Supplementary Methods

Patient characteristics and incidence

Age was assessed at the index and determined from the date of birth records. DLBCL aetiology was determined based on the potentially available diagnosis codes for FL or CLL (i.e.. C82* or C91.1*, respectively) prior to the first DLBCL diagnosis code. Absence of FL/CLL diagnosis codes was considered as a *de novo* DLBCL or a patient with missing record.

The other variables were assessed with records as close to the index as possible. However, results more than 90 days apart from the index were excluded.

IPI risk score and stage were text mined from patient texts. For stage, both the Ann Arbor [1] and TNM (Tumor-Node-Metastasis) [2] stagings were accepted. The IPI score was manually calculated [1] for the remaining patients for whom age, stage, LDH, and ECOG/Zubrod information were available. ECOG/Zubrod values were text-mined. The number of extranodal sites was not structurally available for any of the remaining patients, and therefore, the maximum manually calculated IPI score was 4. The correlation between the manually calculated and text mined IPI scores of the patients for whom both methods were applicable (238/587) was 0.75 (Pearson correlation, p<0001).

Structured data was utilised in the assessment of the COO subtypes and number of performed FISH analyses of *MYC*/*BCL2*/*BCL6* gene rearrangements. Few observations related to the FISH analyses of either multiple myeloma, chronic lymphocytic leukaemia, or minimal residual disease were separately excluded.

Patients who received SCT were detected based on the procedure codes of the stem cell transplantation (110AL1, 110AL2, or 110AU3).

Median age of the cohort and assessed subcohorts were reported, while other determined characteristics were reported as proportions (%) to the assessed patients for whom the respective data was available.

Fisher's exact test was used to compare differences in the characteristics of the subcohorts.

Annual DLBCL incidence per 100000 at the HDSF was determined and tested for time dependent trend by testing statistical significance of the Kendall’s tau correlation coefficient between the year of index and the incidence using the Mann-Kendall trend test.

Additional time-to-event analyses

In the assessment of OS stratified by ICT LOT, the patients were followed since the beginning of the last detected LOT until death (event) or end of follow-up (March 31, 2019; censoring event). In TTNT analyses stratified by ICT LOT, the patients were followed from the beginning of each LOT until the beginning of the next LOT (event), death (event), or end of follow-up (censoring event).

Duration of treatment lines

Kaplan-Meier estimates and corresponding numerical estimates were used in the assessment of the duration of ICT treatment lines. The patients were followed until end of each ICT LOT, death (a censoring event), or the end of follow-up (a censoring event). The duration of treatment line was censored if the treatment line ended in a 60-day window prior the end of follow up, death or beginning of the next treatment line. Otherwise, the end of LOT was observed as an event. The rationale is that in the case of death or end of follow up the treatment would have likely continued if the patient had not deceased or the end of follow up had not been reached. Similarly, in the case of refractory patients, the treatment would have otherwise likely continued. Censoring treatment line lengths like this overall slightly increases the Kaplan-Meier estimates.

Determination of immunochemotherapy treatments

Corticosteroids (prednisolone, prednisone, and dexamethasone) and rituximab were not considered in the initial screening of the distinct ICT treatment lines, except if rituximab was the only administered drug (in addition to potentially used corticosteroids). This was due to the fact that corticosteroids may be used to treat many other conditions, and rituximab could not be used to distinguish LOTs since it may be included in multiple treatment regimens. Similarly, carmustine and melphalan were not used to define treatment lines as these correspond to single administrations related to SCTs.

In the initial screening of the distinct treatment lines, single administrations of the same drug 60 days apart at most were combined into continuous administrations. Administrations of different drugs were combined into one LOT if they had a complete overlap or a partial overlap of at least 60 days. If the overlap was less than 60 days, administrations of different drugs were separated into different LOTs with previous LOT considered to begin on the day of the first administration of the earlier drug and to end the day before the administration of the latter drug. If one drug changed or the administration was terminated without a sufficient overlap with the other drugs even while the administrations of other drugs were continued, the administrations of changed/terminated drug were separated into different LOTs. In any other case, the end of the treatment line was considered as the last recorded dose included in the treatment.

Drugs related to DLBCL were screened 90 days at most prior to index to also include patients whose index was perhaps not correctly captured.

Administrations of corticosteroids and rituximab were searched for 30 days at most prior to the beginning of each treatment line and potential data was included in the corresponding treatment line. Carmustine and melphalan (BEAM regimen) administrations, that were detected between LOTs, were considered to be a part of a previous LOT. The BEAM regimen is typically followed by a SCT. Furthermore, all drug administrations between potential SCT event (as defined above) and preceding records on stem cell mobilisation and/or BEAM were combined into a previous LOT similarly to the BEAM. Single cytarabine and methotrexate administrations were also treated similarly to the BEAM regimen since they are typically used as a prophylactic treatment against CNS relapses after the main curative treatment regimen.

Treatment regimens utilised in each LOT were determined by searching for administrations of the individual drugs belonging to the prespecified treatment regimens. Although rituximab and corticosteroids were part of most prespecified treatment regimens, they were not included in the initial LOT detection process. Rather, the proportion of rituximab and corticosteroid-based treatment regimens were reported separately. (R-)CHOEP/(R-)DA-EPOCH, (R-)COP, and (R-)CEOP regimens were considered mutually exclusive, while all other regimens were not considered mutually exclusive within a LOT.

The distribution of patients across the treatment lines and treatment regimens was visualised in a flow chart. End of follow-up (EOF) and death were included as separate outcomes with death considered as death in the chart if it occurred during the first two years from the end of the last detected LOT, otherwise the event was considered as EOF.

Supplementary references

1. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A Predictive Model for Aggressive Non-Hodgkin's Lymphoma. *N Engl J Med*. 329(14), 987–994 (1993).

2. Amin MB, Edge S, Greene F, *et al.*, editors. AJCC Cancer Staging Manual. 8th ed. Springer International Publishing.