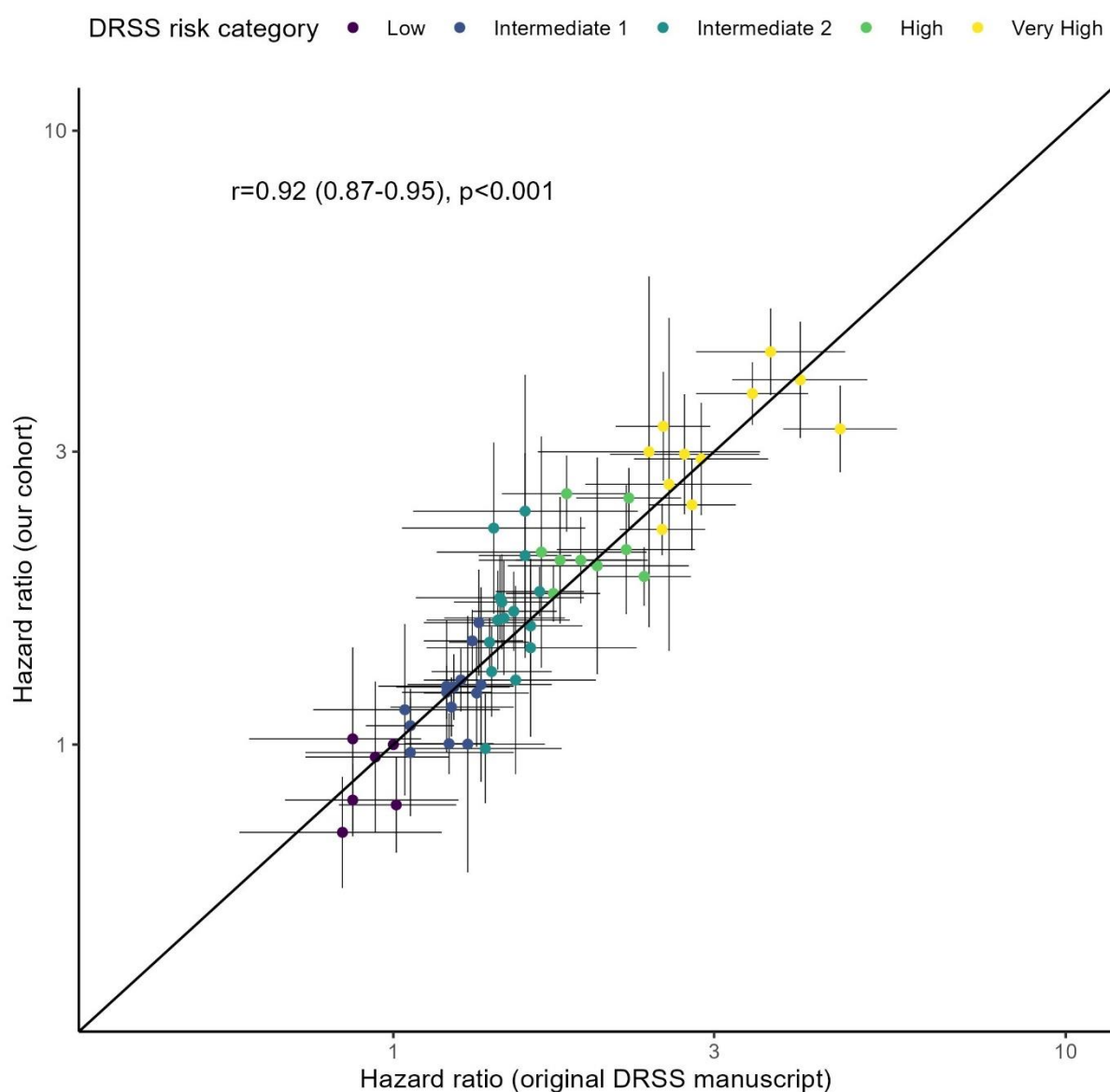


Figure 1: hazard ratios per disease category according to the original DRSS and in our cohort, horizontal lines are confidence intervals in DRSS, vertical lines confidence intervals in our cohort, diagonal line is the line of perfect agreement, calculated correlation is on log HR scale. The reference category is AML, with intermediate risk cytogenetics, and without FLT3-ITD or an NPM1 mutation.