Clinical Epidemiology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

The Danish Lymphoid Cancer Research (DALY-CARE) Data Resource: The Basis for Developing Data-Driven Hematology

Christian Brieghel[1,2,*], Mikkel Werling[1,2,*], Casper Møller Frederiksen[1,2], Mehdi Parviz[1,3], Thomas Lacoppidan[1], Tereza Faitova[1,2], Rebecca Svanberg Teglgaard[1,4], Noomi Vainer[1,2], Caspar da Cunha-Bang[1,2], Emelie Curovic Rotbain[1,2], Rudi Agius[1], Carsten Utoft Niemann[1–3]

[1]Department of Hematology, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark; [2]Danish Cancer Institute, Copenhagen, Denmark; [3]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; [4]Department of Clinical Immunology, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark

*These authors contributed equally to this work

Correspondence: Christian Brieghel, Rigshospitalet, Department of Hematology, Blegdamsvej 9, Building 7054, Copenhagen Ø, 2100, Denmark, Tel +45 35453741, Email christian.brieghel@regionh.dk

Video abstract

Point your SmartPhone at the code above. If you have a QR code reader the video abstract will appear. Or use: https://youtu.be/0eA58JWF6LY

Background: Lymphoid-lineage cancers (LC; International Classification of Diseases, 10th edition [ICD10] C81.x-C90.x, C91.1-C91.9, C95.1, C95.7, C95.9, D47.2, D47.9B, and E85.8A) share many epidemiological and clinical features, which favor meta-learning when developing medical artificial intelligence (mAI). However, access to large, shared datasets is largely missing and limits mAI research.

Aim: Creating a large-scale data repository for patients with LC to develop data-driven hematology.

Methods: We gathered electronic health data and created open-source processing pipelines to create a comprehensive data resource for Danish LC Research (DALY-CARE) approved for epidemiological, molecular, and data-driven research.

Results: We included all Danish adults registered with LC diagnoses since 2002 (n=65,774) and combined 10 nationwide registers, electronic health records (EHR), and laboratory data on a high-powered cloud-computer to develop a secure research environment. Among other, data include treatments (ie 21,750 cytoreductive treatment plans, 21.3M outpatient prescriptions, and 12.7M in-hospital administrations), biochemical analyses (77.3M), comorbidity (14.8M ICD10 codes), pathology codes (4.5M), treatment procedures (8.3M), surgical procedures (1.0M), radiological examinations (3.3M), vital signs (18.3M values), and survival data. We herein describe the data infrastructure and exemplify how DALY-CARE has been used for molecular studies, real-world evidence to evaluate the efficacy of care, and mAI deployed directly into EHR systems.

Conclusion: The DALY-CARE data resource allows for the development of near real-time decision-support tools and extrapolation of clinical trial results to clinical practice, thereby improving care for patients with LC while facilitating streamlining of health data infrastructure across cohorts and medical specialties.

Keywords: large-scale database, data-driven medicine, chronic lymphocytic leukemia, multiple myeloma, lymphoma, machine learning

Introduction

Lymphoma, chronic lymphocytic leukemia (CLL), plasma cell dyscrasia (PCD) and their precursors are a highly heterogenous group of hematological malignancies in the blood, bone marrow and/or lymphoid tissues.[1] They span from precursor states such as monoclonal B-cell lymphocytosis (MBL) and monoclonal gammopathy of uncertain significance (MGUS), over indolent B-cell lymphomas (BCL), smoldering multiple myeloma (MM) and CLL, to

Received: 25 July 2024
Accepted: 19 December 2024
Published: 20 February 2025

aggressive high-grade BCL and PCD such as amyloid light-chain amyloidosis. Lymphoid-lineage cancers (LC) account for approximately 5% of all newly diagnosed cancers and more than 4% of all cancer deaths worldwide, whereas precursors such as MGUS and MBL may be identified in 5% to 12% of healthy, older individuals.[2,3]

Despite the pathological, molecular, and clinical heterogeneity of LC, they share several common features. First, all LC derive from or are directly caused by clonal lymphoid-lineage expansion in blood, bone marrow and/or lymphoid tissues. Second, they are diagnosed by histological tissue staining, flow cytometry, and/or molecular analyses, and the diagnostic workup typically includes biochemical blood tests, a bone marrow examination, and computer tomography (CT) or positron emission tomography (PET)/CT imaging. Third, LC share common epidemiology such as older age at diagnosis and male predominance, whereas socioeconomic status and lifestyle exposures are typically not well-correlated with LC. Fourth, LC develops due to dysregulation of immune function, stimulation and signaling in conjunction with acquisition of driver mutations.[4] Lastly, patients with LC precursor states, CLL, indolent lymphoma or smoldering MM do not benefit from pre-emptive treatment and are therefore often followed by a watch-and-wait strategy. If or when treatment is required, it usually includes corticosteroids, chemotherapy, and immunotherapy in combination, while targeted therapies in recent years are replacing or added to chemoimmunotherapy.[5] From an epidemiological, molecular, and data-driven perspective, the commonalities between LC offer a unique opportunity for identifying common features, which may facilitate modelling disease outcomes using meta-learning and federated learning. Achieving state-of-the-art accuracy of predictive models will require the joint modelling of diseases, ie including training data from patients with lymphoma and multiple myeloma may most likely improve the predictive accuracy for patients with CLL.[6]

The main driver in the recent successes of AI for image analysis and natural language processing outside of medicine is the critical mass of researchers working on the same benchmarks competitively and collaboratively. This ensures a continuous improvement of state-of-the-art performance on the given datasets.[7] By contrast, access to most medical data is heavily restricted. For instance, the performance of prognostic models in CLL has plateaued over the last 20 years likely due to lack of (1) multiple data modalities, (2) time-series modelling, and (3) a diverse set of researchers with access to the same benchmarks.[6] In turn, most mAI is developed on outcomes that are easily and publicly accessible, but not necessarily clinically useful in broader clinical setting (eg the MIMIC III and IV datasets).[6,8]

We therefore aim to address the aforementioned issues with lack of multimodality, common access, and clinically valid outcomes, and provide an overview of the DALY-CARE data resource, its infrastructure, and data formats.

Ethical Approvements

The DALY-CARE protocol has been approved by the Danish Health Data Authority and National Ethics Committee (approvals P-2020-561 and 1804410, respectively). Specifically, DALY-CARE is approved for collection of 1) biobank samples and 2) electronic health data (EHD) retrospectively and prospectively for all Danish adult residents diagnosed with LC in order to study the integrated clinicopathological profile of LC based on medical history, clinical and paraclinical (ie biochemical, cellular, microbiological, pathological, radiology, molecular, and genetic) routine data, while molecular omics (eg chip-array techniques, genomics, transcriptomics, proteomics, and microbiomics) and functional analyses (eg immune assays, phospho-flow cytometry, in vitro drug screening and mouse xenograft models) of adjoined biobank samples are covered by the protocol for correlations with the course of treatment and clinical outcome based on clinical epidemiology and data-driven mAI approaches.

Data Sources

The DALY-CARE data resource is a collection of EHD primarily gathered from existing data sources as summarized in Table 1. From the Danish Clinical Quality Program – National Clinical Registries (RKKP) registers, these include the Danish National Lymphoma Registry (LYFO)[9] since 2005, the Danish National Multiple Myeloma Database (DaMyDa)[10] since 2005, and the Danish National Chronic Lymphocytic Leukemia Registry (DCLLR)[11] since 2008. Exhaustive lists of variables in the RKKP registers are described elsewhere.[12] From the Danish Health Data Authority (SDS) since 2002, existing data sources included the Register of Pharmaceutical Sales covering prescription drug data (LSR),[13] the National Hospital Medication Register covering in-hospital medication (SMR, coverage since 2004),[14] the Danish National Pathology Register (PATOBANK) including pathology notes (free text) and SNOMED codes,[15] the

Table 1 Data sources in DALY-CARE

Data Source	Official Name (Abbreviation)		Dataset Name	Key Variable	Content of Primary Variable
RKKP	Danish National Lymphoma Registry (LYFO)		LYFO	patientid, Date_diagnosis, Date_treatment_1st_start	Contains date of firstline treatment
	Danish National Multiple Myeloma Database (DaMyDa)		MM	patientid, Date_diagnosis, Date_treatment_1_start_MPB	Contains date of firstline treatment
	Danish National Chronic Lymphocytic Leukemia Register (CLL)		CLL	patientid, Date_diagnosis, Date_treatment	Contains date of firstline treatment
SDS	Register of Pharmaceutical Sales (LSR)		EPIKUR	patientid, Date_diagnosis, atc	Contains ATC codes for prescriptions
			EKOKUR	atc	Contains ATC codes for prescriptions
	National Hospital Medication Register (SMR)		indberetningmedpris	c_atc	Contains atc codes for in-hospital medication
	Danish National Patient Registry (LPR)	Administrative data	t_adm (Backbone for LPR)	c_adiag	Contains primary ICD10 diagnosis codes
		Diagnoses	t_diag	c_diag	Contains ICD10 diagnoses codes
		Treatment - surgery	t_sksopr	c_opr	Contains surgical SKS K procedure codes
		Other procedures	t_sksube	c_opr	Contains SKS A, B, D, E, F, N, U, W, Y, Z codes
		Health care provider	t_udtilsgh	c_udtilsgh	Contains SHAK codes for health care provider
	Danish National Pathology Register (PATOBANK)		pato	c_snomedkode	Contains SNOMED pathology codes
		Microscopy	t_mikro	v_fritekst	Contains free-text pathology reports
		Conclusion/summary	t_konk	v_fritekst	Contains free-text pathology summaries
	Clinical Laboratory Information System Database (LABKA)		lab_forsker	analysiscode	Contains NPU codes from clinical biochemistry
	Danish Register of Causes of Death (DAR)		t_doedsaarsag	c_dod_1a	Contains ICD10 codes for primary cause of death
	Danish Cancer Register (DCR)		t_tumor	c_icd10	Contains cancer ICD10 codes
	Danish National Patient Registry 3 (LPR3)	Administrative data	kontakter (Backbone for LPR3)	dw_ek_kontakt;aktionsdiagnose	Contains ICD10 diagnoses
		Courses	forloeb (Backbone for LPR3)	dw_ek_forloeb	
		Diagnoses	diagnoser	diagnosekode	Contains ICD10 diagnoses codes
		Course markers	forloebsmarkoerer	markoer	Contains patient consent to research biopsy
		Orginasations	organisationer	sorenhed_tekst	Contains regional codes
		Treatment - surgery	procedure_kirurgi	procedurekode	Contains surgical SKS K procedure codes
		Other procedures	procedure_andre	procedurekode	Contains other SKS A, B, D, E, F, N, U, W, Y, Z codes
		Results	resultater	triggerkode	Contains SKS A, B, D, K, U and Z codes

(*Continued*)

https://doi.org/10.2147/CLEP.S479672

Table 1 (Continued).

Data Source	Official Name (Abbreviation)		Dataset Name	Key Variable	Content of Primary Variable
SP	Administrered Medicine (AdmMed)		Administreret_Medicin	atc	Contains ATC code for in-hospital administered treatment
	Admissions (ADT) - Hospitalization		ADT_Haendelser	admision_type;patient_class	Contains ctype and class of admission-dicharge-transfers
	Active In-hospital Diagnososes (ActiveDx)		Aktive_Problemliste_Diagnoser	current_icd10_list	Contains ICD10 diagnoses codes
	All Test Results (LABKA+)		AlleProvesvar	component	Contains NPU codes from clinical biochemistry
	Visits and Diagnoses (Visits And Dx)		Behandlingskontakter_diagnoser	skskode	Contains ICD10 diagnoses codes
	Treatment Plans 1 (Tx_plans1)		Behandlingsplaner_del1	protokol_navn	Contains treatment protocol name
	Treatment Plans 2 (Tx_plans2)		Behandlingsplaner_del2	serie_navn	Contains cycle number
	Microbiology charts 1 (Micro1)		Bloddyrkning_del1	komponentnavn	Contains name of microorganism
	Microbiology charts 2 (Micro2)		Bloddyrkning_del2	resultat_tekst	Contains microorganism found
	Microbiology charts 3 (Micro3)		Bloddyrkning_del3	organisme	Contains microorganism tested for resistance
	Microbiology charts 4 (Micro4)		Bloddyrkning_del4	kliniske_oplysninger	Contains clinical history
	Flytningshistorik (Transfers)		Flytningshistorik	haendelse	Contains admission event
	Intensive Care Unit (ICU)		ITAOphold	icu_stay_start	Contains date of ICU admission
	Medical Notes (Notes 1)		Journalnotater_del1	notat_text	Contains free-text of medical note
	Medical Notes (Notes 2)		Journalnotater_del2	handling	Contains signature of medical note
	Prescribed Medicine (RxMed)		OrdineretMedicin	atc	Contains ATC codes for hospital prescriptions
	PatientInfo		PatientInfo	sex_c	Contains information on sex
	Social History		SocialHX	ryger;drikker	Contains history of smoking and alchol consumption
	Vital Signs (VitalSigns)		VitaleVaerdier	displayname	Contains name of vital sign
	Do-not-resuscitate (DNR)		Behandlingsniveau	behanldingsniveau	Contains DNR orders
	Medical imaging 1		BilleddiagnostikeUndersøgelser_Del1	bestillingsnavn	Contains type of imaging
	Medical imaging 2		BilleddiagnostikeUndersøgelser_Del2	beskrivelse	Contains free-text descriptions
PERSIMUNE	Danish Microbiology Database (MiBa)	microbiology_analysis	microbiology_analysis	micanalysiscode	Contains microbiology analysiscode
		microbiology_culture	microbiology_culture	c_microorganismidentificationcode	Contains MORG code from cultured result

https://doi.org/10.2147/CLEP.S479672

		microbiology_culture_resistance	microbiology_culture_resistance	susceptibility	Contains information on antimicrobial resistence
		microbiology_microscopy	microbiology_microscopy	resultsummary	Contains mircoscopy result
	Clinical Laboratory Information System Research Database (LABKA)		biochemistry	c_analysiscode	Contains NPU codes from clinical biochemistry
Laboratory	IGHV analyses		IGHVIMGT	IGHV	Contains IGHV mutational status
	Flow cytometry		Flowcytometry	CONCLUSION	Contains summary from flow cytometry analysis
	FISH analyses		FISH	INIT#3	Contains summary from FISH analysis
	CLL NGS panel		CLLPANEL_WIDE	N.drivers	Contains no. of drivers mutated
	Biobank samples		BIOBANK_SAMPLES	date_samplecollection	Contains date of sample collection for available biobank samples
Curated	CLL_2ND_LINE_TREATMENT		CLL_TREAT	DATE2LINE	Contains date of initiating 2nd line treatment
	MM_TREATED_WITH_DARATUMUAB		MM_TREAT_DARA	linjestart	Contains date of initiating daratumumab
	CLL_TREATED_WTIH_IBRUTINIB		CLL_TREAT_IBRUTINIB	ibrutinib_start	Contains date of initiating ibrutinib
	DANRICHT		DANRICHT	date_RT	Contains date of Richter transformation

https://doi.org/10.2147/CLEP.S479672

Clinical Laboratory Information System Database with routine laboratory results (LABKA),[16] the Danish Register of Causes of Death (DAR),[17] the Danish National Patient Registry (LPR) versions 1 and 3,[18,19] and the Danish Cancer Registry (DCR).[20]

We further retrieved EHD from 16 modules in the EPIC®-based EHR system in eastern Denmark (*Sundhedsplatformen* [SP]) including 1) administered medicine, 2) admissions (ADT), 3) active in-hospital diagnoses, 4) all test results, 5) hematology/oncology treatment plans, 6) microbiology charts, 7) transfers between departments, 8) intensive care unit admissions, 9) medical notes (as free text), 10) prescribed medicine, 11) out- and in-patient visits and diagnoses, 12) anthropometrics, 13) a social history of smoking and alcohol consumption, and 14) vital signs, 15) do-not-resuscitate (DNR) orders, and 16) medical imaging (Supplementary Table S1 and Supplementary Material).

Additionally, we gathered EHD indirectly from laboratory systems at our institution. These included cleaned versions of the Danish Microbiology Database (MiBa)[21] and LABKA retrieved through the Personalised Medicine of Infectious Complication in Immune Deficiency (PERSIMUNE) covering the Capital Region of Denmark (Table 1).[22] In addition, we added summary reports and results directly from the laboratory systems for routine flow cytometry analyses,[23] immunoglobulin heavy-chain variable gene (IGHV) analyses including stereotypic subset designation,[24] targeted next-generation sequencing (tNGS),[25] and fluorescence in situ hybridization (FISH) reports. Manually curated datasets collected through EHR reviews include information on patients receiving second-line CLL treatment, patients with CLL treated with ibrutinib, patients with MM treated with daratumumab, and patients with CLL developing Richter's transformation from previously published cohorts.[26–29]

Finally, the DALY-CARE protocol specifically allows for functional and molecular analyses including omics of biobank samples from individuals in the cohort. For this purpose, we gathered available biobank information from four large Danish biobanks, namely the CLL biobank, the Copenhagen Hospital Biobank, the Danish Cancer Biobank, and PERSIMUNE biobank (Table 1).[30] Imputed genotypes from peripheral blood at time of first hospital blood-workup are availble in more than 11,000 patients (Supplementary Material).[31]

Main Variables

Variables in existing data sources have been described elsewhere (Supplementary Table S2).[9–11,13–17,19,21] In short, baseline demographic and prognostic information (eg clinical stage, biochemical, and molecular data), as well as information on treatment, response, and survival are assembled from the RKKP registers. The main variables use the Danish Medical Classification System (SKS) encoding[32] that, in short, covers diagnoses (D-SKS or ICD10 codes), medicine (M-SKS codes or anatomical therapeutic chemical [ATC]), surgical (K-SKS codes), radiological (UX SKS codes) and treatment procedures (B-SKS codes) including specific antineoplastic, immunotherapy, and radiotherapy codes (ie BWH, BOJH1, and BWGC SKS codes, respectively).[19] Further, the biochemical data use nomenclature property units (NPU) codes, health care providers use SHAK codes, and pathology data use SNOMED codes and free-text notes (Supplementary Material). These codes facilitate clear definitions and easy mapping across different datasets (Supplementary Tables S3 and S4). All diagnoses, medicine, biochemistry, and procedures across all available datasets have been gathered into independent aggregated tables. Data do not include sensitive information on ethnicity, religion, or socio-economic status.

Database Infrastructure

The DALY-CARE data resource is based on open source scalable PostgreSQL located inside a secure ISO 27001 certified private cloud on a high-performance supercomputer at the Danish National Genome Center (NGC) research infrastructure. The server is accessed via a Federal Information Processing Standard (FIPS) complaint virtual desktop interface (VDI) that provides full separation between users. Built specifically with data security and privacy in mind, the platform uses 2-factor authentication (2FA) and layered security approach. The platform is on-premises and does not use any components that are hosted outside the infrastructure, which ensures data security and compliance with the statutory acts.[33]

All data are pseudonymized based on a single patient ID linked to the Danish Civil Registration System.[34] This allows for unique identification of all patients and enables easy data linkage across all datasets while protecting individual patients' data privacy (Figure 1a). All raw data are available directly in the DALY-CARE database. Newly

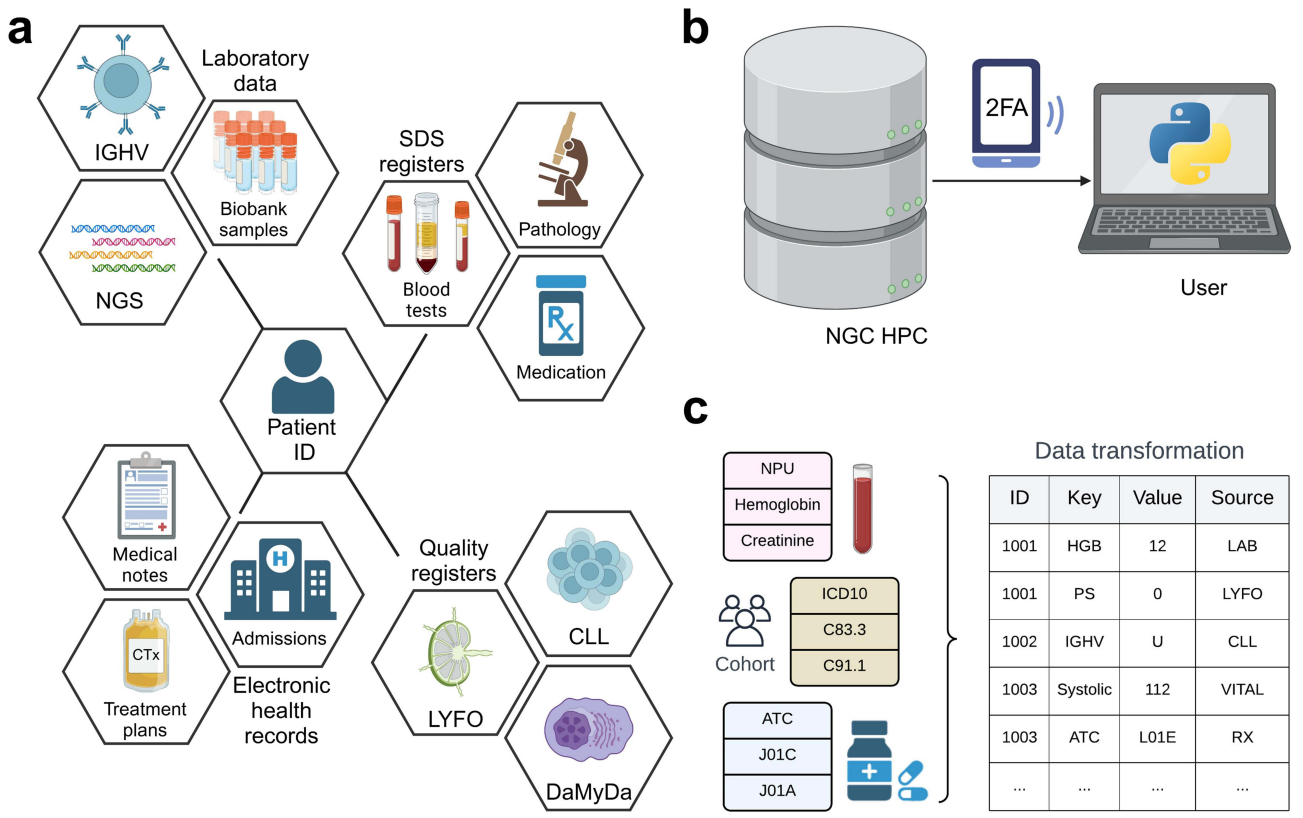

Figure 1 Schematic overview of the Danish Lymphoid Cancer Research (DALY-CARE) data resource. Data from various sources including Danish nationwide registers, hematological quality registers, electronic health record data, and hematological laboratories may be linked using a single pseudonymized patient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

retrieved data are quality assessed, and previous data cuts are logged to ensure the reproducibility of all studies. All RKKP, SDS, and EHR data are updated once yearly, and EHR data may be updated upon request.

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

1. *import*
   a. *public* contains raw data retrieved from RKKP, SDS, EHR, and PERSIMUNE.
   b. *laboratory* contains data from hematological laboratories including flow cytometry, FISH, IGHV, and NGS data.
   c. *lookup_tables* contain tables to map encoded data including SNOMED, SKS, ATC, ICD10, NPU, and SHAK.
2. *core*
   a. *public* contains cleaned versions of raw datasets and aggregated data from multiple data sources.
   b. *curated* contains manually curated datasets obtained from manual EHR reviews and NGS data.
   c. *lookup_tables* contain cleaned tables to map standard coding.

We created data processing pipelines and functions for loading and processing the data in both R and Python software (Figure 1b and Supplementary Material). This includes pipelines that transform the raw data into a format ready to use by predictive models (feature generation) and scripts that define and extract clinical outcomes from the raw datasets. We grouped functions based on their specific purpose inside or general use outside the DALY-CARE server to allow for synergy in larger collaboratives of Danish epidemiologists and data scientists studying similar register data as well as for adaptation to similar international data. Importantly, the pipelines in R and Python software allow for direct database queries to quickly load data using indexed variables (Supplementary Table S3).[35,36]

https://doi.org/10.2147/CLEP.S479672

Study Population

We included all Danish adult residents registered with an LC diagnosis in 1) the RKKP registers, 2) the SDS registers, and 3) EHR data since 1 January 2002 using International Classification of Diseases version 10 (ICD10; ie C81.x-C90.x, C91.1-C91.9, C95.1, C95.7, C95.9, D47.2, D47.9B, and E85.8A) and Systematized Nomenclature of Medicine (SNOMED) codes, which could be mapped to ICD10 codes for LCs (Supplementary Table S5).[37] The nationwide coverage for malignant LC diagnoses within RKKP has previously been estimated to 99%.[10,11,38] The cutoff date for inclusion of data was 15 Nov 2023.

In total, we included 65,744 individuals with an LC of whom 35,399 (53.8%) had more than one LC diagnosis: 15,838 (24.1%), 10,420 (15.8%), 5315 (8.1%), and 3826 (5.8%) patients had 2, 3, 4, and >4 different LC diagnoses, respectively (Supplementary Figure S1). Most patients with multiple LC diagnoses were initially diagnosed with unspecified LC (eg C81.9, C82.9, C85.5, C85.9, and C91.9) followed by specified subclassification, whereas fewer patients had multiple diseases, progressed from precursor states to overt LC or transformed to more aggressive disease (Figure 2). The 20 most common LC diagnoses could be attributed to 62,946 patients (95.7%) and are summarized in Table 2. Please see an exhaustive list in Supplementary Table S6 and note that each patient may be registered with multiple LC diagnoses.

At time of first LC diagnosis for patients with common LC entities (Supplementary Table S7), the median age was 70.3 years (interquartile range [IQR] 60.7;77.9) and 56.1% were male (Table 3). A detailed description of the patients and baseline characteristics is available in the Supplementary Material. We further calculated baseline polypharmacy defined as <5 vs ≥5 prescriptions in the year prior to first LC diagnosis, Charlson comorbidity index (CCI) scores (Supplementary Table S8), and disease-specific international prognostic indices (Supplementary Table S9).[39,40] From the cause of death register, the RKKP registers, and EHR data, we compiled dates of last follow-up or death and calculated time from first LC diagnosis to death or

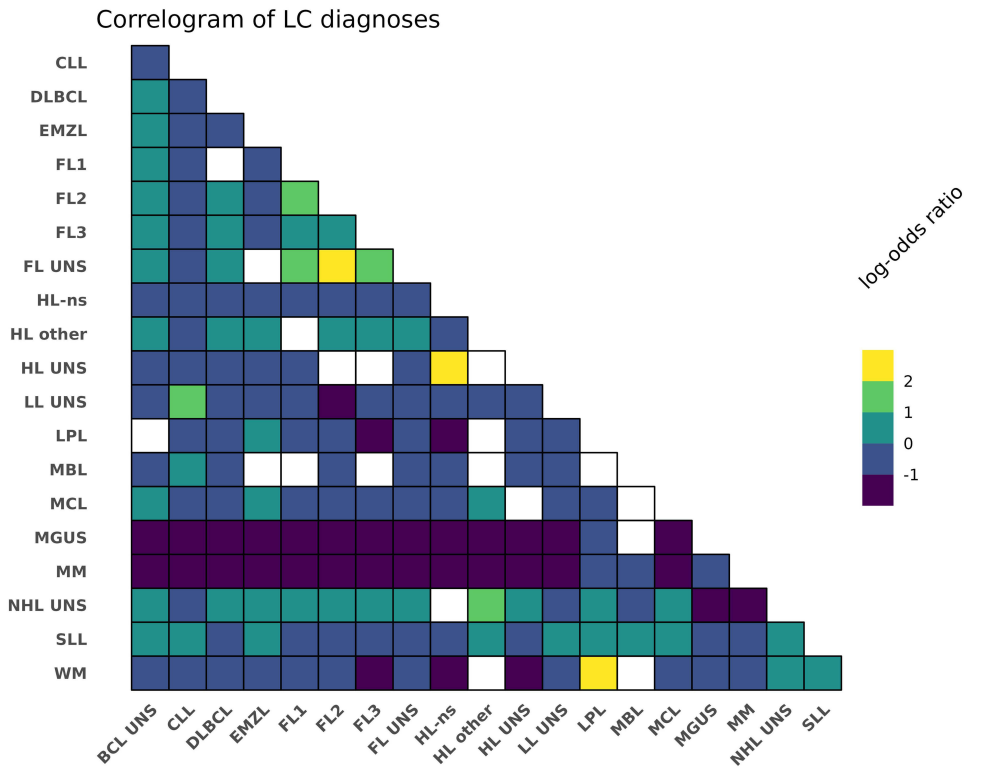

Figure 2 Correlations between lymphoid-lineage cancer (LC) ICD10 diagnoses shown as pairwise Fisher's exact tests. The log odds ratio is indicated when the false detection rate (FDR) was above 0.1. ICD10 pairs with an FDR ≥0.1 are indicated in white.

Abbreviations: BCL, B-cell lymphoma; CLL, chronic lymphocytic leukemia; DLBCL, diffuse large B-cell lymphoma; EMZL, extranodal marginal zone lymphoma; FL, follicular lymphoma; HL, Hodgkin lymphoma; LL, lymphoid leukemia; LPL, lymphoplasmacytic lymphoma; MCL, mantle cell lymphoma; MGUS, monoclonal gammopathy of uncertain significance; MM, multiple myeloma; MBL, monoclonal B-cell lymphocytosis; NHL, non-Hodgkin lymphoma; ns, nodular sclerosis; NOS, not otherwise specified; SLL, small lymphocytic lymphoma; UNS, unspecified; WM, Waldenström macroglobulinemia.

Table 2 Common DALY-CARE diagnoses

Disease	ICD10	n	freq
Non-Hodgkin lymphoma, unspecified	DC85.9	21321	32.4%
Monoclonal gammopathy of undetermined significance (MGUS)	DD47.2	15399	23.4%
Diffuse large B-cell lymphoma	DC83.3	11813	18.0%
Chronic lymphocytic leukaemia of B-cell type	DC91.1	11724	17.8%
Multiple myeloma	DC90.0	11410	17.3%
B cell lymphoma, UNS	DC85.1	8107	12.3%
Follicular lymphoma, unspecified	DC82.9	5713	8.7%
Small cell B-cell lymphoma	DC83.0	4988	7.6%
Hodgkin lymphoma, unspecified	DC81.9	3416	5.2%
Lymphoplasmacytic lymphoma	DC83.0B	3162	4.8%
Waldenström macroglobulinaemia	DC88.0	3080	4.7%
Follicular lymphoma grade II	DC82.1	2761	4.2%
Mucosa-associated lymphoid tissue (MALT) lymphoma	DC88.4	2105	3.2%
Nodular sclerosis classical Hodgkin lymphoma	DC81.1	1915	2.9%
Mantle cell lymphoma	DC83.1	1802	2.7%
Other specified types of non-Hodgkin lymphoma	DC85.7	1726	2.6%
Follicular lymphoma grade I	DC82.0	1697	2.6%
Monoclonal B cell lymphocytosis (MBL)	DD479.B	1560	2.4%
Lymphoid leukaemia, unspecified	DC91.9	1361	2.1%
Follicular lymphoma grade III, unspecified	DC82.2	1281	1.9%

Table 3 Patient characteristics

Variable	Level	Total
Age, years	median [iqr]	70.3 [60.7, 77.9]
Sex	Female	28,867 (43.9)
	Male	36,907 (56.1)
CCI score	median [iqr]	2 [2, 3]
No. of ATC	median [iqr]	6 [3, 10]
	missing	3008
Polypharmacy	Yes	38,501 (61.3)
	No	24,265 (38.7)
	missing	3008

https://doi.org/10.2147/CLEP.S479672

end of follow-up. As a result, survival data are readily available (Supplementary Figures S2 and S3). We underscore that by including LC diagnosis codes from in-patient registers (ie LPR and EHR), the observed number of patients will represent the combination of incident and prevalent cases.

Biobank Material

On top of the register and EHR data, we also included laboratory data as indicated in Table 1. To facilitate large-scale omic studies in DALY-CARE, we searched four large Danish biobanks and identified 40,863 biobank samples either stored on different dates or collected from different tissues from 18,528 individuals.[30] This included 21,431 samples among 9636 patients with lymphoma (ICD10 codes C81.x-C89.x), 11,043 samples in 3809 patients with CLL (ICD10 code C91.1), and 6206 samples in 3216 patients with PCD (ICD10 codes C90.0-C90.3, D47.2, and E85.8A). To create a large genetic repository for future studies, genotyping (n>12,000), whole-genome sequencing (n>2000), and proteomics (n>800) is ongoing.

Disparities in Data and Coverage

To avoid bias when linking different datasets, we here provide an overview of coverage over time for data from RKKP, SDS, EHR, and PERSIMUNE datasets. Temporal coverage and number of unique patients for 45 datasets are provided in Supplementary Figures S4–S9. Generally, more recent data show better coverage than older data: 13 out of 54 data sources plotted date back to 2002, while all 54 data sources have data available for 2019 (Figure 3 and Supplementary Figure S4). While all data are population-based, data collection was centered around eastern Denmark and our institution in specific. We thus calculated the prevalence of CLL, lymphoma, and PCD disease categories in the data available in DALY-CARE to assess regional disparities, which would likely represent differences in data coverage (Figure 4). For instance, differences in regional coverage over time reveal better coverage in the Capital Region for biochemical data such as LABKA (Supplementary Figures S4–S8 panels F1 and H1).

Previous DALY-CARE Usage

Due to the variety of data sources and data modalities, the DALY-CARE data resource may facilitate a diverse set of research projects. To provide an overview of the unique possibilities of DALY-CARE, we highlight examples of previously published work in Supplementary Table S10 including use of biobank data to find novel prognostic markers, investigation of real-world evidence of prognostic factors influencing the efficacy of health care, and development and deployment of mAI algorithms (Supplemental Material). Notably, we developed the data driven model CLL-TIM for predicting infection and/or CLL treatment in newly diagnosed CLL. This model was implemented into the ongoing PreVent-ACaLL clinical trial (NCT03868722) and deployed into the EHR system of eastern Denmark.[41]

Data Availability

According to Danish legislation, EHD may not be shared publicly. Within the approved protocol (available on Github), the data may be shared on a collaborative basis based on a data processing agreement (please see DPA template available on Github). All analyses must be performed on the DALY-CARE data resource at the NGC. Except for free-text tables (ie pathology, medical imaging, and medical notes), dummy tables and the underlying code for this study is publicly available on Github.

Improvements in patient care are intrinsically tied to the ability to access and share data easily.[42] We here demonstrate how DALY-CARE allows researchers to identify complex patterns across different datasets, patient cohorts, and subgroups on a collaborative basis. By making the DALY-CARE data resource FAIR (findable, accessible, interoperable, and reusable), we provide the opportunity to share definitions, outcomes and ultimately data that will allow for competitive modeling on outcomes that likely increase the predictive performance of prognostic models. To translate these data-driven perspectives into benefits for patients on a national and international level, we believe it is crucial to break down the arbitrary distinction between primary and secondary use of EHD. Thus, utilizing daily routine EHR data and national registers for the purposes of clinical epidemiology, precision medicine and to enrich omic analyses with clinical data within the approved protocol.

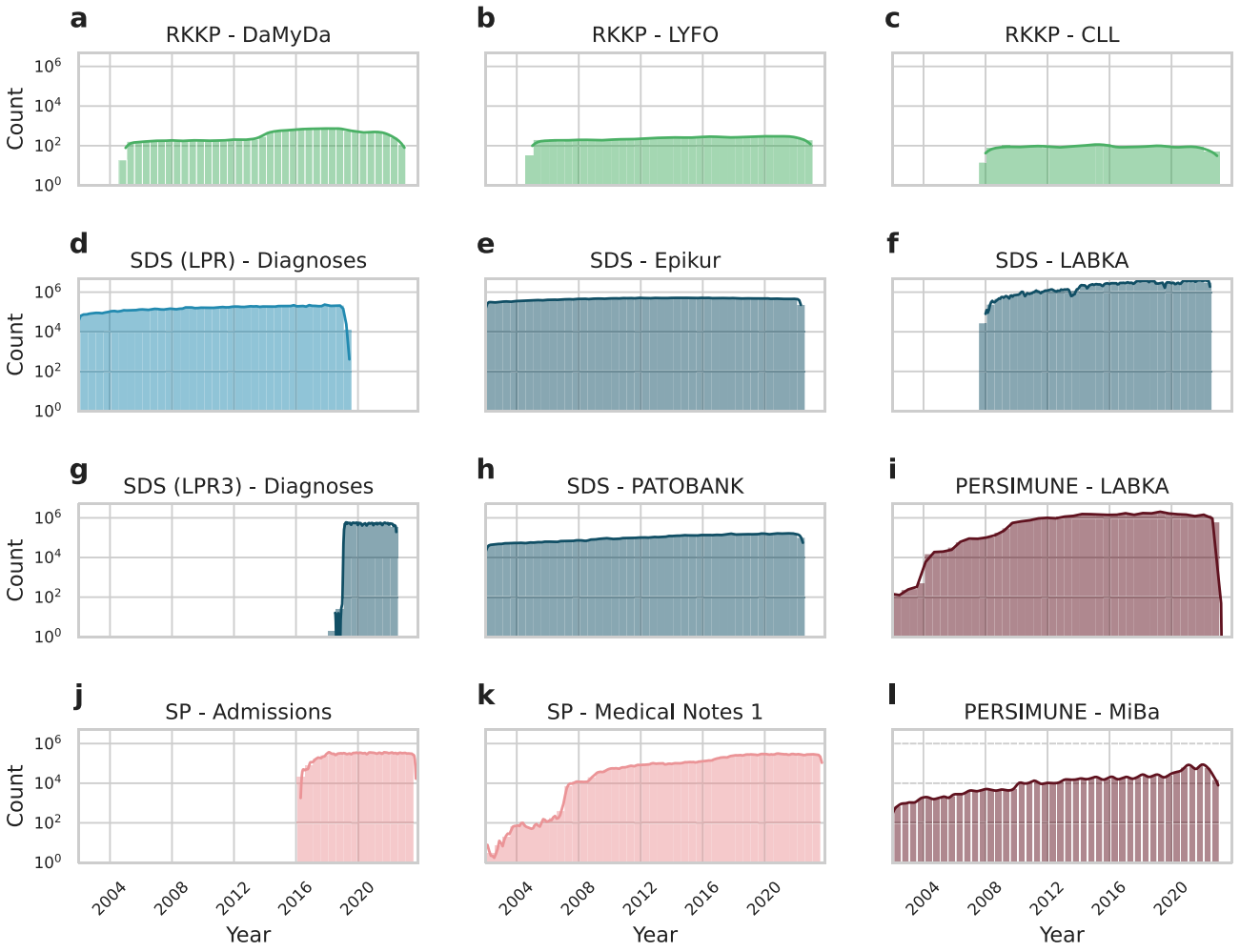

Figure 3 Overview of temporal coverage in 12 key datasets in the DALY-CARE data resource. RKKP hematological quality registers are wide format datasets, where data are typically entered upon diagnosis/registration, upon treatment, relapse, and follow-up (**a**–**c**). SDS LPR register was replaced by LPR3 in Feb 2019 (**d** and **g**). Epikur (prescriptions), LABKA (blood tests), and PATOBANK (pathology requisitions) were independent of this update (**e, f** and **h**). We underscore that similar data from different data sources may have different time coverage (**f** and **i**). The electronic health record (EHR) system of eastern Denmark (SP) went live in Mar 2016 (**j**). Even so, medical notes antedating go-live dates from the previous EHR system were imported and are available as historic notes (**k**). A steadily increasing number of antimicrobial analyses was observed (**l**). Each panel shows the time coverage of a single dataset (**a**–**l**). Note that the y-axis shows the number of observations in three-month intervals on a log-scale. Lines above bins represent kernel density estimates of the counts. All 45 main datasets are detailed in Supplementary Figure S4.

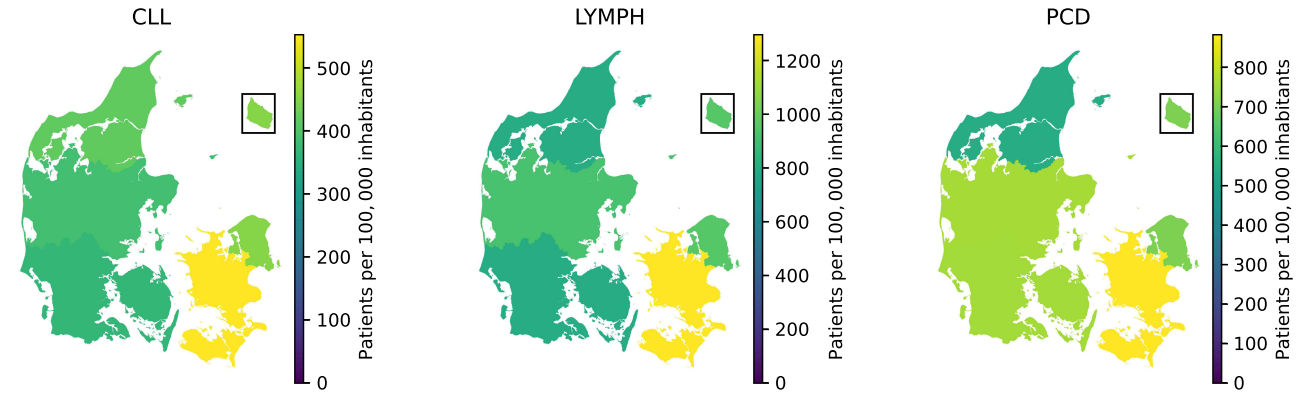

Figure 4 Number of patients in each disease category per 100,000 residents for patients with chronic lymphocytic leukemia (CLL), lymphoma (LYMPH), and plasma cell dyscrasia (PCD). Numbers included all patients in the DALY-CARE data resource regardless of vital status. Please see definition of disease category in the Supplementary Material.

https://doi.org/10.2147/CLEP.S479672

Methodological Considerations

The DALY-CARE cohort is truly population-based as nearly all Danish adult LC patients are referred to one of eight hematological centers and cancer registration is mandatory by law. This results in a register coverage of 99%.[38] Even so, we found large regional disparities in the available data for patients, mainly owing to EHR data being available for only two out of five Danish Regions, covering half of the Danish population. The use of ICD10 codes upon admissions may also differ widely across institutions. This skew in availability and usage could limit the scope of the possible studies or produce unintentional biases in data driven algorithms if they are not accounted for. Like other EHD archives, data in DALY-CARE are affected by shifts in systems such as the transition from LPR to LPR3 (ie 2 Feb 2019) or implementation of EPIC® in eastern Denmark (from May 2016 through Nov 2017; Supplementary Material). These time dependent changes in data and data formats are important to recognize to avoid biases that could lead data driven algorithms astray. To this end, we focused on 1) describing when and where the data are available, 2) standardization and quality control for measurements (ie making the same measures using different units commensurable), and 3) using robust clinical definitions to calculate features based on domain-knowledge such as prognostic indices.

By highlighting previously published work (Supplementary Table S10), we believe that the DALY-CARE data resource will serve as an important first step towards standardizing and creating similar data resources for other research groups. This would help alleviating the problem of limited training data by enabling the use of techniques that can leverage information learned from training on other cohorts, such as meta-learning,[43] and have so far helped to pool larger international cohorts for rare diseases and events.[44–49] Beyond the national generalization (eg SKS coding), we mapped 90% of our curated data from medicine, laboratory measurements and diagnoses to OMOP standardization to facilitate international collaborations. This will also serve as a critically important step for implementing state-of-the-art mAI that aggregate information across countries using federated learning.

With access to free-text medical notes for 22,708 patients within the DALY-CARE data resource, we have initiated work to test natural language processing for identification of specific events such as infections. Ongoing research in this field continues to show promise.[50,51] From our preliminary results, it is, however, unclear whether unstructured text data will be generalizable across different hospitals and diseases.[52]

We recognize the lack of large, shared datasets as the biggest obstacle for development in mAI research and call for the integration and use of primary health care data in mAI research on a coordinated national level. By creating a secure research environment within the DALY-CARE data resource, we here deliver proof-of-concept to overcome this barrier, thus paving the way for expanding into different domains (outside hematology) for other institutions.

Medical interventions for LC patients are often implemented without scientific evidence, and many interventions will never be investigated in randomized clinical trials (RCT).[53] DALY-CARE provides means to qualify and inform the impact of changes in health care practice that have not, cannot, and likely will never be tested sufficiently in RCTs. This is exemplified by monitoring overall survival, health care utilizations and infections for patients ending specialized follow-up or receiving immunoglobulin replacement therapy.

Conclusion

The DALY-CARE data resource includes multimodal health data for more than 65,000 patients with LC, gathered and stored on a super-computer cloud ensuring data privacy, storing capacity, and the processing power needed without compromising the risk of data migration.[54] Based on published studies using the DALY-CARE data resource, we provide examples of how DALY-CARE can lead to (1) finding novel prognostic markers using biobank data, (2) using real-world evidence studies to evaluate the efficacy of routine health care, and (3) deploying mAI algorithms directly into EHR systems. Thus, DALY-CARE allows for development of near real-time decision-support tools and extrapolation of clinical trial results into clinical practice, thereby improving care for patients with LC and paving the way for truly data-driven hematology, while facilitating streamlining of health data infrastructure across cohorts and medical domains.

Acknowledgments

We would like to thank Dr Peter de Nully Brown for founding the Danish National Lymphoma Registry and leading the way for Danish quality registers in hematology. We sincerely thank Alexander Djupnes Fuglkjær from Aalborg University for contributing to the code base and external validation hereof. We thank Anders Christian Riis-Jensen and Anton Kokholm Andersen from CØK within the Capital Region of Denmark for retrieving and providing EHR data. We further thank Prof Jens Lundgren leading PERSIMUNE for providing laboratory and microbiology data, Prof Sisse Rye Ostrowski for collaboration with the Copenhagen Hospital Biobank, Dr Ida Schjødt for kindly providing flow cytometry data from the Surface Marker Laboratory at Rigshospitalet, Dr Mette Klarskov Andersen for kindly providing cytogenetics data from the Department of Clinical Genetics, Rigshospitalet, and Lone Bredo Pedersen for performing and providing IGHV analyses from the CLL laboratory, Rigshospitalet. This paper has been uploaded to MedRxiv as a preprint: https://www.medrxiv.org/content/10.1101/2024.04.11.24305663v1.

Author Contributions

CUN conceived the project. CUN, CB, and CMF collected the data. CMF created the database. CB, CMF, MW, and TL created the database infrastructure. MW and TL performed data quality control. CB and MW wrote the draft manuscript. All authors contributed to and approved the final manuscript. All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

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

Disclosure

CB received travel grants from Octapharma outside this study. CMF received funding from Octapharma. TL received travel grants from AbbVie outside this study. CF reports grants for research funding from Octapharma, during the conduct of the study. MP reports grants from AstraZeneca and CLL-CLUE under ERA PerMed, outside the submitted work. RST reports grants from AstraZeneca, outside the submitted work. NV received consultancy fees and funding from AstraZeneca outside of this work. CCB reports personal fees from Janssen, Octapharma, Astra Zeneca; support for attending meeting from AbbVie, outside the submitted work. ECR received consultancy fees and/or travel grants from AbbVie, Janssen, and AstraZeneca outside of this work. CUN received research funding and/or consultancy fees from AstraZeneca, Janssen, AbbVie, Beigene, Genmab, CSL Behring, Octapharma, Takeda, Eli Lily, MSD, and Novo Nordisk Foundation. All other authors declare no competing interests to disclose for this work.

References

1. Alaggio R, Amador C, Anagnostopoulos I, et al. The 5th edition of the World Health Organization classification of haematolymphoid tumours: lymphoid neoplasms. *Leukemia*. 2022;36(7):1720–1748. doi:10.1038/s41375-022-01620-2
2. Ferlay J, Colombet M, Soerjomataram I, et al. Cancer statistics for the year 2020: an overview. *Int J Cancer*. 2021;149:778–789. doi:10.1002/ijc.33588

3. Kyle RA, Larson DR, Therneau TM, et al. Long-term follow-up of monoclonal gammopathy of undetermined significance. *N Engl J Med*. 2018;378(3):241–249. doi:10.1056/NEJMoa1709974
4. Niemann CU, Wiestner A. B-cell receptor signaling as a driver of lymphoma development and evolution. *Semin Cancer Biol*. 2013;23(6):410–421. doi:10.1016/j.semcancer.2013.09.001
5. Greer JP, Arber DA, Glader BE, et al. *Wintrobe's Clinical Hematology*. 14th ed. Wolters Kluwer Health Pharma Solutions (Europe) Ltd.; 2018.
6. Agius R, Parviz M, Niemann CU. Artificial intelligence models in chronic lymphocytic leukemia - recommendations toward state-of-the-art. *Leuk Lymphoma*. 2022;63(2):265–278. doi:10.1080/10428194.2021.1973672
7. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211–252. doi:10.1007/s11263-015-0816-y
8. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi:10.1038/sdata.2016.35
9. Arboe B, Josefsson P, Jørgensen J, et al. Danish National Lymphoma Registry. *Clin Epidemiol*. 2016;8:577–581. doi:10.2147/clep.S99470
10. Gimsing P, Holmstrom MO, Klausen TW, et al. The Danish National Multiple Myeloma Registry. *Clin Epidemiol*. 2016;8:583–587. doi:10.2147/clep.s99463
11. da Cunha-Bang C, Geisler CH, Enggaard L, et al. The Danish National Chronic Lymphocytic Leukemia Registry. *Clin Epidemiol*. 2016;8(1):561–565. doi:10.2147/clep.s99486
12. RKKP. Dokumentation af de nationale kliniske kvalitetsdatabaser. Available from: https://www.rkkp-dokumentation.dk/. Accessed October 3, 2023. Danish.
13. Johannesdottir SA, Horváth-Puhó E, Ehrenstein V, Schmidt M, Pedersen L, Sørensen HT. Existing data sources for clinical epidemiology: the Danish National Database of Reimbursed Prescriptions. *Clin Epidemiol*. 2012;4:303–313. doi:10.2147/clep.S37587
14. SDS. Sundhedsdatastyrelsen: the national hospital medication register. Available from: https://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/sygdomme-laegemidler-og-behandlinger/sygehusmedicinregisteret. Accessed September 21, 2023.
15. Erichsen R, Lash TL, Hamilton-Dutoit SJ, Bjerregaard B, Vyberg M, Pedersen L. Existing data sources for clinical epidemiology: the Danish National Pathology Registry and Data Bank. *Clin Epidemiol*. 2010;2:51–56. doi:10.2147/clep.s9908
16. Grann AF, Erichsen R, Nielsen AG, Frøslev T, Thomsen RW. Existing data sources for clinical epidemiology: the clinical laboratory information system (LABKA) research database at Aarhus University, Denmark. *Clin Epidemiol*. 2011;3:133–138. doi:10.2147/clep.S17901
17. Helweg-Larsen K. The Danish Register of Causes of Death. *Scand J Public Health*. 2011;39(7 Suppl):26–29. doi:10.1177/1403494811399958
18. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. *Scand J Public Health*. 2011;39(7 Suppl):30–33. doi:10.1177/1403494811401482
19. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;7:449–490. doi:10.2147/clep.S91125
20. Gjerstorff ML. The Danish Cancer Registry. *Scand J Public Health*. 2011;39(7 Suppl):42–45. doi:10.1177/1403494810393562
21. Voldstedlund M, Haarh M, Mølbak K. The Danish Microbiology Database (MiBa) 2010 to 2013. *Euro Surveillance*. 2014;19(1):20667. doi:10.2807/1560-7917.es2014.19.1.20667
22. Persimune. Persimune: centre of excellence for personoalised medicine of infectious complications in immune deficiency. 2024. Available from: www.persimune.dk/. Accessed January 23, 2025.
23. van Dongen JJ, Lhermitte L, Bottcher S, et al. EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Comparative Study Research Support, Non-U.S. Gov't. *Leukemia*. 2012;26(9):1908–1975. doi:10.1038/leu.2012.120
24. Agathangelidis A, Chatzidimitriou A, Chatzikonstantinou T, et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: the 2022 update of the recommendations by ERIC, the European Research Initiative on CLL. *Leukemia*. 2022;36(8):1961–1968. doi:10.1038/s41375-022-01604-2
25. Brieghel C, da Cunha-Bang C, Yde CW, et al. The number of signaling pathways altered by driver mutations in chronic lymphocytic leukemia impacts disease outcome. *Clin Cancer Res*. 2020;26(6):1507–1515. doi:10.1158/1078-0432.Ccr-18-4158
26. Vainer N, Aarup K, Andersen MA, et al. Real-world outcomes upon second-line treatment in patients with chronic lymphocytic leukaemia. *Br J Haematol*. 2023;201(5):874–886. doi:10.1111/bjh.18715
27. Szabo AG, Klausen TW, Levring MB, et al. The real-world outcomes of multiple myeloma patients treated with daratumumab. *PLoS One*. 2021;16(10):e0258487. doi:10.1371/journal.pone.0258487
28. Aarup K, Rotbain EC, Enggaard L, et al. Real-world outcomes for 205 patients with chronic lymphocytic leukemia treated with ibrutinib. *Eur J Haematol*. 2020;105(5):646–654. doi:10.1111/ejh.13499
29. Ben-Dali Y, Hleuhel MH, da Cunha-Bang C, et al. Richter's transformation in patients with chronic lymphocytic leukaemia: a nationwide epidemiological study. *Leuk Lymphoma*. 2020;61(6):1435–1444. doi:10.1080/10428194.2020.1719092
30. Laugesen K, Mengel-From J, Christensen K, et al. A review of major Danish Biobanks: advantages and possibilities of health research in Denmark. *Clin Epidemiol*. 2023;15:213–239. doi:10.2147/clep.S392416
31. Jensson BO, Arnadottir GA, Katrinardottir H, et al. Actionable genotypes and their association with life Span in Iceland. *N Engl J Med*. 2023;389(19):1741–1752. doi:10.1056/NEJMoa2300792
32. Danish-Health-Data-Authority. Classifications. Available from: https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/classifications. Accessed November 18, 2023.
33. The_Danish_National_Genome_Center. Research projects on the Danish National Genome Center's supercomputer. Available from: https://eng.ngc.dk/research-and-international-collaboration/research-projects-on-the-danish-national-genome-centers-supercomputer. Accessed February 19, 2024.
34. Pedersen CB. The Danish Civil Registration System. *Scand J Public Health*. 2011;39(7 Suppl):22–25. doi:10.1177/1403494810387965
35. Bahlo J, Kutsch N, Bauer K et al. An international prognostic index for patients with chronic lymphocytic leukaemia (CLL-IPI): a meta-analysis of individual patient data. *Lancet Oncol*. 2016;17(6):779–790. doi:10.1016/s1470-2045(16)30029-8
36. Rossum G, Drake JF. Python reference manual. *Centrum voor Wiskunde en Informatica Amsterdam*. 1995.
37. ZfKD. Zentrum für Krebsregisterdaten: conversion table/coding manual. https://www.krebsdaten.de/Krebs/EN/Content/Methods/Coding_manual/coding_manual_node.html. Accessed October 19, 2023.

38. Arboe B, El-Galaly TC, Clausen MR, et al. The Danish National Lymphoma Registry: coverage and data quality. *PLoS One*. 2016;11(6):e0157999. doi:10.1371/journal.pone.0157999
39. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? A systematic review of definitions. *BMC Geriatr*. 2017;17(1):230. doi:10.1186/s12877-017-0621-2
40. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol*. 2011;173(6):676–682. doi:10.1093/aje/kwq433
41. Agius R, Riis-Jensen AC, Wimmer B, et al. Deployment and validation of the CLL treatment infection model adjoined to an EHR system. *NPJ Digit Med*. 2024;7(1):147. doi:10.1038/s41746-024-01132-6
42. Bauchner H, McDermott MM, Butte AJ. Data sharing enters a new era. *Ann Intern Med*. 2023;176(3):400–401. doi:10.7326/m22-3479
43. Zhang XS, Tang F, Dodge HH, Zhou J, Wang F. MetaPred: meta-learning for clinical risk prediction with limited patient electronic health records. *Kdd*. 2019;2019:2487–2495. doi:10.1145/3292500.3330779
44. Agathangelidis A, Chatzidimitriou A, Gemenetzi K, et al. Higher-order connections between stereotyped subsets: implications for improved patient classification in CLL. *Blood*. 2021;137(10):1365–1376. doi:10.1182/blood.2020007039
45. Mansouri L, Thorvaldsdottir B, Sutton LA, et al. Different prognostic impact of recurrent gene mutations in chronic lymphocytic leukemia depending on IGHV gene somatic hypermutation status: a study by ERIC in HARMONY. *Leukemia*. 2023;37(2):339–347. doi:10.1038/s41375-022-01802-y
46. Del Giudice I, Cappelli LV, Delgado J, et al. Spontaneous regression in chronic lymphocytic leukaemia. Clinical features of 50 cases from the ERIC registry and review of the literature. *Br J Haematol*. 2023;201(2):353–356. doi:10.1111/bjh.18696
47. Chatzikonstantinou T, Kapetanakis A, Scarfò L, et al. COVID-19 severity and mortality in patients with CLL: an update of the international ERIC and Campus CLL study. *Leukemia*. 2021;35(12):3444–3454. doi:10.1038/s41375-021-01450-8
48. Baliakas P, Mattsson M, Hadzidimitriou A, et al. No improvement in long-term survival over time for chronic lymphocytic leukemia patients in stereotyped subsets #1 and #2 treated with chemo(immuno)therapy. *Haematologica*. 2018;103(4):e158–e161. doi:10.3324/haematol.2017.182634
49. Sutton LA, Young E, Baliakas P, et al. Different spectra of recurrent gene mutations in subsets of chronic lymphocytic leukemia harboring stereotyped B-cell receptors. *Haematologica*. 2016;101(8):959–967. doi:10.3324/haematol.2016.141812
50. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med*. 2023;29(9):2307–2316. doi:10.1038/s41591-023-02504-3
51. Boonstra MJ, Weissenbacher D, Moore JH, Gonzalez-Hernandez G, Asselbergs FW. Artificial intelligence: revolutionizing cardiology with large language models. *Eur Heart J*. 2024;45(5):332–345. doi:10.1093/eurheartj/ehad838
52. Parviz M, Niemann CU. Investigating bias in Danish Medical Notes: infection classification of long texts using transformer and LSTM Architectures coupled with BERT [in preparation]. 2024.
53. Djulbegovic B. A framework to bridge the gaps between evidence-based medicine, health outcomes, and improvement and implementation science. *J Oncol Pract*. 2014;10(3):200–202. doi:10.1200/jop.2013.001364
54. Zhang J, Morley J, Gallifant J, et al. Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation. *Lancet Digit Health*. 2023;5(10):e737–e748. doi:10.1016/s2589-7500(23)00157-7