

The Danish Lymphoid Cancer Research (DALY-CARE) data resource: the basis for developing data-driven hematology

Christian Brieghel^{1*}, Mikkel Werling^{1*}, Casper Møller Frederiksen¹, Mehdi Parviz^{1,2}, Caspar da Cunha-Bang¹, Tereza Faitova¹, Rebecca Svanberg Teglgard^{1,3}, Noomi Vainer¹, Thomas Lacoppidan¹, Emelie Rotbain^{1,4}, Rudi Agius¹, and Carsten Utoft Niemann^{1,2}.

1. Department of Hematology, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark

2. Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

3. Department of Clinical Immunology, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark.

4. Danish Cancer Institute, Copenhagen, Denmark

*Co-first authors

Running head: The DALY-CARE data resource

Keywords: DALYCARE; data-driven medicine; Chronic Lymphocytic Leukemia (CLL), Multiple Myeloma (MM); lymphoma, machine learning; hematology.

Corresponding author:

Carsten Utoft Niemann

Rigshospitalet

Department of Hematology

Blegdamsvej 9, building 7054

2100 Copenhagen Ø

Denmark

Phone: +45-35458878

e-mail: carsten.utoft.niemann@regionh.dk

Abstract word count: 148/150

Text word count: 4401/4000-4500 words

Insert count: 10/10 displays (6 figures + 4 tables)

Reference count: 52/60

33 Abstract

34 Lymphoid-lineage cancers (LC: lymphoma, chronic lymphocytic leukemia, multiple myeloma, and their precursors) share many
35 epidemiological and clinical features. To develop data-driven hematology, we gathered electronic health data and created open-
36 source data processing pipelines to create a comprehensive data resource for Danish LC Research (DALY-CARE) approved for
37 epidemiological, molecular, and data-driven research. We included all Danish adults registered with LC diagnoses since 2002
38 (n=65,774) and combined 10 nationwide registers, electronic health records (EHR), and laboratory data on a high-powered cloud-
39 computer to develop a secure research environment. We herein exemplify how DALY-CARE has been used to develop novel
40 prognostic markers using biobank data, real-world evidence to evaluate the efficacy of care, and medical artificial intelligence
41 algorithms deployed directly into EHR systems. The DALY-CARE data resource allows for development of both near real-time
42 decision-support tools and extrapolation of clinical trial results to clinical practice, thereby improving care for patients with LC.

43 Background & Summary

44 Lymphoma, chronic lymphocytic leukemia (CLL), plasma cell dyscrasia (PCD) and their precursors are a highly heterogenous group
45 of hematological malignancies in the blood, bone marrow and/or lymphoid tissues¹. They span from precursor states such as
46 monoclonal B-cell lymphocytosis (MBL) and monoclonal gammopathy of uncertain significance (MGUS), over indolent B-cell
47 lymphomas (BCL), smoldering multiple myeloma (MM) and CLL, to aggressive high-grade BCL and PCD such as amyloid light-
48 chain amyloidosis. Lymphoid-lineage cancers (LC) accounts for approximately 5% of all newly diagnosed cancers and more than 4%
49 of all cancer deaths worldwide, whereas precursors such as MGUS and MBL may be identified in 5% to 12% of healthy, older
50 individuals^{2,3}.

51 Despite the pathological, biochemical, molecular, and clinical heterogeneity of LC, they share several common features. First, all LC
52 derive from or are directly caused by monoclonal lymphoid-lineage expansion in blood, bone marrow and/or lymphoid tissues.
53 Second, they are diagnosed by either histological tissue staining, flow cytometry, and/or molecular analyses, and the diagnostic
54 workup typically includes biochemical blood tests, a bone marrow examination, and computer tomography (CT) or positron emission
55 tomography (PET)/CT imaging. Third, different LC share common epidemiology such as older age at diagnosis and male
56 predominance, whereas socioeconomic status and lifestyle exposures are generally not correlated with risk of developing LC. Fourth,
57 LC develops due to dysregulation of immune function, stimulation and signaling in conjunction with accumulation of driver
58 mutations⁴. Lastly, most patients with LC precursor states, CLL, indolent lymphoma or smoldering MM do not benefit from pre-
59 emptive treatment and are therefore followed by a watch-and-wait strategy. If or when treatment is required, it usually includes
60 corticosteroids, chemotherapy or immunotherapy in combination, while targeted therapies are rapidly substituting or adding to
61 chemoimmunotherapy combinations for most LC⁵. From an epidemiological, molecular, and data-driven perspective, the
62 commonalities between LC offer a unique opportunity for identifying common features across the different LC. This opens new
63 possibilities regarding modelling disease outcomes using meta-learning and federated learning along with optimized transfer of
64 molecular and genetic findings between the different LC. Achieving state-of-the-art accuracy of predictive models will likely require
65 the joint modelling of diseases i.e., including training data from patients with lymphoma and multiple myeloma may improve
66 predictive performance for patients with CLL.

67 To facilitate a shared data infrastructure to study LC, we developed the Danish LC Research (DALY-CARE) data resource allowing
68 for clinical epidemiology, day-to-day monitoring of clinical outcomes, omics and functional analyses of available biobank samples⁶,
69 and data-driven medical artificial intelligence (mAI) research⁷. Specifically, DALY-CARE has been approved by the Danish
70 National Ethics Committee to collect 1) biobank samples and 2) electronic health data (EHD) retrospectively and prospectively for
71 all Danish adult residents diagnosed with LC. DALY-CARE is approved for the study of the integrated clinicopathological profile of
72 LC based on medical history, clinical and paraclinical (i.e. biochemical, cellular, microbiological, pathological, radiology, molecular,
73 and genetic) routine data, while molecular omics (e.g. chip-array techniques, genomics, transcriptomics, proteomics, and
74 microbiomics) and functional analyses (e.g. immune assays, phospho-flow cytometry, in vitro drug screening and mouse xenograft

75 models) of adjoined biobank samples are covered by the protocol for correlations with the course of treatment and clinical outcome
76 based on clinical epidemiology and data-driven mAI approaches.

77 The main driver in the recent successes of AI for image analysis and natural language processing outside of medicine is the critical
78 mass of researchers working on the same benchmarks competitively and collaboratively. This ensures a continuous improvement of
79 state-of-the-art performance on the given datasets⁸. By contrast, access to most medical data is heavily restricted. For instance, the
80 performance of prognostic models in CLL has plateaued over the last 20 years likely due to lack of (1) multiple data modalities, (2)
81 time-series modelling, and (3) a diverse set of researchers with access to the same benchmarks⁹. In turn, most mAI is developed on
82 outcomes that do not have the highest clinical impact - but instead on the restricted data sets that are easily and publicly accessible to
83 data scientists (e.g. the MIMIC III and IV datasets)^{9,10}. With the DALY-CARE resource, we thus aim to address the aforementioned
84 issues of lack of multimodality, common access, and clinically valid outcomes. Within DALY-CARE, we will create benchmark
85 datasets for outcomes with high clinical impact (e.g. death, treatment response, cardiac events, adverse events and infections) and
86 provide access to the multimodal data necessary to model these. We expect this pipeline to accelerate data-driven research for
87 hematological malignancies and the subsequent development of state-of-the-art decision support tools. The DALY-CARE data
88 resource provides the setup for automatized near real-time capture and monitoring of different outcomes upon changes in clinical
89 care as well as upon deploying decision support tools, whether developed based on the DALY-CARE cohort or based on external
90 datasets.

91 In this paper, we provide an overview of the DALY-CARE data resource, its infrastructure, and data formats. We exemplify the
92 potential of the data resource by providing examples of published studies and algorithms directly based on the DALY-CARE data
93 resource.

94 **Methods**

95 **Study population**

96 We included all Danish adult residents registered with a LC diagnosis since 1 January 2002 using International Classification of
97 Diseases version 10 (ICD10; i.e. C81.x-C90.x, C91.1-C91.9, C95.1, C95.7, C95.9, D47.2, D47.9B, and E85.8A) and Systematized
98 Nomenclature of Medicine (SNOMED) codes, which could be mapped to ICD10 codes for LCs (Supplemental Table S1)¹¹. Patients
99 were identified from three data sources: 1) the Danish Clinical Quality Program – National Clinical Registries (RKKP), 2) the Danish
100 Health Data Authority (SDS), and 3) the EPIC®-based EHR system in eastern Denmark (*Sundhedsplatformen* [SP]). Nearly all
101 Danish patients with LC are referred to one of eight Danish hematology departments, which are placed in five Danish regions
102 (Supplementary Information: Hematological centers and regional assignment to patients), and the nationwide coverage for malignant
103 LC diagnoses within RKKP has previously been estimated to 99%¹²⁻¹⁴. The cutoff date for follow-up and inclusion of data and newly
104 diagnosed patients with LC was 15 Nov 2023.

105 Data sources

106 EHD were primarily gathered from existing data sources as summarized in Table 1. From the RKKP registers, these include the
107 Danish National Lymphoma Registry (LYFO)¹⁵ since 2005, the Danish National Multiple Myeloma Database (DaMyDa)¹⁴ since
108 2005, and the Danish National Chronic Lymphocytic Leukemia Registry (DCLLR)¹² since 2008. Exhaustive lists of variables in the
109 RKKP registers are described elsewhere (Supplemental Table S2)¹⁶. From SDS starting from 2002, existing data sources included the
110 Register of Pharmaceutical Sales covering prescription drug data (LSR)¹⁷, the National Hospital Medication Register covering in-
111 hospital medication (SMR, coverage since 2004)¹⁸, the Danish National Pathology Register (PATOBANK) including pathology
112 notes (free text) and SNOMED codes¹⁹, the Clinical Laboratory Information System Database with routine laboratory results
113 (LABKA)²⁰, the Danish Register of Causes of Death (DAR)²¹, the Danish National Patient Registry (LPR) versions 1 and 3 with
114 diagnosis and procedure codes^{22,23}, and the Danish Cancer Registry (DCR)²⁴ (Supplemental Table S2).

115 We retrieved EHD from 14 modules in the EPIC® (or SP) EHR system of eastern Denmark including 1) administered medicine, 2)
116 admissions (ADT), 3) active in-hospital diagnoses, 4) all test results, 5) hematology/oncology treatment plans, 6) microbiology
117 charts, 7) transfers between departments, 8) intensive care unit admissions, 9) medical notes (as free text), 10) prescribed medicine,
118 11) out- and in-patient visits and diagnoses, 12) anthropometrics, 13) a social history including smoking and alcohol consumption,
119 and 13) vital signs (Supplemental Table S3; Supplementary Information: Go-live dates).

120 Additionally, we gathered EHD indirectly from laboratory systems at our institution. These included cleaned versions of the Danish
121 Microbiology Database (MiBa)²⁵ and LABKA retrieved through the Personalised Medicine of Infectious Complication in Immune
122 Deficiency (PERSIMUNE) covering the Capital Region of Denmark (Table 1)²⁶. In addition, we added summary reports and results
123 directly from the laboratory systems for routine flow cytometry analyses²⁷, immunoglobulin heavy-chain variables gene (IGHV)
124 analyses including stereotypic subset designation²⁸, targeted next-generation sequencing (tNGS)²⁹, and fluorescence in situ
125 hybridization (FISH) reports. Manually curated datasets collected through EHR reviews include information on patients receiving
126 second line CLL treatment, patients with CLL treated with ibrutinib, patients with MM treated with daratumumab, and patients with
127 CLL developing Richter's transformation from previously published cohorts³⁰⁻³³.

128 Finally, the DALY-CARE protocol specifically allows for functional and molecular analyses including omics of biobank samples
129 from individuals in the cohort. For this purpose, we gathered available biobank information from four large Danish biobanks, namely
130 the CLL biobank, the Copenhagen Hospital Biobank, the Danish Cancer Biobank, and PERSIMUNE biobank (Table 1)³⁴. Imputed
131 genotypes from peripheral blood at time of first hospital blood-workup are available in 9,320 individuals (see Supplementary
132 Information: Future data perspectives)³⁵.

133

134 Ethical approvals

135 The DALY-CARE protocol has been approved by the Danish Health Data Authority and National Ethics Committee (approvals P-
136 2020-561 and 1804410, respectively). According to Danish legislation, the collection of electronic health register data is mandatory,
137 and data were collected for research purposes in accordance with the approved protocol (see Supplemental Appendix). The Danish
138 National Ethics Committee granted an exemption for patients to provide informed consent in order to share electronic health data.
139 The exemption was based on the potential high impact for the patient group in question, thus considered to outweigh the issues raised
140 by an exemption. This exemption was also extended to allow analyses of biobank samples including extensive molecular analyses for
141 the retrospective part of the cohort, while for such prospective sampling, written informed consent was provided by patients to collect
142 and analyze biobank samples and electronic health data.

143

144 Data Records

145 Main variables

146 Variables in existing data sources have been described elsewhere (Supplemental Table S2)^{12,14,15,17-21,23,25}. In short, baseline
147 demographic, prognostic, common biochemical, and molecular data, as well as clinical data on treatment, response, and survival are
148 assembled from the RKKP registers, which are kept as wide-format datasets and cleaned through protocols available in the DALY-
149 CARE data resource (Supplementary Information: Software). The main variables use encoding according to the Danish Medical
150 Classification System^a (SKS)³⁶. In brief, SKS encoding covers diagnoses, pathology, biochemistry, locations as well as surgical,
151 radiological and treatment procedures using ICD10, SNOMED for pathology, nomenclature property units (NPU) for laboratory,
152 health care provider location (SHAK), and anatomical therapeutic chemical (ATC) codes for medicine (Supplementary Information:
153 Codes and formats). These codes facilitate clear definitions and easy mapping of data when linking datasets in DALY-CARE
154 (Supplemental Table S4-S5). All diagnoses, medicine, and biochemistry across all available datasets have been gathered into
155 independent aggregated views.

156 We underscore that data do not include any sensitive information such as names, addresses, religion, ethnicity, political views,
157 financial data, or sexual orientation.

158

159 Database infrastructure

160 The DALY-CARE data resource is based on open source scalable PostgreSQL located inside a secure ISO 27001 certified private
161 cloud on a high-performance supercomputer at the Danish National Genome Center (NGC) research infrastructure. The server
162 is accessed via a Federal Information Processing Standard (FIPS) compliant virtual desktop interface (VDI) that provides full

^aSundhedsvæsenets Klassifikations System

163 separation between users. Built specifically with data security and privacy in mind, the platform uses 2-factor authentication (2FA)
164 and layered security approach. The platform is on-premises and does not use any components that are hosted outside the
165 infrastructure, which ensures data security and compliance with the statutory acts³⁷.

166 All data are pseudonymized based on a single patient ID linked to the Danish Civil Registration System's CPR number³⁸. This allows
167 for unique identification of all patients and enables easy data linkage across all datasets while protecting individual patients' data
168 privacy (Figure 1a). All raw data are available directly in the DALY-CARE database. Newly retrieved data is quality assessed and
169 previous data cuts are logged to ensure the reproducibility of all studies.

170 The DALY-CARE server is divided into two databases with three schemas each:

171 1. *import*

- 172 a. *public* contains raw data retrieved from RKKP, SDS, EHR, and PERSIMUNE.
- 173 b. *laboratory* contains data from hematological laboratories including flow cytometry, FISH, IGHV, and NGS data.
- 174 c. *lookup_tables* contain tables to map encoded data including SNOMED, SKS, ATC, ICD-10, NPU, and SHAK.

175 2. *core*

- 176 a. *public* contains cleaned versions of raw datasets and aggregated data from multiple data sources.
- 177 b. *curated* contains manually curated datasets obtained from manual EHR reviews and NGS data.
- 178 c. *lookup_tables* contain cleaned tables to map standard coding.

179 We created data processing pipelines and functions for loading and processing the data in both R and Python software (Figure 1b;
180 Supplementary Information: Software). This includes pipelines that transform the raw data into a format ready to use by predictive
181 models (feature generation), and scripts that define and extract clinical outcomes from the raw datasets. We grouped functions based
182 on their specific purpose inside or general use outside the DALY-CARE server to allow for synergy in larger collaboratives of
183 Danish epidemiologists and data scientists studying similar register data as well as for adaptation to similar international data.
184 Importantly, the pipelines in R and Python software allow for direct database queries to quickly load data using indexed encoded
185 variables (Supplemental Table S4)^{39,40}.

186

187 Biobank material

188 On top of the electronic register and EHR data, we also included laboratory data as indicated in Table 1. To facilitate large-scale
189 omics studies in DALY-CARE, we searched four large Danish biobanks and identified 40,863 biobank samples either stored on
190 different dates or collected from different tissues from 18,528 individuals. This included 21,431 samples among 9,636 patients with
191 lymphoma (ICD10 codes C81.x-C89.x), 11,043 samples in 3,809 patient with CLL (ICD10 code C91.1), and 6,206 samples in 3,216
192 patients with PCD (ICD10 codes C90.0-C90.3, D47.2, and E85.8A). The different tissues included bone marrow (n=2,535), lymph

193 node (n=2,624), peripheral blood (n=24,068), DNA from peripheral mononuclear cells (n=2300), viably frozen cells (n=826), and
194 plasma (n=8,510). Most peripheral blood samples were drawn upon first hospital visit and may in most instances serve as germline
195 (normal) samples. To create a large genetic repository in the future, genotyping (n>12,000), whole genome sequencing (n>2,000),
196 and proteomics (n>800) is ongoing.

197

198 Technical validation

199 Patients

200 We identified all adult patients with ICD10 codes or SNOMED codes that could be mapped to LC diagnoses (Supplemental Table S1
201 and S6). In total, we included 65,744 patients with a LC of whom 35,399 (53.8%) had more than one diagnosis: 15,838 (24.1%),
202 10,420 (15.8%), 5315 (8.1%), and 3826 (5.8%) patients had 2, 3, 4, and >4 different LC diagnoses, respectively. Most patients with
203 multiple LC diagnoses were initially diagnosed with unspecified LC (e.g. C81.9, C82.9, C85.5, C85.9, and C91.9) followed by
204 specified subclassification, whereas fewer patients had multiple diseases, progressed from precursor states to overt LC or transformed
205 to more aggressive disease (Figure 2; Supplementary Information: Detailed patient information). The 20 most common LC diagnoses
206 could be attributed to 62,946 patients (95.7%) and are summarized in Table 2.

207 At time of first LC diagnosis, the median age was 70.3 years (interquartile range [IQR] 60.7;77.9) and 56.1% were male (Table 3).
208 From the cause of death register, RKKP and EHR data, we compiled dates of last follow-up (31 Dec 2020, 15 Nov 2023, and 3 Sep
209 2023, for each source respectively) or death to calculate time from first LC diagnosis to death or end of follow-up. In 1763 (2.7%)
210 patients the last date of follow-up antedated the first diagnosis due to registration lag time. As a result, Kaplan-Meier days and
211 survival status are readily available in the DALY-CARE data resource for survival analyses. The median follow-up time was 8.5
212 years (IQR, 4.9;13.4). For the 10 most common LCs, the 5-year unadjusted OS ranged from 50.0% in MM to 74.8% in FL (Figure 3;
213 Supplemental Table S7). A detailed description of the patients and baseline characteristics is available in the Supplementary
214 Information (please see Detailed patient information). We next used clinical baseline characteristics including polypharmacy defined
215 by prescriptions in the year prior to first LC diagnosis (Supplemental Table S8) to perform multivariable Cox regression analyses in
216 LC subtypes with available information on disease-specific international prognostic indices (IPI; Supplemental Table S9). Adjusted
217 for age, sex, IPI, Charlson comorbidity index (CCI) score (2 vs >2), and polypharmacy (<5 vs ≥5 drugs), we demonstrated an
218 independent association with shorter OS for disease-specific IPI, CCI score, and polypharmacy (Supplemental Figure S3)^{41,42}.
219 However, polypharmacy intervals (0-3, 4-6, 7-9, 10-12, and >12 different ATC codes) further demonstrated a notable dose-response
220 effect on OS, which likely underscores an improved ability of prescriptions to explain the survival impact of comorbidity better than
221 ICD10 codes obtained from hospital admissions used to calculate CCI scores (Figure 4).

222 Data coverage

223 To avoid bias when linking different datasets, we provide an overview of coverage over time for data from RKKP, SDS, EHR, and
224 PERSIMUNE datasets (Figure 5 and Supplementary Figure S4). While all data are population-based, data collection was centered

225 around eastern Denmark and our institution in specific. We thus calculated the prevalence of CLL, DLBCL, and MM in the data
226 available in DALY-CARE to assess regional differences, which would likely represent differences in data coverage (Figure 6). In
227 addition to the spatial biases in the data, Figure 5 shows a clear temporal bias. Generally, more recent data has better coverage than
228 older data: 13 out of 54 data sources plotted date back to 2002, while all 54 data sources have data available for 2019. Furthermore,
229 difference in regional coverage over time reveal better coverage in the Capital Region for LABKA data sources (Supplemental
230 Figures S5-S9 panels F1 and H1).

231 Previous data usage

232 Due to the variety of data sources and data modalities, the DALY-CARE data resource may facilitate a diverse set of research
233 projects. Here we highlight some of the unique possibilities of DALY-CARE by giving examples of previously published work
234 including use of biobank data to find novel prognostic markers, investigation of real-world evidence of different prognostic factors
235 influencing the efficacy of health care, and development and deployment of mAI algorithms. To avoid extreme self-citing,
236 publications have been summarized in Table 4.⁴³

237 Using biobank data, we have validated and identified molecular risk factors associated with shorter time to first treatment and disease
238 progression. In collaboration with the [European Research Initiative on CLL \(ERIC\)](#), we have contributed to demonstrating adverse
239 outcome for certain stereotypic B-cell receptor (BcR) subsets and identifying BcR subsets in approximately 40% of patients with
240 CLL. In other ERIC collaborations, recurrent gene mutations influence time to first treatment differently in patients with mutated and
241 unmutated IGHV status. Additionally, we have shown that recurrent gene mutations in CLL may be assigned to key signaling
242 pathways to improve prognostication of time to first treatment. We have further demonstrated that low variant allele frequency *TP53*
243 gene mutations do not impact overall survival time from CLL diagnosis per se, but instead have high clinical impact from time of
244 CLL treatment, whereas patients with multi-hit *TP53* aberrations being treated with ibrutinib demonstrate shorter progression-free
245 survival. By accessing large Danish biobanks, we have further been able to detect CLL clones in samples decades before diagnosis of
246 CLL. In two other studies, we investigated the microbiome of CLL patients, and used fresh blood samples to demonstrate immune
247 response improvement in clinical trial patients treated with targeted therapy. In brief, these studies highlight the potential of using
248 multimodal approaches to achieve personalized medicine for patients based on for instance gene mutations, immune phenotype, and
249 microbiome analyses.

250 With large cohorts of patients with LC included in DALY-CARE, the data resource is highly suited for real-world evidence
251 epidemiological studies. By describing the Danish CLL register and validating the international prognostic index for CLL (CLL-IPI)
252 in the Danish CLL population and for Binet A stage patients, we have identified variables that may supplement and improve the
253 predictive performance of CLL-IPI. These variables include driver mutations in cell signaling pathways, eosinophil counts,
254 hypogammaglobulinemia, comorbidity scores (CLL-CI), type 2 diabetes, and an index to identify patients without need of treatment
255 (CLL-WONT). For CLL-WONT, we subsequently showed that it was both feasible and safe to end specialized follow-up for patients

256 with lower risk CLL. DALY-CARE was also used to gather data from a wide range of treatment cohorts to describe real-world
257 outcomes upon first and second line chemoimmunotherapy and ibrutinib in CLL as well as upon daratumumab in MM. In other
258 studies, we defined and assessed the risk of clinically relevant outcomes such as second primary malignancies in CLL associated with
259 exposure to chemoimmunotherapy, Richter's transformation associated with CLL high risk features, atrial fibrillation upon ibrutinib,
260 and blood stream infections in CLL and MM. Such standardized epidemiological definitions of adverse outcomes and events are
261 available in DALY-CARE. Prior to LC diagnosis, we have also been able to demonstrate a decade long increased prescription of
262 antimicrobials, and lymphocyte slopes up until CLL diagnosis further informing disease trajectories. Finally, during the corona virus
263 disease pandemic, we could quickly assess and monitor mortality in a wide range of LCs, and we further demonstrated poor vaccine
264 responses in CLL and better clinical outcomes during the omicron era. This underscores how frequently updated data allows for near
265 real-time monitoring of large-scale real-world clinical outcomes.

266 We also employed DALY-CARE to develop data-driven studies using mAI. For instance, we showed that adding paraclinical
267 information to CLL-IPI variables improved predictive performance, whereas recurrent gene mutation information did not. This
268 highlights the potential within DALY-CARE to estimate which classes of variables and modelling approaches are most informative
269 for a given outcome. Notably, we have developed and deployed the data driven model CLL-TIM for predicting whether patients with
270 newly diagnosed CLL will have an infection or will be treated within 2 years. This model was implemented into the ongoing
271 PreVent-ACaLL clinical trial ([NCT03868722](https://clinicaltrials.gov/ct2/show/study/NCT03868722)) and the EHR system of eastern Denmark, and we have shared recommendations and
272 considerations for easing transition from development to deployment for future mAI algorithms⁴⁴. In conclusion, these studies
273 highlight that DALY-CARE facilitates transformation of detailed EHD and biobank-based analyses into actionable models and
274 prognostic indices with direct impact on daily clinical practice.

275

276 Discussion

277 The DALY-CARE data resource is a collection of national health registries, EHRs, routine and special laboratory data as well as
278 analyses from adjoined biobanks. Including more than 65,000 patients with LC, this data resource was gathered and stored on a
279 super-computer cloud ensuring both the data privacy and safety, storing capacity, and the processing power needed to handle
280 sensitive data without compromising the risk of data migration⁴⁵. Based on published studies using the DALY-CARE data resource,
281 we provide examples of how DALY-CARE can lead to (1) finding novel prognostic markers using biobank data, (2) using real-world
282 evidence studies to evaluate the efficacy of routine health care, and (3) deploying mAI algorithms directly into EHR systems.

283 The DALY-CARE cohort is truly population-based as nearly all Danish adult LC patients are referred to one of eight hematological
284 centers and cancer registration is mandatory by law. This results in a register coverage of 99%¹³. Even so, we found large regional
285 disparities in the available data for patients mainly owing to EHR data being available for only two out of five Danish Regions,
286 covering half of the Danish population. The use of ICD10 codes upon admissions may also differ widely across institutions. This

287 skew in availability and usage could limit the scope of the possible studies or produce unintentional biases in data driven algorithms
288 if they are not accounted for. Like other EHD archives, data in DALY-CARE are affected by shifts in systems such as the transition
289 from LPR to LPR3 (i.e. 2 Feb 2019) or implementation of EPIC® in eastern Denmark (from May 2016 through Nov 2017;
290 Supplementary Information: Go-live dates). These time dependent changes in data and data formats are important to recognize to
291 avoid biases that could lead data driven algorithms astray. To this end, we focused on 1) describing when and where the data are
292 available and identified sub-cohorts according to ICD10 diagnoses, 2) standardization and quality control for measurements (i.e.
293 making the same measures using different units commensurable), and 3) using robust clinical definitions to calculate features based
294 on domain-knowledge such as prognostic indices.

295 By highlighting previously published work (Table 4), we hope that DALY-CARE will serve as an important first step towards
296 standardizing and creating similar data resources for other research groups. This would help alleviating the problem of limited
297 training data by enabling the use of techniques that can leverage information learned from training on other cohorts such as meta-
298 learning⁴⁶. Beyond the national generalization (e.g. SKS coding), we mapped 90% of our curated data from medicine, laboratory
299 measurements and diagnosis to OMOP standardization to facilitate international collaborations. This will also serve as a critically
300 important step for implementing state-of-the-art mAI that aggregate information across countries using federated learning.

301 Most research in mAI is almost exclusively focusing on developing novel methods and algorithms, whereas reports of deploying,
302 monitoring, improving, or maintaining existing algorithms are anecdotal^{47,48}. Outside of image analysis used for clinical care, mAI is
303 rarely deployed directly into clinical patient care. The DALY-CARE data resource provides real-world clinical data and is regularly
304 updated. Essential parts of the codebase are publicly available on [Github](#) and regularly updated. We believe that access to near real-
305 time real-world clinical data will ease the transition from development to production, aiding the issue of deploying algorithms.

306 Medical interventions for LC patients are often implemented without scientific evidence, and many interventions will never be
307 investigated in randomized clinical trials (RCT)⁴⁹. DALY-CARE provides means to qualify and inform the impact of changes in
308 health care practice that have not, cannot, and likely will never be tested sufficiently in RCTs. This is exemplified by monitoring
309 overall survival, health care utilizations and infections for patients ending specialized follow-up or receiving immunoglobulin
310 replacement therapy. Having access to large-scale multimodal data allows not only for monitoring of clinical impact after clinical
311 practice changes in management and supportive care but also when introducing novel treatments tested in RCT to RCT-ineligible
312 patients. For instance, half of all Danish patients with newly diagnosed MM did not qualify for inclusion in RCT⁵⁰, and Danish
313 patients on ibrutinib outside clinical trials only have a median duration to discontinuation of 3 years as compared to 7 years in clinical
314 trials^{31,50,51}.

315 Improvements in patient care is intrinsically tied to the ability to access and share data easily⁵². Here, we demonstrate how DALY-
316 CARE allows researchers to identify complex patterns across different patient cohorts and subgroups on a collaborative basis. The
317 identification of such patterns may provide new data-driven hypotheses and prognostic markers as well as inform and qualify the

318 impact of implemented management of LC patients, while providing the basis for development of improved decision support tools
319 for LC patient care. Providing the ability to share definitions, outcomes and ultimately data will allow for competitive modeling on
320 outcomes that would most likely increase predictive performance of prognostic models. To translate these data-driven perspectives
321 into benefits for patients on a national and international level, we believe it is crucial to break down the arbitrary division between
322 primary and secondary use of EHD. That is, utilizing daily routine EHR data and national registers for the purpose of clinical
323 epidemiology, precision medicine and to enrich omics analyses with clinical data.

324 In conclusion, we integrated real-world data from quality hematology registers, nationwide health registers, electronic health record
325 data, analyses of biobank samples, and data from specialized hematological laboratories to facilitate cutting edge research in clinical
326 epidemiology, large-scale observational studies, omics analyses, and development of decision support tools based on machine-
327 learning algorithms. Capturing and utilizing routine EHD to evaluate changes in medical practice that are not based on evidence from
328 randomized clinical trials can now be monitored in near real-time. DALY-CARE thus paves the way for truly data-driven
329 hematology and provide proof-of-concept for improved data-driven health care.

330 Usage Notes

331 Due to pseudonymized but not fully anonymized nature of data, access to the data resource will be based on a Data Usage Agreement
332 (named Data Processor Agreement [DPA], see Supplemental Appendix 1 for template). The DPA will specify the analyses
333 performed on a collaborative basis, these analyses should be within the approved DALY-CARE protocol (Supplemental Appendix
334 2). All analyses must be performed on the DALY-CARE data resource at NGC. Except for pathology notes and medical notes,
335 dummy tables are presented in Supplemental Appendix 3.

336 Code availability

337 The underlying code for this study is available on the DALY-CARE data resource and can be accessed on a collaborative basis on
338 reasonable request to the corresponding author. Essential parts of the codebase are publicly available on [Github](#) and regularly
339 updated.

340

341 Funding

342 The project was funded by the Alfred Benzon foundation, the Danish Cancer Society (grant R269-A15924), and the CLL-CLUE
343 project funded by the European Union. This work was based on data analyzed at the national infrastructure for personal medicine
344 hosted at the Danish National Genome Center, which is supported by the Novo Nordisk Foundation (grant agreement
345 NNF18SA0035348 and grant agreement NNF19SA0035486). This work was supported by Danish Data Science Academy, which is
346 funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516). The PERSIMUNE project
347 contributed data and achieved funding from the Danish National Research Foundation (#126). WGS is achieved through
348 collaboration with deCODE genetics (Reykjavík, Iceland).

349 Author contributions

350 C.U.N. conceived the project. C.U.N., C.B., and C.M.F. collected the data. C.M.F. created the database. C.B., C.M.F, M.W., and T.L.
351 created the database infrastructure. M.W. and T.L. performed data quality control. C.B. and M.W. wrote the draft manuscript. All
352 authors contributed to and approved the final manuscript.

353 Acknowledgements

354 We would like to thank dr. Peter de Nully Brown for founding the Danish National Lymphoma Registry and leading the way for
355 Danish quality registers in hematology. We sincerely thank Anders Christian Riis-Jensen and Anton Kokholm Andersen from CØK
356 within the Capital Region of Denmark for retrieving and providing EHR data. We further thank Professor Jens Lundgren leading
357 PERSIMUNE for providing laboratory and microbiology data, Professor Sisse Rye Ostrowski for collaboration with the Copenhagen
358 Hospital Biobank, dr. Ida Schjødt for kindly providing flow cytometry data from the Surface Marker Laboratory at Rigshospitalet, dr.
359 Mette Klarskov Andersen for kindly providing cytogenetics data from the Department of Clinical Genetics, Rigshospitalet, and Lone
360 Bredo Pedersen for performing and providing IGHV analyses from the CLL laboratory, Rigshospitalet.

361 Competing interests

362 CB received travel grants from Octapharma. CMF received funding from Octapharma. TL received travel grants from AbbVie
363 outside this study. NV received consultancy fees and funding from AstraZeneca outside of this work. ECR received consultancy fees
364 and/or travel grants from Abbvie, Janssen, and AstraZeneca outside of this work. CUN received research funding and/or consultancy
365 fees from AstraZeneca, Janssen, AbbVie, Beigene, Genmab, CSL Behring, Octapharma, Takeda, Eli Lilly, MSD, and Novo Nordisk
366 Foundation. All other authors declare no competing interests to disclose.

367 References

- 368 1 Alaggio, R. *et al.* The 5th edition of the World Health Organization Classification of
369 Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia* **36**, 1720-1748 (2022).
370 <https://doi.org:10.1038/s41375-022-01620-2>
- 371 2 Ferlay, J. *et al.* Cancer statistics for the year 2020: An overview. *Int J Cancer* (2021).
372 <https://doi.org:10.1002/ijc.33588>
- 373 3 Kyle, R. A. *et al.* Long-Term Follow-up of Monoclonal Gammopathy of Undetermined Significance. *N*
374 *Engl J Med* **378**, 241-249 (2018). <https://doi.org:10.1056/NEJMoa1709974>
- 375 4 Niemann, C. U. & Wiestner, A. B-cell receptor signaling as a driver of lymphoma development and
376 evolution. *Semin Cancer Biol* (2013). [https://doi.org:S1044-579X\(13\)00088-6](https://doi.org:S1044-579X(13)00088-6) [pii]
10.1016/j.semcancer.2013.09.001
- 377 5 Greer, J. P. *et al.* *Wintrobe's clinical hematology*. Fourteenth edition edn, (Wolters Kluwer Health
378 Pharma Solutions (Europe) Ltd., 2018).
- 379 6 Hasin, Y., Seldin, M. & Lusic, A. Multi-omics approaches to disease. *Genome Biol* **18**, 83 (2017).
380 <https://doi.org:10.1186/s13059-017-1215-1>
- 381 7 Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc J* **6**,
382 94-98 (2019). <https://doi.org:10.7861/futurehosp.6-2-94>
- 383 8 Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of*
384 *Computer Vision* **115**, 211–252 (2015). <https://doi.org:https://doi.org/10.1007/s11263-015-0816-y>
- 385 9 Agius, R., Parviz, M. & Niemann, C. U. Artificial intelligence models in chronic lymphocytic leukemia
386 - recommendations toward state-of-the-art. *Leuk Lymphoma* **63**, 265-278 (2022).
387 <https://doi.org:10.1080/10428194.2021.1973672>
- 388 10 Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).
389 <https://doi.org:10.1038/sdata.2016.35>
- 390 11 ZfKD. *Zentrum für Krebsregisterdaten: Conversion table/coding manual*,
391 <[https://www.krebsdaten.de/Krebs/EN/Content/Methods/Coding_manual/coding_manual_node.h](https://www.krebsdaten.de/Krebs/EN/Content/Methods/Coding_manual/coding_manual_node.htm)
392 [tm](https://www.krebsdaten.de/Krebs/EN/Content/Methods/Coding_manual/coding_manual_node.htm)> (
393 [https://www.krebsdaten.de/Krebs/EN/Content/Methods/Coding_manual/coding_manual_node.h](https://www.krebsdaten.de/Krebs/EN/Content/Methods/Coding_manual/coding_manual_node.htm)
394 [tm](https://www.krebsdaten.de/Krebs/EN/Content/Methods/Coding_manual/coding_manual_node.htm)>)
- 394 12 da Cunha-Bang, C. *et al.* The Danish National Chronic Lymphocytic Leukemia Registry. *Clin*
395 *Epidemiol* **8**, 561-565 (2016). <https://doi.org:10.2147/clep.s99486>
- 396 13 Arboe, B. *et al.* The Danish National Lymphoma Registry: Coverage and Data Quality. *PLoS One* **11**,
397 e0157999 (2016). <https://doi.org:10.1371/journal.pone.0157999>
- 398 14 Gimsing, P. *et al.* The Danish National Multiple Myeloma Registry. *Clin Epidemiol* **8**, 583-587 (2016).
399 <https://doi.org:10.2147/clep.s99463>
- 400 15 Arboe, B. *et al.* Danish National Lymphoma Registry. *Clin Epidemiol* **8**, 577-581 (2016).
401 <https://doi.org:10.2147/clep.S99470>
- 402 16 RKKP. *Dokumentation af de nationale kliniske kvalitetsdatabaser*, <[https://www.rkkp-](https://www.rkkp-dokumentation.dk/)
403 [dokumentation.dk/](https://www.rkkp-dokumentation.dk/)> (2023).
- 404 17 Johannesdottir, S. A. *et al.* Existing data sources for clinical epidemiology: The Danish National
405 Database of Reimbursed Prescriptions. *Clin Epidemiol* **4**, 303-313 (2012).
406 <https://doi.org:10.2147/clep.S37587>
- 407 18 SDS. *Sundhedsdatastyrelsen: The national hospital medication register*,
408 <[https://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-](https://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/sygdomme-laegemidler-og-behandling/sygehusmedicinregisteret)
409 [sundhedsregistre/sygdomme-laegemidler-og-behandling/sygehusmedicinregisteret](https://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/sygdomme-laegemidler-og-behandling/sygehusmedicinregisteret)> (2023).
- 410 19 Erichsen, R. *et al.* Existing data sources for clinical epidemiology: the Danish National Pathology
411 Registry and Data Bank. *Clin Epidemiol* **2**, 51-56 (2010). <https://doi.org:10.2147/clep.s9908>
- 412 20 Grann, A. F., Erichsen, R., Nielsen, A. G., Frøslev, T. & Thomsen, R. W. Existing data sources for
413 clinical epidemiology: The clinical laboratory information system (LABKA) research database at
414 Aarhus University, Denmark. *Clin Epidemiol* **3**, 133-138 (2011). <https://doi.org:10.2147/clep.S17901>

- 415 21 Helweg-Larsen, K. The Danish Register of Causes of Death. *Scand J Public Health* **39**, 26-29 (2011).
416 <https://doi.org:10.1177/1403494811399958>
- 417 22 Lyng, E., Sandegaard, J. L. & Rebolj, M. The Danish National Patient Register. *Scand J Public Health*
418 **39**, 30-33 (2011). <https://doi.org:10.1177/1403494811401482>
- 419 23 Schmidt, M. *et al.* The Danish National Patient Registry: a review of content, data quality, and
420 research potential. *Clin Epidemiol* **7**, 449-490 (2015). <https://doi.org:10.2147/clep.S91125>
- 421 24 Gjerstorff, M. L. The Danish Cancer Registry. *Scand J Public Health* **39**, 42-45 (2011).
422 <https://doi.org:10.1177/1403494810393562>
- 423 25 Voldstedlund, M., Haarh, M. & Molbak, K. The Danish Microbiology Database (MiBa) 2010 to 2013.
424 *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable*
425 *disease bulletin* **19** (2014). <https://doi.org:10.2807/1560-7917.es2014.19.1.20667>
- 426 26 PERSIMUNE. PERSIMUNE: CENTRE OF EXCELLENCE FOR PERSONALISED MEDICINE OF INFECTIOUS
427 COMPLICATIONS IN IMMUNE DEFICIENCY, <www.persimune.dk> (2023).
- 428 27 van Dongen, J. J. *et al.* EuroFlow antibody panels for standardized n-dimensional flow cytometric
429 immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia* **26**, 1908-1975
430 (2012). <https://doi.org:10.1038/leu.2012.120>
- 431 28 Agathangelidis, A. *et al.* Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia:
432 the 2022 update of the recommendations by ERIC, the European Research Initiative on CLL.
433 *Leukemia* **36**, 1961-1968 (2022). <https://doi.org:10.1038/s41375-022-01604-2>
- 434 29 Brieghel, C. *et al.* The Number of Signaling Pathways Altered by Driver Mutations in Chronic
435 Lymphocytic Leukemia Impacts Disease Outcome. *Clin Cancer Res* (2020).
436 <https://doi.org:10.1158/1078-0432.ccr-18-4158>
- 437 30 Vainer, N. *et al.* Real-world outcomes upon second-line treatment in patients with chronic
438 lymphocytic leukaemia. *Br J Haematol* **201**, 874-886 (2023). <https://doi.org:10.1111/bjh.18715>
- 439 31 Aarup, K. Real-World Outcome for 205 Danish Patients with Chronic Lymphocytic Leukemia Treated
440 with Ibrutinib. *Ash Annual Meeting Abstract* **#1767** (2019).
- 441 32 Ben-Dali, Y. *et al.* Risk Factors Associated with Richter's Transformation in Patients with Chronic
442 Lymphocytic Leukemia. *Ash Annual Meeting Abstract Paper abstract* **#1697** (2018).
- 443 33 Szabo, A. G. *et al.* The real-world outcomes of multiple myeloma patients treated with
444 daratumumab. *PLoS One* **16**, e0258487 (2021). <https://doi.org:10.1371/journal.pone.0258487>
- 445 34 Laugesen, K. *et al.* A Review of Major Danish Biobanks: Advantages and Possibilities of Health
446 Research in Denmark. *Clin Epidemiol* **15**, 213-239 (2023). <https://doi.org:10.2147/clep.S392416>
- 447 35 Jensson, B. O. *et al.* Actionable Genotypes and Their Association with Life Span in Iceland. *N Engl J*
448 *Med* **389**, 1741-1752 (2023). <https://doi.org:10.1056/NEJMoa2300792>
- 449 36 Danish-Health-Data-Authority. *Classifications*,
450 <https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/classifications> (2023).
- 451 37 The_Danish_National_Genome_Center. *Research projects on the Danish National Genome Center's*
452 *supercomputer*, <[https://eng.ngc.dk/research-and-international-collaboration/research-projects-](https://eng.ngc.dk/research-and-international-collaboration/research-projects-on-the-danish-national-genome-centers-supercomputer)
453 [on-the-danish-national-genome-centers-supercomputer](https://eng.ngc.dk/research-and-international-collaboration/research-projects-on-the-danish-national-genome-centers-supercomputer)> (2024).
- 454 38 Pedersen, C. B. The Danish Civil Registration System. *Scand J Public Health* **39**, 22-25 (2011).
455 <https://doi.org:10.1177/1403494810387965>
- 456 39 An international prognostic index for patients with chronic lymphocytic leukaemia (CLL-IPI): a meta-
457 analysis of individual patient data. *Lancet Oncol* **17**, 779-790 (2016). [https://doi.org:10.1016/s1470-](https://doi.org:10.1016/s1470-2045(16)30029-8)
458 [2045\(16\)30029-8](https://doi.org:10.1016/s1470-2045(16)30029-8)
- 459 40 Rossum, G. v. & Drake, J. F. Python reference manual. . *Centrum voor Wiskunde en Informatica*
460 *Amsterdam* (1995).
- 461 41 Masnoon, N., Shakib, S., Kalisch-Ellett, L. & Caughey, G. E. What is polypharmacy? A systematic
462 review of definitions. *BMC Geriatr* **17**, 230 (2017). <https://doi.org:10.1186/s12877-017-0621-2>

463 42 Quan, H. *et al.* Updating and validating the Charlson comorbidity index and score for risk
464 adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* **173**, 676-
465 682 (2011). <https://doi.org:10.1093/aje/kwq433>
466 43 Van Noorden, R. & Singh Chawla, D. Hundreds of extreme self-citing scientists revealed in new
467 database. *Nature* **572**, 578-579 (2019). <https://doi.org:10.1038/d41586-019-02479-7>
468 44 Agius, R. *et al.* Implementation of the CLL Treatment Infection Model Adjoined to an Electronic
469 Health Record System - Guidelines for Practical Implementation of Data-Driven Models. Preprints
470 with the Lancet. (2023). <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4555192>.
471 45 Zhang, J. *et al.* Mapping and evaluating national data flows: transparency, privacy, and guiding
472 infrastructural transformation. *Lancet Digit Health* **5**, e737-e748 (2023).
473 [https://doi.org:10.1016/s2589-7500\(23\)00157-7](https://doi.org:10.1016/s2589-7500(23)00157-7)
474 46 Zhang, X. S., Tang, F., Dodge, H. H., Zhou, J. & Wang, F. MetaPred: Meta-Learning for Clinical Risk
475 Prediction with Limited Patient Electronic Health Records. *Kdd* **2019**, 2487-2495 (2019).
476 <https://doi.org:10.1145/3292500.3330779>
477 47 Seneviratne, M. G., Shah, N. H. & Chu, L. Bridging the implementation gap of machine learning in
478 healthcare. *BMJ Innovations* **6**, 45-47 (2020). <https://doi.org:doi:10.1136/bmjinnov-2019-000359>
479 48 Tomašev, N. *et al.* Use of deep learning to develop continuous-risk models for adverse event
480 prediction from electronic health records. *Nature protocols* **16**, 2765-2787 (2021).
481 <https://doi.org:10.1038/s41596-021-00513-5>
482 49 Djulbegovic, B. A framework to bridge the gaps between evidence-based medicine, health
483 outcomes, and improvement and implementation science. *J Oncol Pract* **10**, 200-202 (2014).
484 <https://doi.org:10.1200/jop.2013.001364>
485 50 Klausen, T. W. *et al.* The majority of newly diagnosed myeloma patients do not fulfill the inclusion
486 criteria in clinical phase III trials. *Leukemia* **33**, 546-549 (2019). [https://doi.org:10.1038/s41375-018-](https://doi.org:10.1038/s41375-018-0272-0)
487 [0272-0](https://doi.org:10.1038/s41375-018-0272-0)
488 51 Aarup, K. *et al.* Real-world outcomes for 205 patients with chronic lymphocytic leukemia treated
489 with ibrutinib. *Eur J Haematol* **105**, 646-654 (2020). <https://doi.org:10.1111/ejh.13499>
490 52 Bauchner, H., McDermott, M. M. & Butte, A. J. Data Sharing Enters a New Era. *Ann Intern Med* **176**,
491 400-401 (2023). <https://doi.org:10.7326/m22-3479>

492

493 Tables

494 Please see separate excel files.

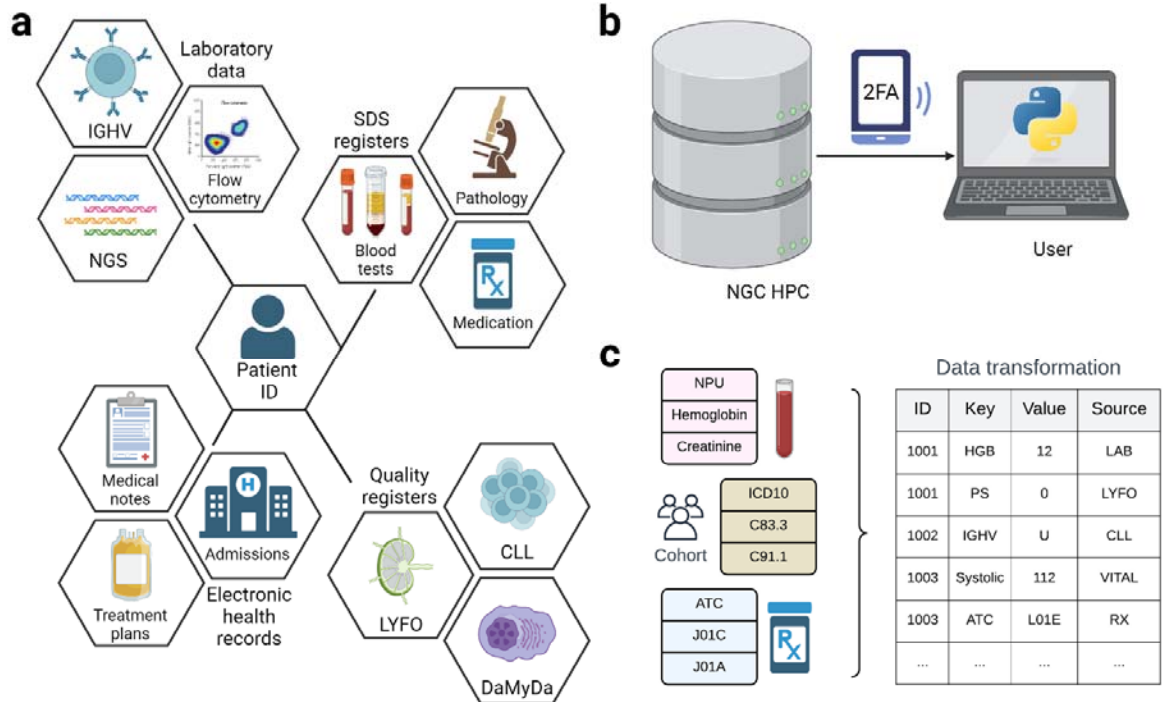
495 Table 1. Overview of datasets in the DALY-CARE data resource

496 Table 2. Most common lymphoid-lineage cancer disorders in the DALY-CARE data resource.

497 Table 3. Patient characteristics.

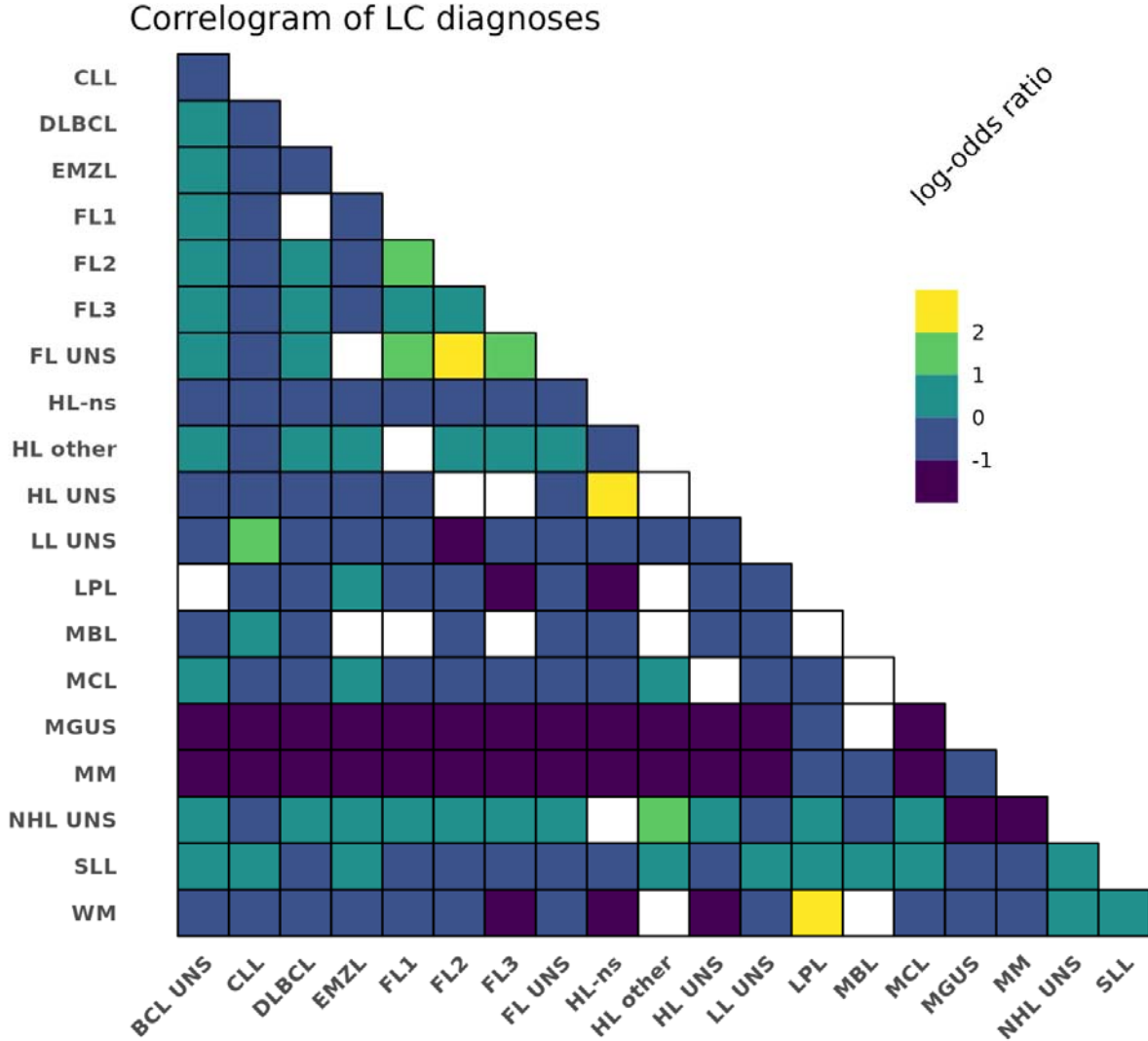
498 Table 4. DALY-CARE data resource usage.

499



501

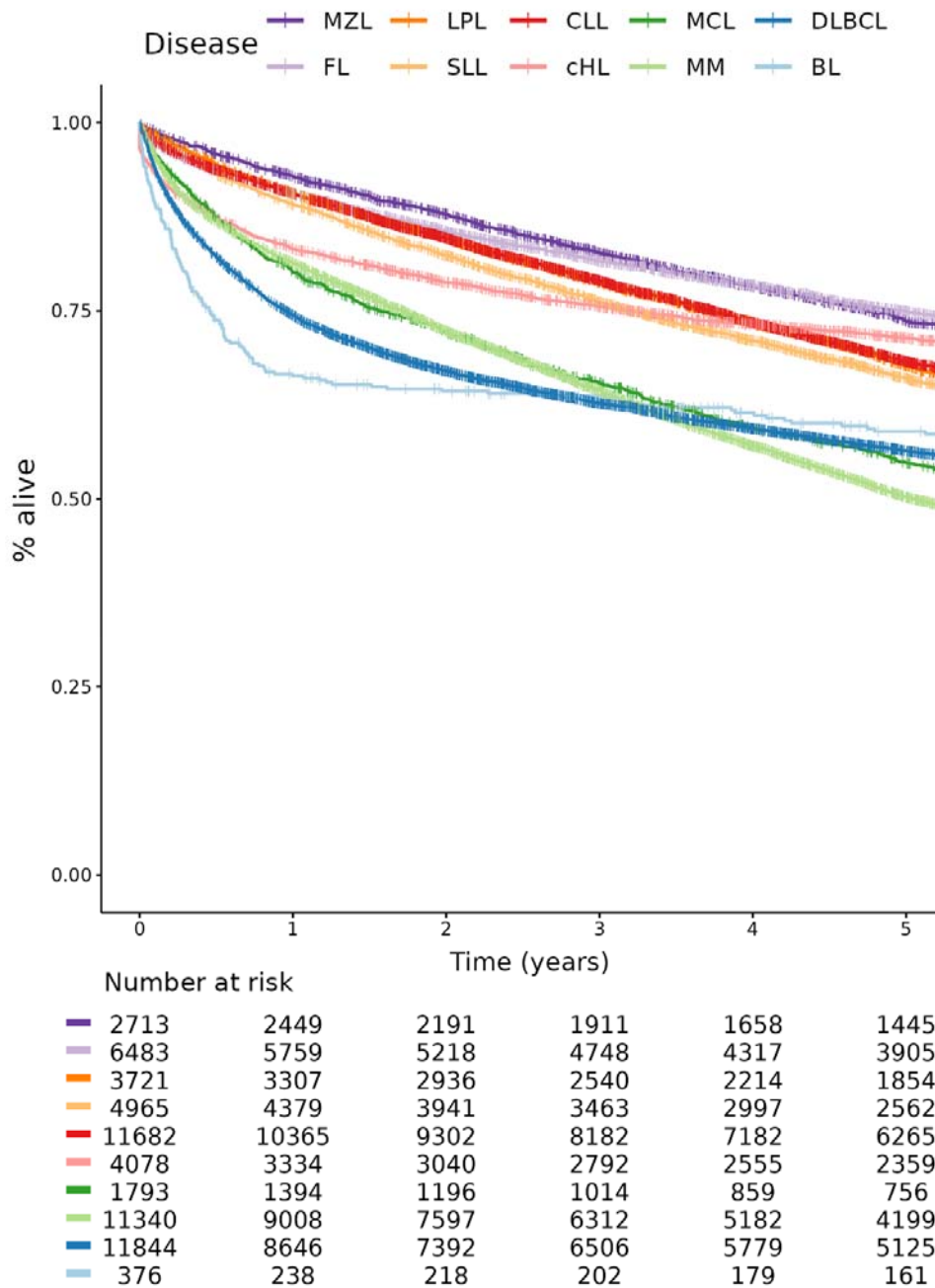
502 **Figure 1.** Schematic overview of the Danish Lymphoid Cancer Research (DALY-CARE) data resource. Data
 503 from various sources including Danish nationwide registers, hematological quality registers, electronic
 504 health record data, and hematological laboratories may be linked using a single pseudonymized patient ID
 505 (a). Hosted on the Danish National Genome Center high-performance computer (HPC) cloud, all data are
 506 placed in a PostgreSQL database accessed by 2-factor authentication (2FA) and loaded via R or Python
 507 software (b). Database queries and pipelines to load and transform data based on encoded data (e.g.
 508 patient ID, international classification of disease version 10 [ICD10], nomenclature property units [NPU],
 509 and anatomical therapeutic chemical [ATC]) facilitate easy data transformation commonly used for survival
 510 analyses and data-driven research (c).



511

512 **Figure 2.** Correlations between lymphoid-lineage cancer (LC) ICD10 diagnoses shown as pairwise Fisher's
 513 exact tests. The log odds ratio is indicated when the false detection rate (FDR) was above 0.1. ICD10 pairs
 514 with an FDR ≥ 0.1 are indicated in white. B cell lymphoma, BCL; chronic lymphocytic leukemia, CLL; diffuse
 515 large B cell lymphoma, DLBCL; extranodal marginal zone lymphoma, EMZL; follicular lymphoma, FL;
 516 Hodgkin lymphoma, HL; lymphoid leukemia, LL; lymphoplasmacytic lymphoma, LPL; mantle cell lymphoma,
 517 MCL; monoclonal gammopathy of uncertain significance, MGUS; multiple myeloma, MM; monoclonal B-cell
 518 lymphocytosis, MBL; non-Hodgkin lymphoma, NHL; nodular sclerosis, ns; not otherwise specified, NOS;
 519 small lymphocytic lymphoma, SLL; unspecified, UNS; Waldenström macroglobulinemia, WM.

520

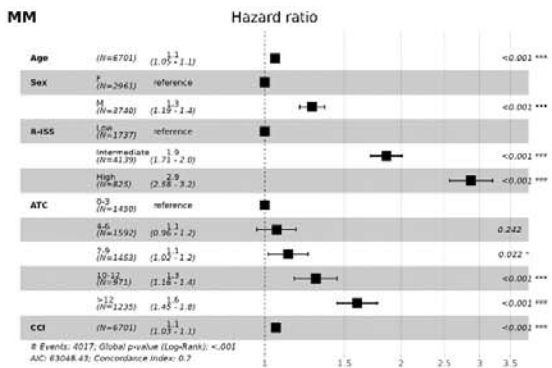
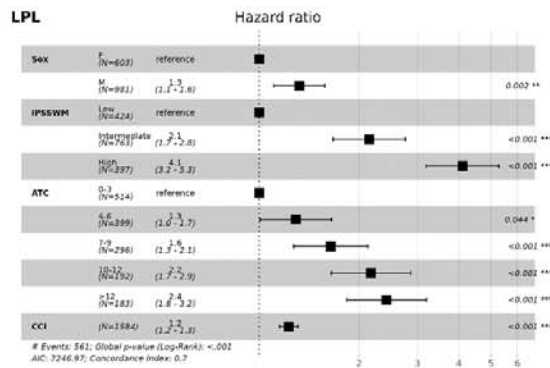
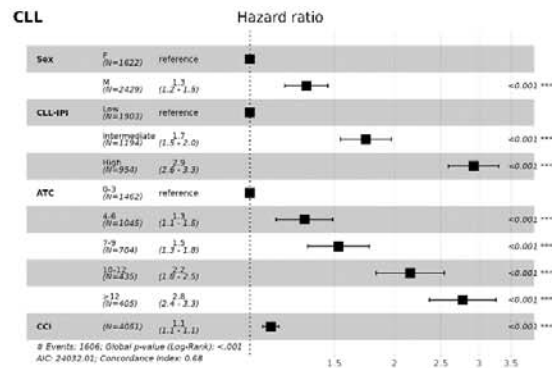
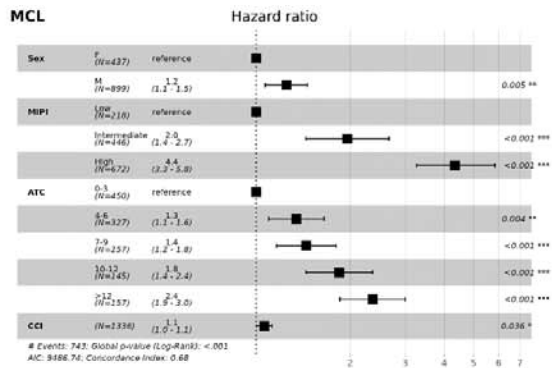
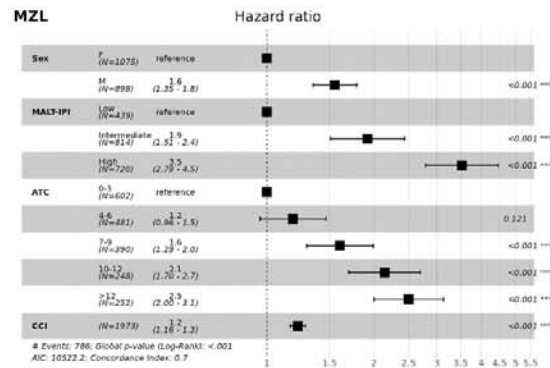
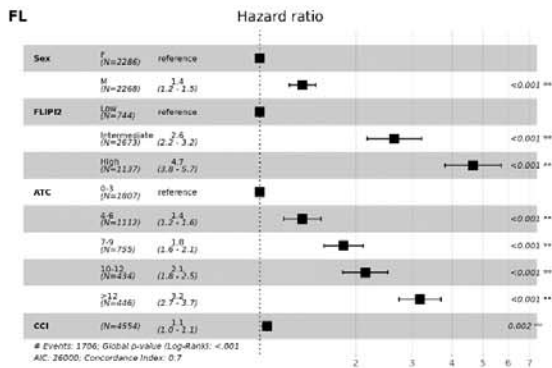
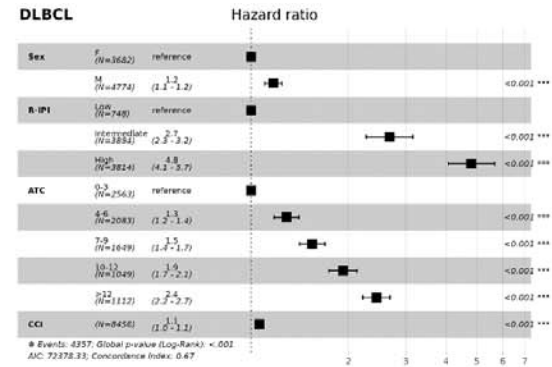
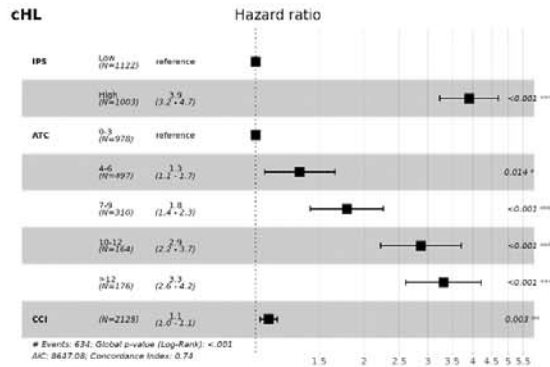
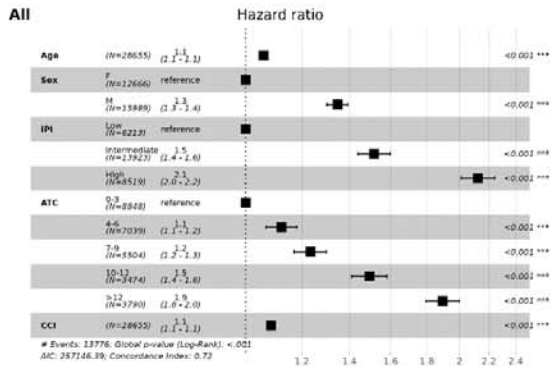


521

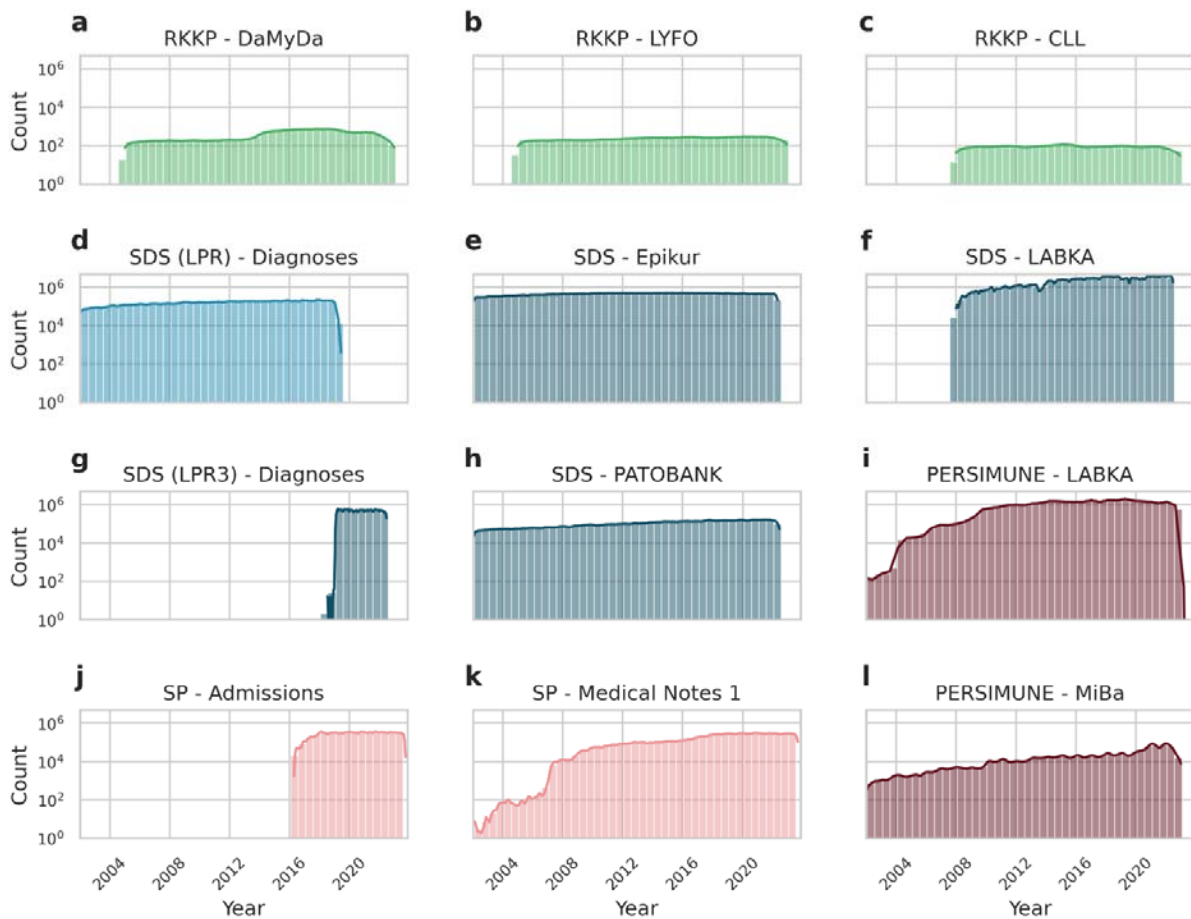
522 **Figure 3.** Unadjusted overall survival in 10 LC diagnoses. Burkitt lymphoma, BL; classical Hodgkin
 523 lymphoma, cHL; chronic lymphocytic leukemia, CLL; diffuse large B cell lymphoma, DLBCL; follicular
 524 lymphoma, FL; lymphoplasmacytic lymphoma, LPL; mantle cell lymphoma, MCL; multiple myeloma, MM;
 525 marginal zone lymphoma, MZL; not otherwise specified, NOS; small lymphocytic lymphoma, SLL.

526

527

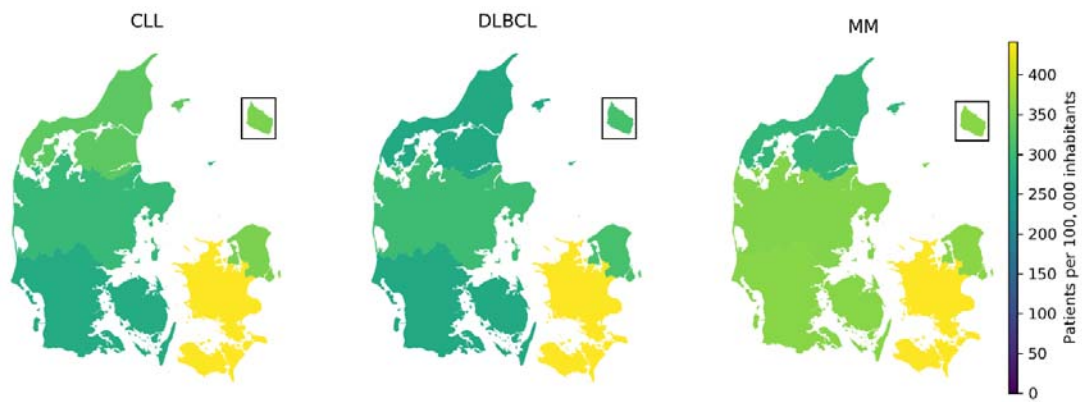


529 **Figure 4.** Multivariable analyses on age, sex, disease-specific international prognostic index (IPI), the number of anatomical therapeutic chemical (ATC)
530 codes from prescription in the year prior to diagnosis, and Charlson comorbidity index (CCI) score in all disease, classical Hodgkin lymphoma (cHL),
531 diffuse large B cell lymphoma (DLBCL), follicular lymphoma (FL), marginal zone lymphoma (MZL), mantle cell lymphoma (MCL), chronic lymphocytic
532 leukemia (CLL), lymphoplasmacytic lymphoma (LPL), and multiple myeloma (MM). In the pooled analysis of all disease, CLL-IPI high and very high risk
533 grouped into high IPI and revised international staging system (R-ISS) I, II, and III were labeled low, intermediate, and high risk, respectively.



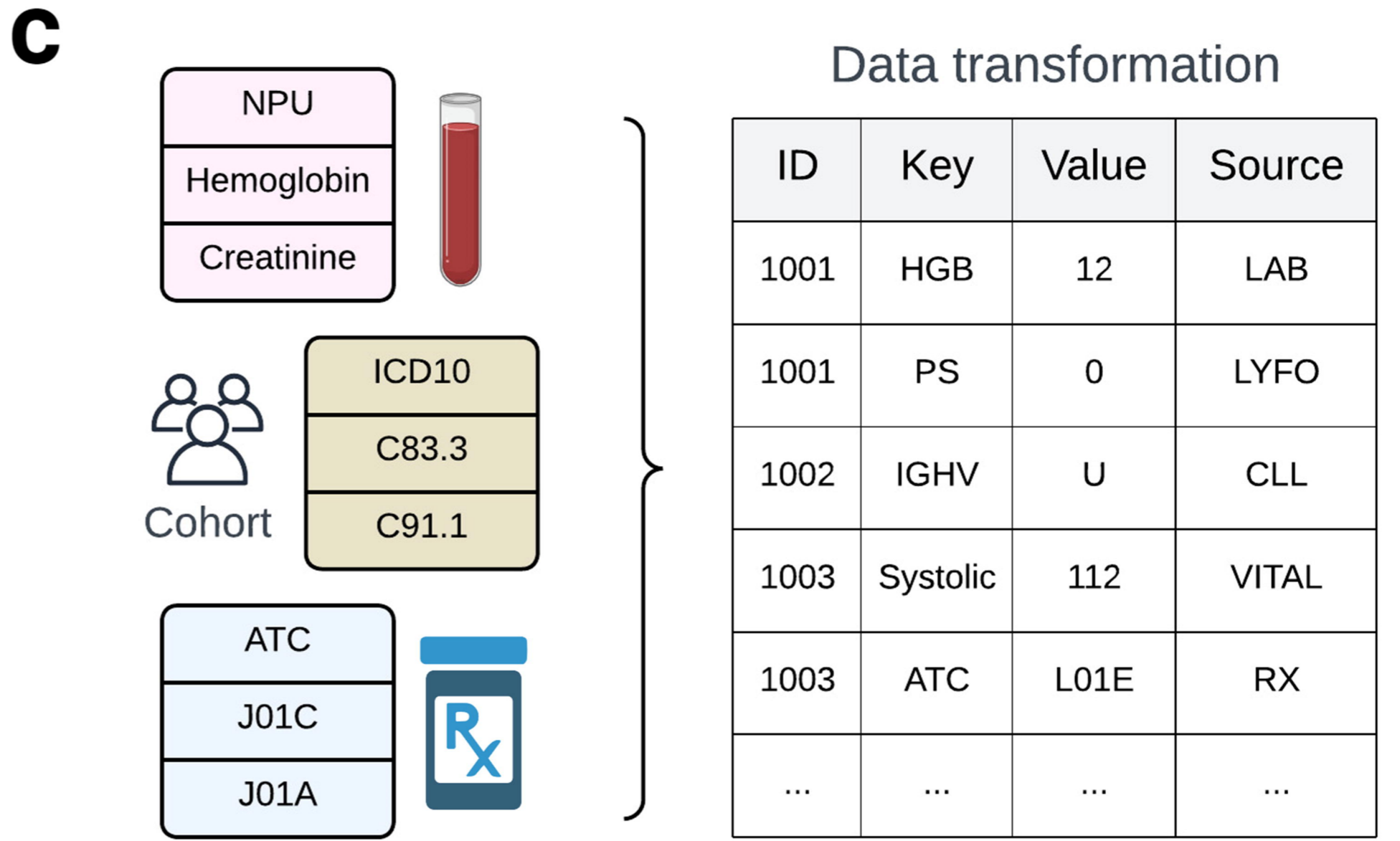
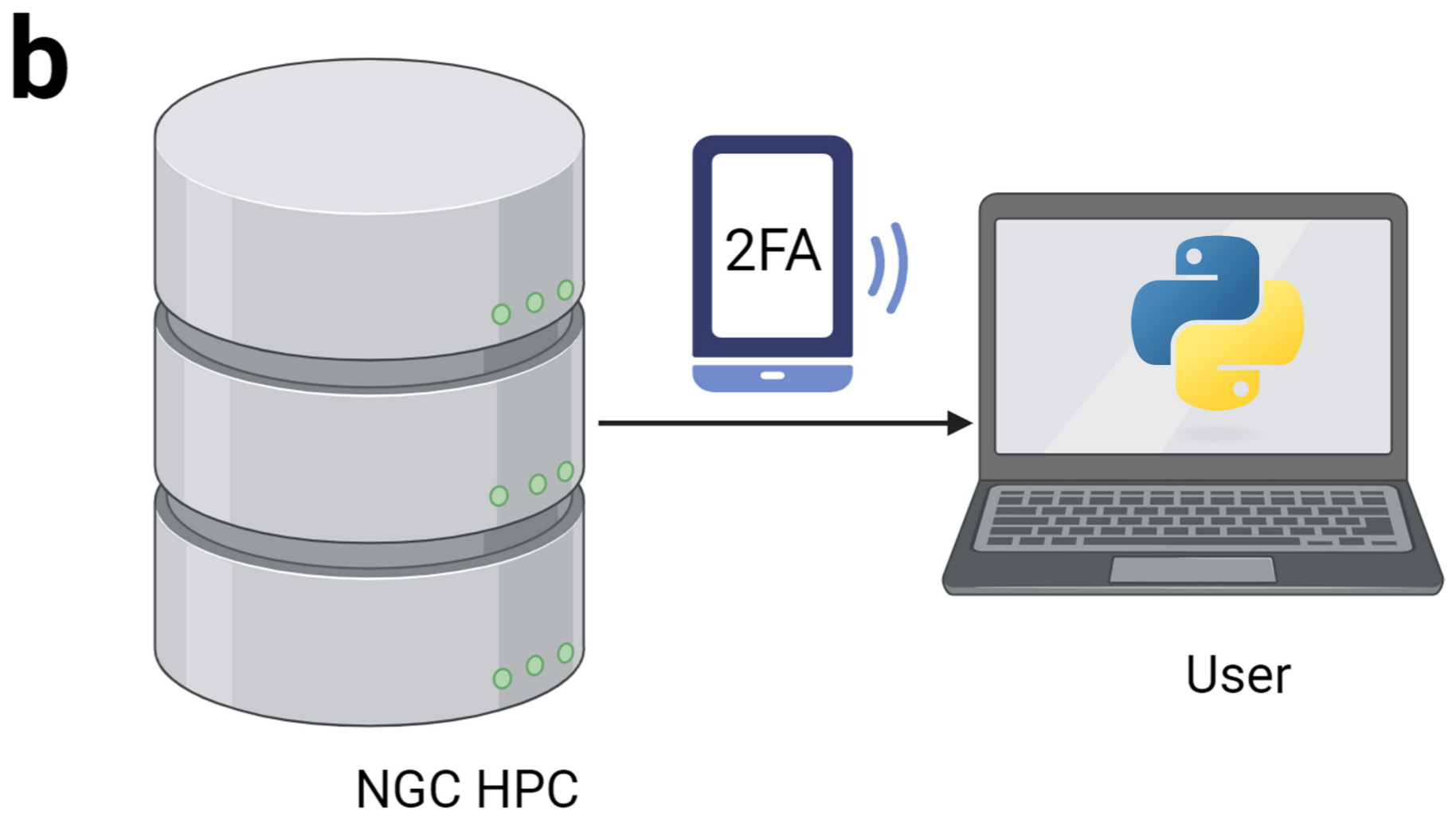
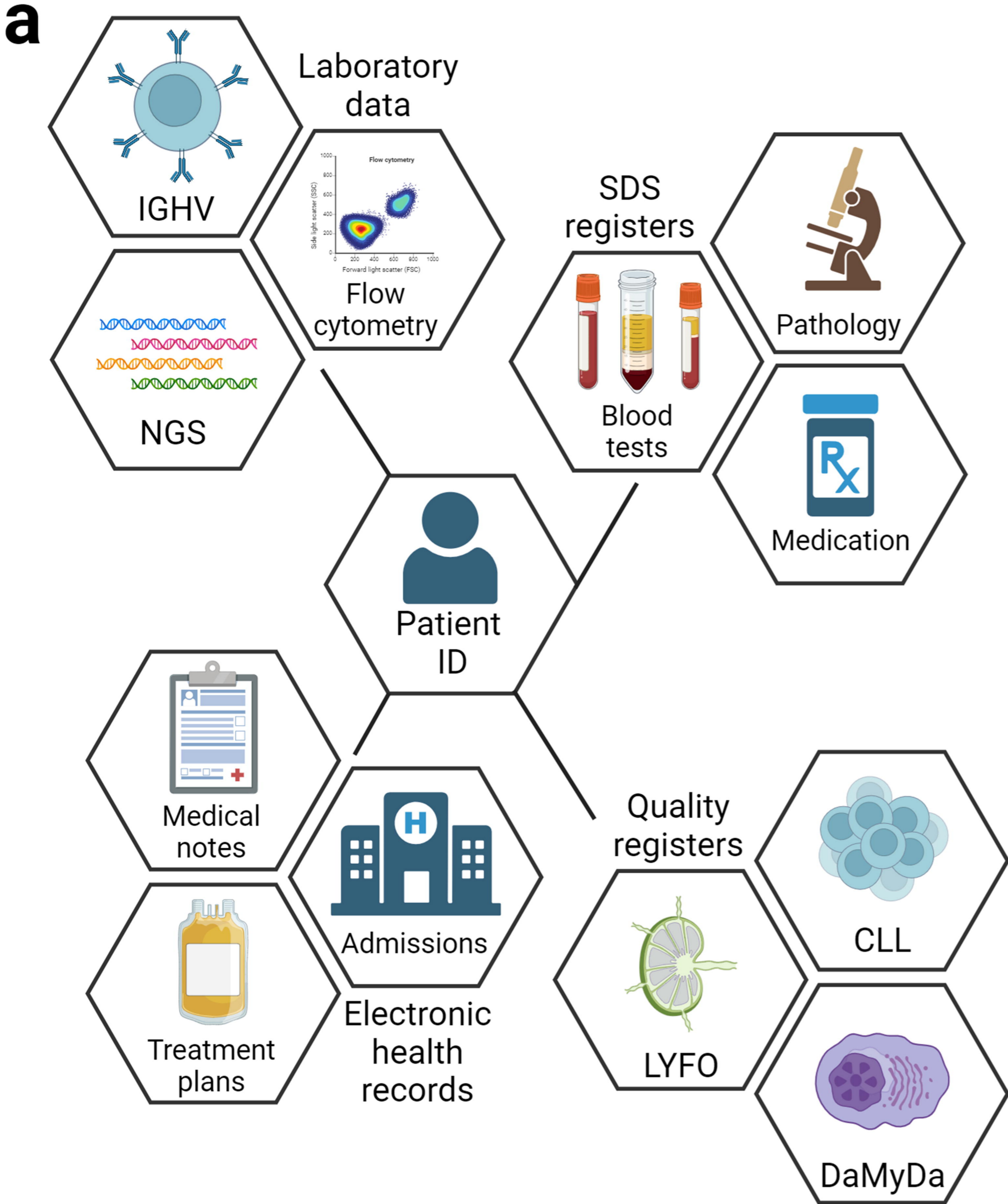
534

535 **Figure 5.** Overview of time coverage in 12 key datasets in the DALY-CARE data resource. RKKP
 536 hematological quality registers are wide format datasets, where data are typically entered upon
 537 diagnosis/registration, upon treatment, relapse, and follow-up (a-c). SDS LPR register was replaced by LPR3
 538 in Feb 2019 (d, g). Epikur (prescriptions), LABKA (blood tests), and PATOBANK (pathology requisitions) were
 539 independent of this update (e, f, h). We underscore that similar data from different data sources may have
 540 different time coverage (f, i). The electronic health record system of eastern Denmark (SP) went live in Mar
 541 2016 (j). Even so, medical notes antedating go-live dates from the previous EHR system were imported and
 542 are available as historic notes (k). A steadily increasing number of antimicrobial analyses was observed (l).
 543 Each panel shows the time coverage of a single dataset (a-l). Note that the y-axis shows the number of
 544 observations in three-month intervals on a log-scale. Lines above bins represent kernel density estimates of
 545 the counts.

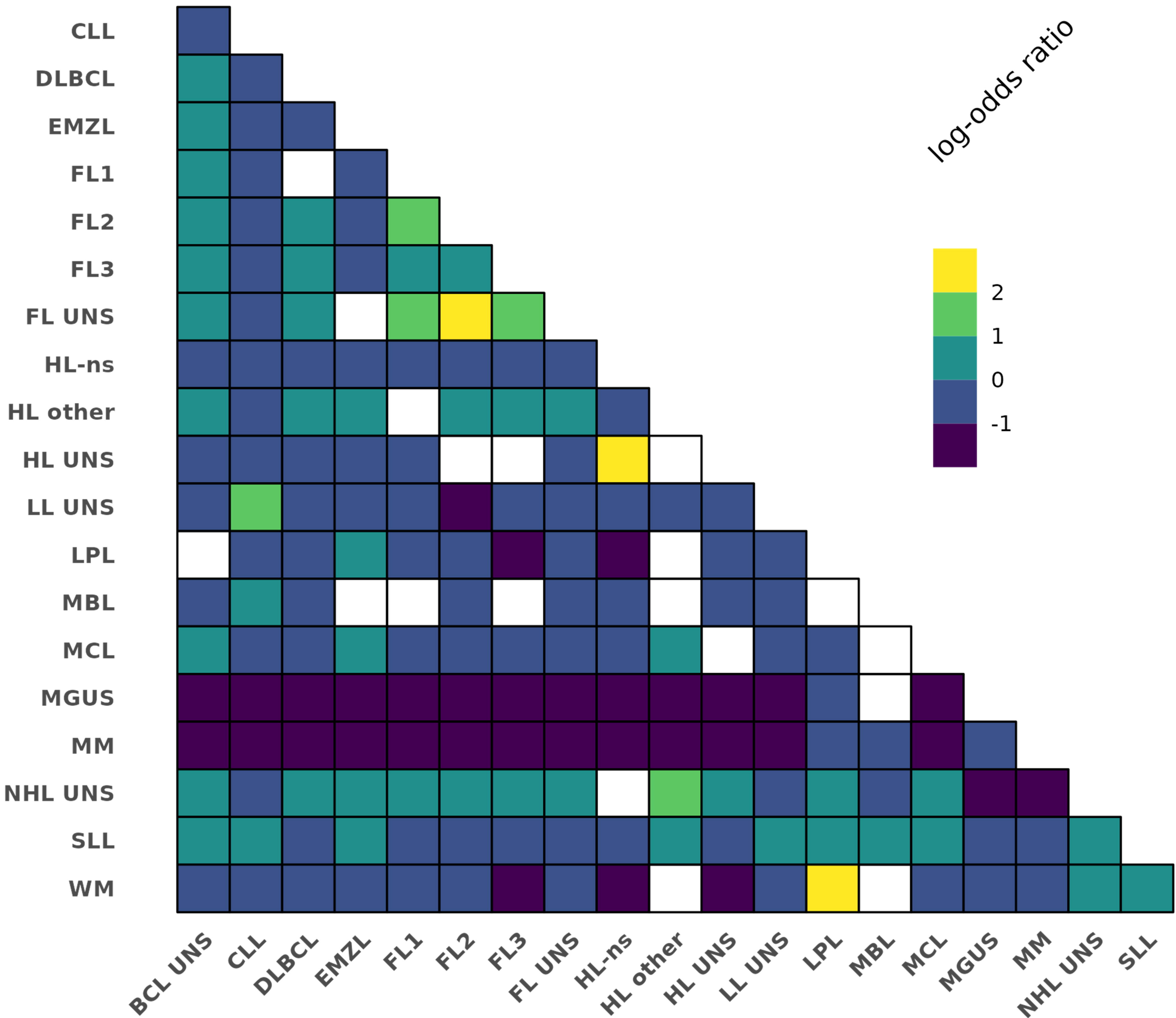


546

547 **Figure 6.** Number of patients per 100,000 residents in a) chronic lymphocytic leukemia (CLL), b) diffuse
 548 large B-cell lymphoma (DLBCL), and c) multiple myeloma (MM). Numbers included all patients in the DALY-
 549 CARE data resource regardless vital status.

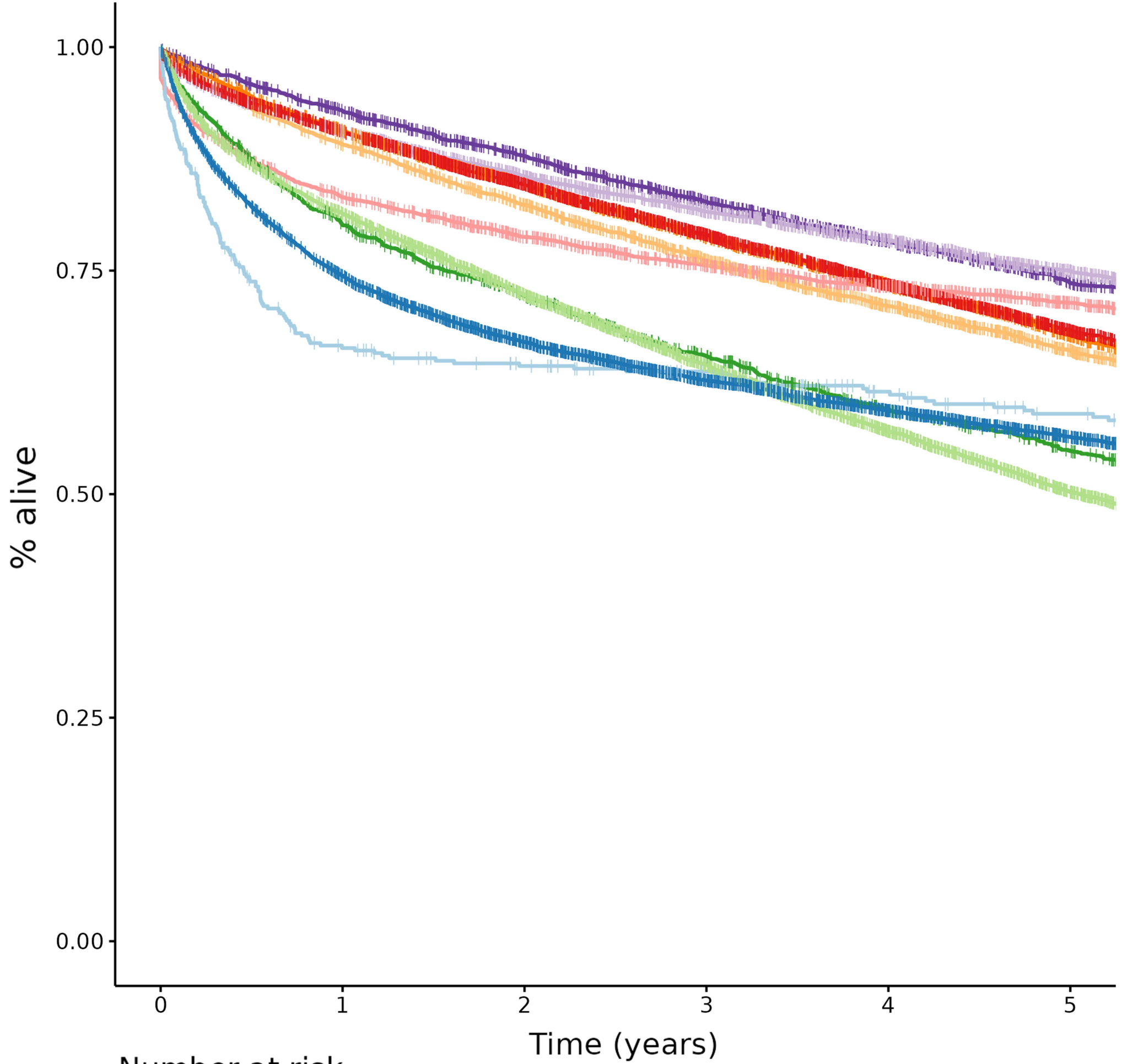


Correlogram of LC diagnoses



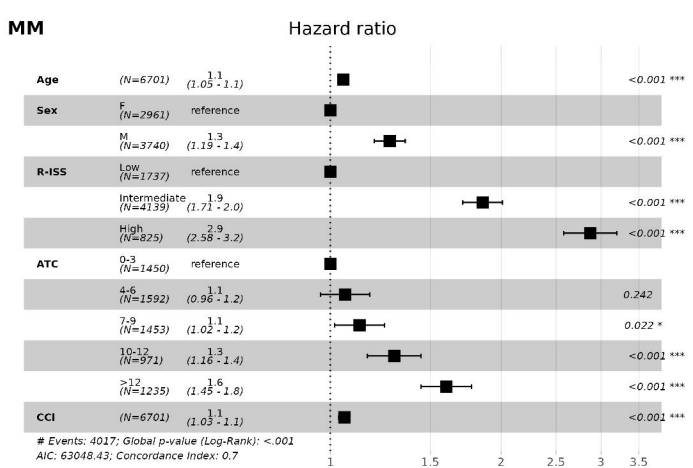
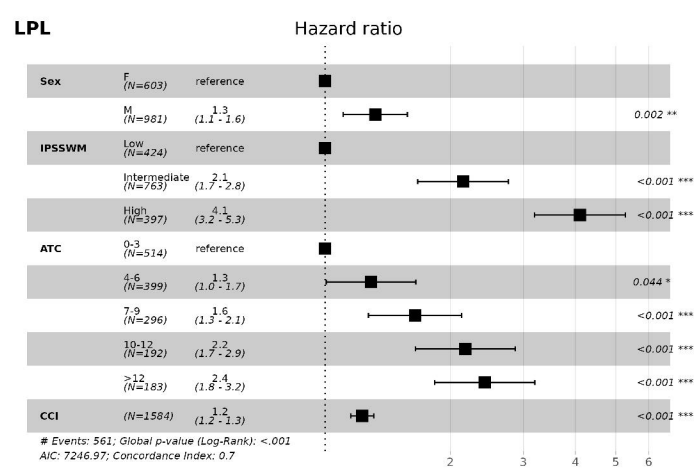
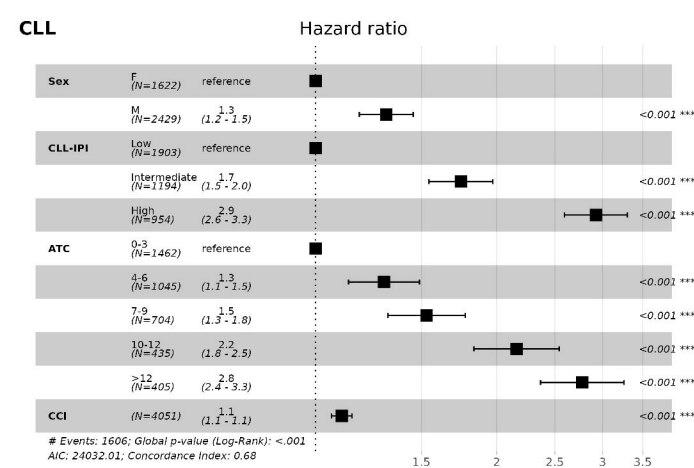
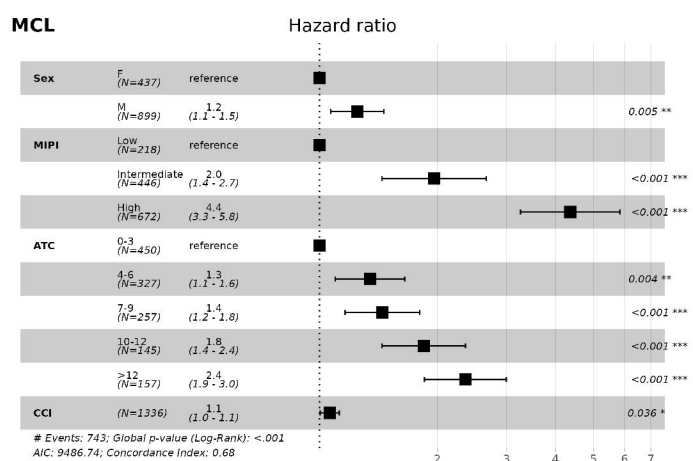
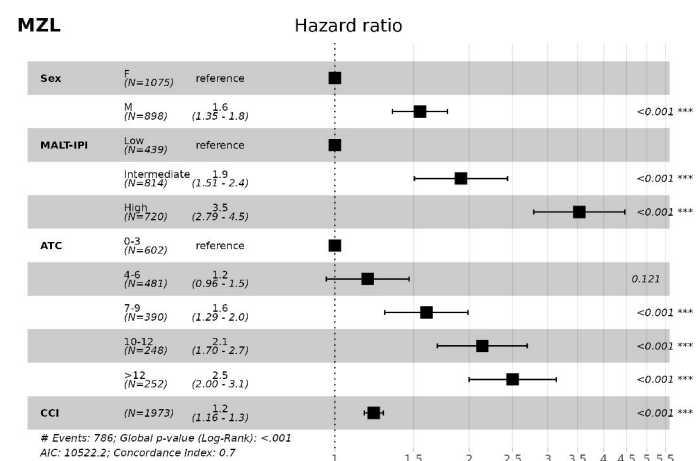
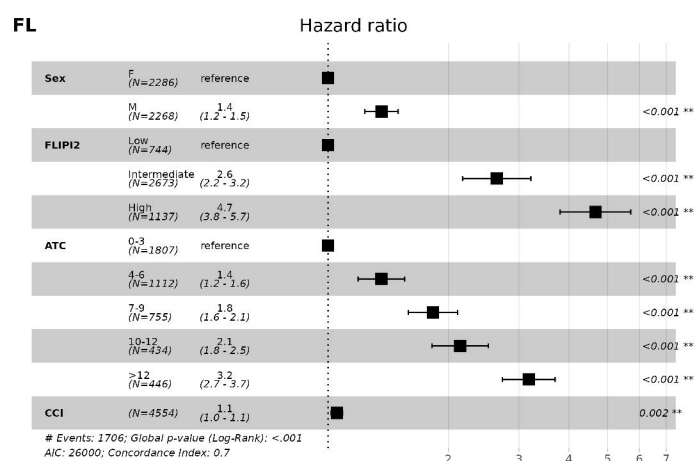
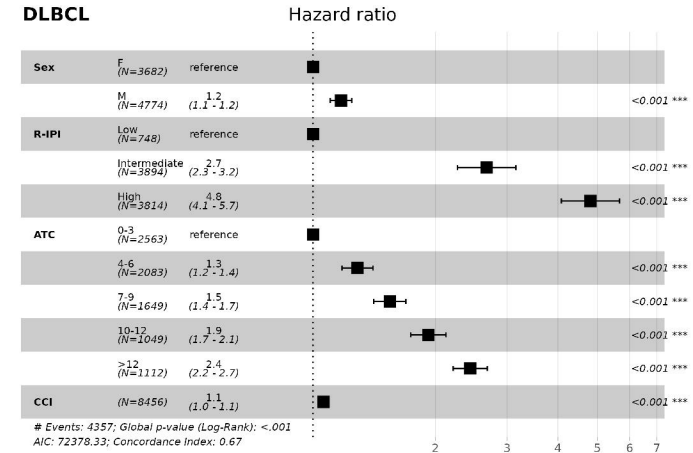
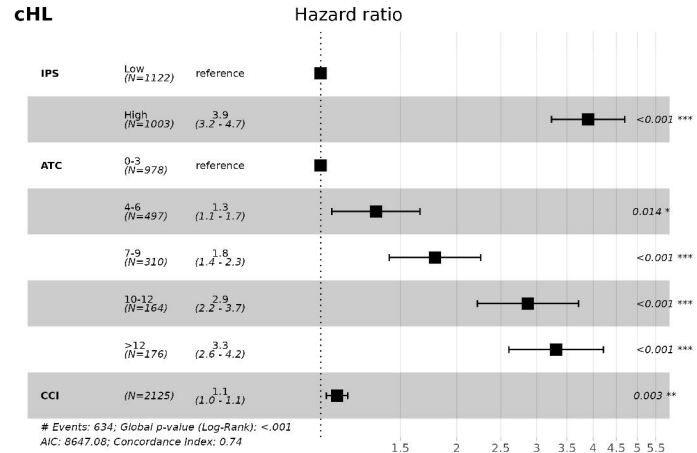
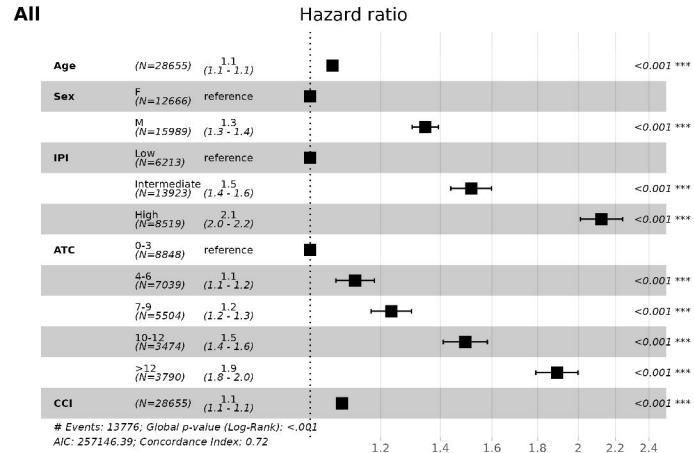
Disease

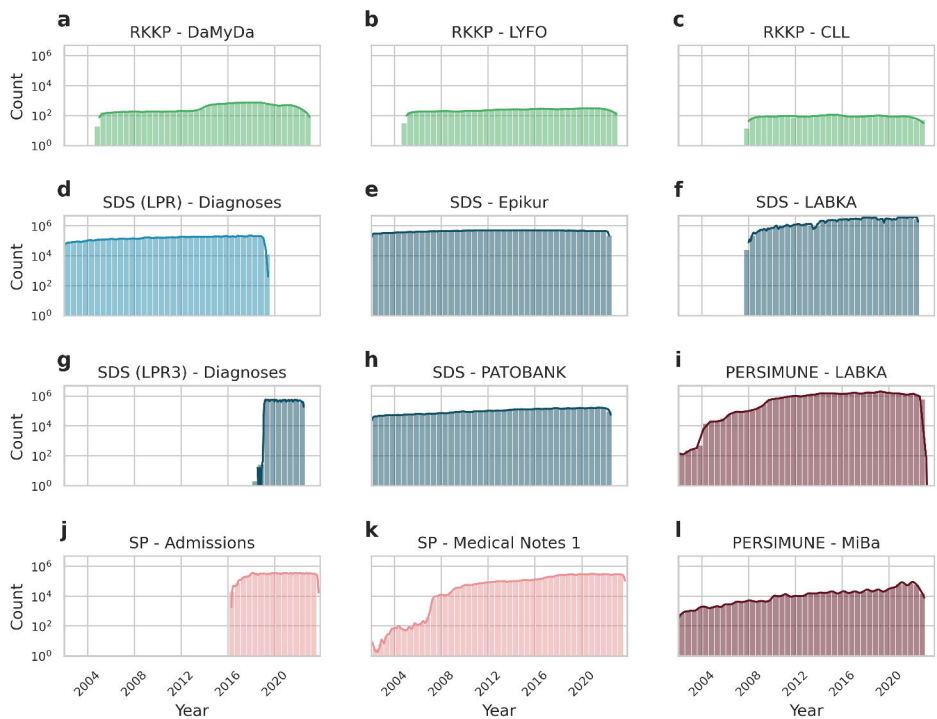
+	MZL	+	LPL	+	CLL	+	MCL	+	DLBCL
+	FL	+	SLL	+	cHL	+	MM	+	BL



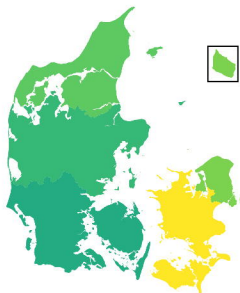
Number at risk

2713	2449	2191	1911	1658	1445
6483	5759	5218	4748	4317	3905
3721	3307	2936	2540	2214	1854
4965	4379	3941	3463	2997	2562
11682	10365	9302	8182	7182	6265
4078	3334	3040	2792	2555	2359
1793	1394	1196	1014	859	756
11340	9008	7597	6312	5182	4199
11844	8646	7392	6506	5779	5125
376	238	218	202	179	161

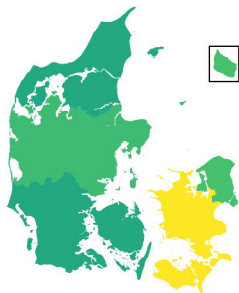




CLL



DLBCL



MM

