RESEARCH ARTICLE

Creation of a pediatric mature B-cell non-Hodgkin lymphoma cohort within the Pediatric Health Information System Database

Rebecca Citrin[1,2]*, Joseph P. Horowitz[3], Anne F. Reilly[1,4], Yimei Li[1,2,4], Yuan-Shung Huang[5], Kelly D. Getz[1,5], Alix E. Seif[1,4], Brian T. Fisher[2,4,5,6], Richard Aplenc[1,2,4,5]

1 Division of Oncology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **2** Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **3** Drexel University College of Medicine, Philadelphia, Pennsylvania, United States of America, **4** Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **5** Center for Pediatric Clinical Effectiveness, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **6** Division of Infectious Diseases, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America

* rebecca.citrin@gmail.com

Abstract

Mature B-cell non-Hodgkin lymphoma (B-NHL) constitutes a collection of relatively rare pediatric malignancies. In order to utilize administrative data to perform large-scale epidemiologic studies within this population, a two-step process was used to assemble a 12-year cohort of B-NHL patients treated between 2004 and 2015 within the Pediatric Health Information System database. Patients were identified by ICD-9 codes, and their chemotherapy data were then manually reviewed against standard B-NHL treatment regimens. A total of 1,409 patients were eligible for cohort inclusion. This process was validated at a single center, utilizing both an institutional tumor registry and medical record review as the gold standards. The validation demonstrated appropriate sensitivity (91.5%) and positive predictive value (95.1%) to allow for the future use of this cohort for epidemiologic and comparative effectiveness research.

Check for updates

OPEN ACCESS

Citation: Citrin R, Horowitz JP, Reilly AF, Li Y, Huang Y-S, Getz KD, et al. (2017) Creation of a pediatric mature B-cell non-Hodgkin lymphoma cohort within the Pediatric Health Information System Database. PLoS ONE 12(10): e0186960. https://doi.org/10.1371/journal.pone.0186960

Editor: Marta M. Alonso, Universidad de Navarra, SPAIN

Received: July 29, 2017

Accepted: October 10, 2017

Published: October 23, 2017

Copyright: © 2017 Citrin et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant analyses performed by our research team using the database are included in the manuscript. Data are available from the Children's Hospital Association (CHA) Pediatric Health Information System (PHIS) database, which is comprised of non-profit, freestanding U.S. children's hospitals (http://www.chca.com). PHIS contains de-identified patient data. However, it is possible to identify institutional-specific data from the hospitals that contribute data to PHIS. Additionally, the agreement between CHA and the investigators does not legally allow for

Introduction

Pediatric mature B cell non-Hodgkin lymphomas (B-NHL) represent 4% of all pediatric malignancies, with an annual incidence of approximately 450 cases in the United States [1]. The relative rarity of these malignancies makes them difficult to study outside of large cooperative group trials. Administrative databases are an alternative data source within which to perform large-scale epidemiologic and comparative effectiveness studies. However, the use of administrative databases to study B-NHL has been hampered by well-described limitations of ICD-9 codes in identifying newly diagnosed pediatric cancer patients [2–4]. To address this

sharing PHIS data with non-participant sites. Therefore, we are able to make the data available to investigators at participating PHIS sites, but legal restrictions prevent us from making the data available to researchers outside of PHIS institutions. Anyone interested in re-analyzing the data may contact the CHA PHIS administrators (analytics@childrenshospitals.org) to be considered for data access. Alternatively, they may contact Rebecca Citrin (rebecca.citrin@gmail.com) who will coordinate with the PHIS administrators to address the request.

Funding: This study was supported by Eagles Fly for Leukemia (www.eaglesfly.org), the NIH Grant R01 CA133881 (R.A.), and by Alex's Lemonade Stand Foundation (www.alexslemonade.org). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

limitation, we developed a multistep process incorporating ICD-9 codes and manual review of chemotherapy billing data to assemble and validate a cohort of pediatric patients with *de novo* B-NHL using the Pediatric Health Information System (PHIS) database.

Materials and methods

Data source

PHIS is a comprehensive administrative database containing inpatient and emergency department data from 48 freestanding children's hospitals throughout the United States. The available data is comprised of patient-level information including demographics, dates of admission, discharge disposition, and all ICD-9 diagnosis and procedure codes linked to a given admission. It also contains billed resource utilization data, which incorporates pharmaceuticals, laboratory tests, imaging studies, procedures, and the dates that these bills were generated. Patients are assigned a unique identifier within this database, allowing for identification across multiple inpatient admissions within the same institution. This study was approved by the Children's Hospital of Philadelphia's (CHOP) Institutional Review Board and was granted a waiver of informed consent.

Cohort assembly

First, all admissions to one of the 48 contributing PHIS sites between January 1, 2004 and September 30, 2015 were screened for the presence of an ICD-9 diagnosis code consistent with possible B-NHL. All relevant ICD-9 codes utilized are listed in Table 1, along with the analogous ICD-10 codes for future database queries. Patients were excluded if they lacked billing data for all thirteen pre-specified chemotherapeutic agents used in standard American B-NHL treatment regimens.

Next, chemotherapy data were manually reviewed for at least 45 days from the first hospital admission associated with the malignant diagnosis code. Patients were included in the final

Table 1. List of ICD-9 and ICD-10 diagnosis codes for B-NHL.

ICD-9 code	Description
200.0x	Reticulosarcoma
200.2x	Burkitt's leukemia or lymphoma
200.5x	Primary CNS lymphoma
200.7x	Large cell lymphoma
200.8x	Other named variants of reticulosarcoma and lymphosarcoma
202.8x	Other malignant lymphomas
202.9x	Other and unspecified malignant neoplasms of lymphoid and histiocytic tissue
ICD-10 code	**Description**
C83.3x	Diffuse large B cell lymphoma
C83.7x	Burkitt's lymphoma
C83.8x	Other non-follicular lymphoma
C83.9x	Non-follicular (diffuse) lymphoma
C85.1x	Unspecified B cell lymphoma
C85.2x	Mediastinal (thymic) B cell lymphoma
C85.8x	Other specified types of non-Hodgkin lymphoma
C85.9x	Non-Hodgkin lymphoma, unspecified
C96.9	Malignant neoplasm of lymphoid, hematopoietic, and related tissue, unspecified
C96.Z	Other specified malignant neoplasms of lymphoid, hematopoietic, and related tissue

https://doi.org/10.1371/journal.pone.0186960.t001 cohort if they had at least two courses of chemotherapy consistent with established B-NHL regimens. Specifically, the majority of these regimens included one COPADM (cyclophosphamide, vincristine, doxorubicin, methotrexate, and either prednisone, prednisolone, or methylprednisolone) or RCOPADM (COPADM + rituximab), with or without a preceding COP cytoreduction cycle (cyclophosphamide, vincristine, and either prednisone, prednisolone, or methylprednisolone). Additionally, patients who received two COPAD courses (COPADM without methotrexate), at least two R-EPOCH cycles with or without a preceding COP reduction cycle (rituximab, etoposide, vincristine, cyclophosphamide, doxorubicin, and either prednisone, prednisolone, or methylprednisolone), or consecutive regimens per the non-Hodgkin Lymphoma Berlin-Frankfurt-Munster (BFM) 95 study (course 1: methotrexate, cytarabine, etoposide, ifosfamide, dexamethasone, +/- vincristine; course 2: cyclophosphamide, doxorubicin, methotrexate, dexamethasone, +/- vincristine) were eligible for cohort inclusion. These regimens were selected based upon their use in the following cooperative group trials during or preceding the study period: Children's Cancer Group 5961, Children's Oncology Group ANHL01P1, Children's Oncology Group ANHL1131, and Berlin-Frankfurt-Munster NHL 95 trial [5–8]. This cohort generation process allowed for subtle deviations in chemotherapy patterns, including any number of additional days of steroid exposure, adjustments in intra-cycle chemotherapy schedules of up to two days, and the absence of up to one chemotherapy class per cycle.

Additionally, patients were eligible for cohort inclusion if they received at least one COP reduction cycle (without the requirement to receive at least two cycles) with no subsequent inpatient encounters in at least the following 45 days. This condition was stipulated in order to capture patients who may not live long enough to meet the full inclusion criteria detailed above. Lastly, patients who received cycles of B-NHL directed therapies that typically appear later in cooperative group protocols underwent additional review, with entries into the PHIS database reviewed for 60 days prior to the initial admission identified by ICD-9 codes. These patients were included in the cohort if the subsequent review identified earlier chemotherapy regimens meeting inclusion criteria. Patients who did not receive chemotherapy as detailed above were excluded from the primary cohort.

Two authors (R.C. and J.P.H) conducted the primary chemotherapy review. To evaluate inter-rater reliability, 10% of the sample was re-reviewed by both evaluators and a κ-statistic was calculated as a measure of inter-rater agreement with regard to cohort inclusion status.

Cohort validation

Cohort validation utilized the medical records of patients with *de novo* B-NHL treated at CHOP as well as an institutional tumor registry maintained by CHOP. In order to calculate the positive predictive value (PPV) of inclusion in the cohort, members of the study team reviewed pathology reports, attending oncologists' admission notes and discharge summaries, and medication administration records of all CHOP treated patients identified in the initial PHIS database ICD-9 code query. The sensitivity was calculated by comparing the patients identified in PHIS to an institutional CHOP registry of patients with B-NHL.

Finally, in order to ensure that this was a representative cohort, demographic characteristics of patients in the final cohort were extracted from PHIS (including age, gender, and race), and were compared with those reported by the National Cancer Institutes' SEER Program from 2004–2013 using Pearson's χ^2 test [9]. All analyses were performed using Microsoft Excel, SAS version 9.3 (Cary, NC), and STATA statistical software version 14.0 (College Station, TX)

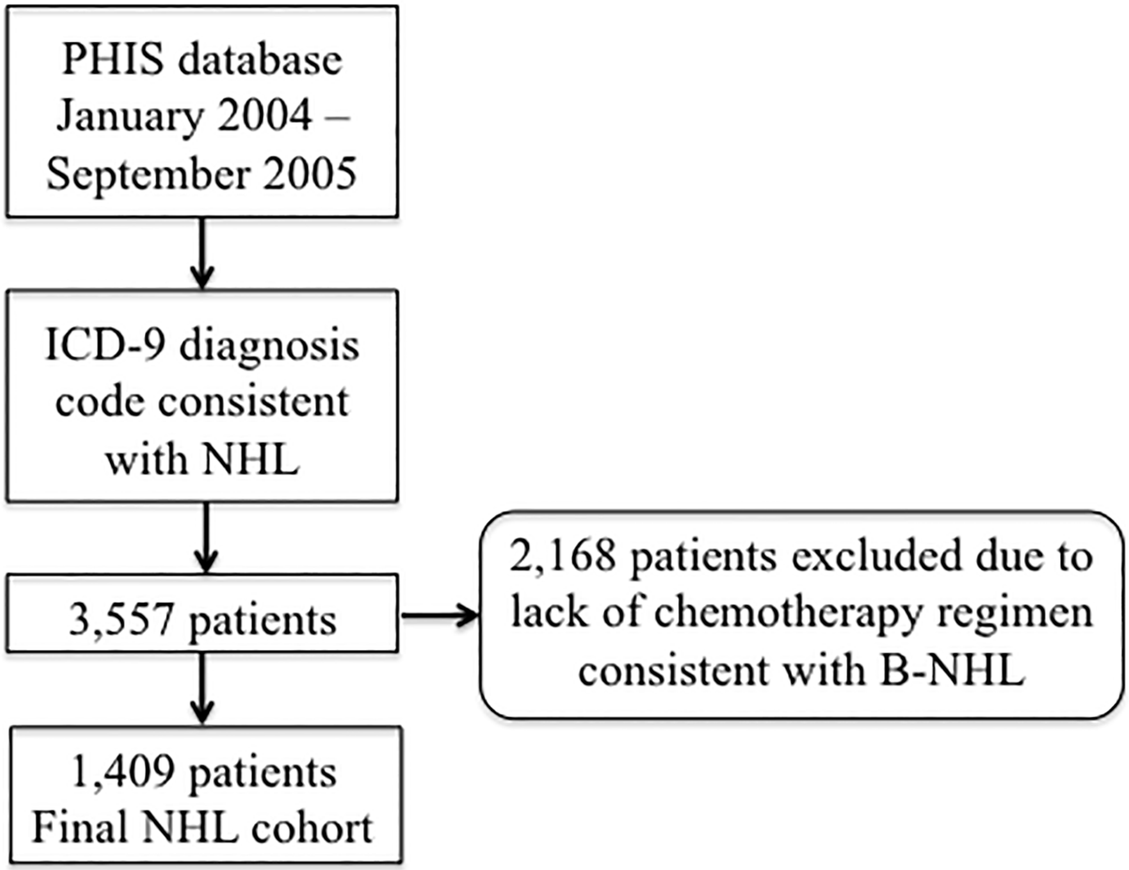
PHIS database January 2004 – September 2005
ICD-9 diagnosis code consistent with NHL
3,557 patients
2,168 patients excluded due to lack of chemotherapy regimen consistent with B-NHL
1,409 patients Final NHL cohort

Fig 1. Flow chart depicting the cohort assembly process.

https://doi.org/10.1371/journal.pone.0186960.g001

Results

A total of 3,557 unique patients with an ICD-9 code for B-NHL were identified in PHIS within the study period. All patients underwent manual chemotherapy review, with 1,409 patients remaining eligible for cohort inclusion. This process is depicted in Fig 1 and Table 2. The 10% of the cohort reviewed by both reviewers had inter-rater concordance of 99.4%.

This two-stage process of cohort generation was validated within the 178 (5%) patients treated at CHOP. As shown in Table 2, using ICD-9 codes alone resulted in a sensitivity of 95.4% with a PPV of 59.5%. Manual chemotherapy review increased the PPV of the primary cohort to 95.1%, with a corresponding decrease in sensitivity to 91.5%. Prior to manual

Table 2. Validation results from each step of the generation process of a pediatric mature B cell non-Hodgkin lymphoma cohort.

	True positives	False positives	False negatives	Sensitivity (95% CI)	Positive predictive value (95% CI)
Step 1: Identification via ICD-9 codes (200.0, 200.2, 200.5, 200.7, 200.8, 202.8, 202.9)	103	70	5	95.4% (89.5–98.5%)	59.5% (51.8–66.9%)
Step 2: Manual chemotherapy review*	97	5	9	91.5% (84.5–96.0%)	95.1% (88.9% - 98.4%)

*The validation cohort for sensitivity and PPV calculations included 178 patients aged 0–21 treated at CHOP during the study period.

https://doi.org/10.1371/journal.pone.0186960.t002

Table 3. Demographic characteristics of patients in PHIS-assembled cohort compared to the National Cancer Institute's SEER program.

	PHIS-assembled (n = 1,395*)		SEER 2004–2013 (n = 1,477)		*P*-value
	Number	%	Number	%	
Age category					
<1	2	0.14	4	0.27	<0.001
≥1 and <5	182	13.1	170	11.51	
≥5 and <10	374	26.8	301	20.38	
≥10 and <15	434	31.1	402	27.22	
≥15 and <20	403	28.9	600	40.62	
Gender					
Male	1,058	75.8	1007	68.18	<0.001
Female	337	24.2	470	31.82	
Race					
White	951	68.2	1131	76.57	0.03
Black	131	9.4	203	13.74	
Other**	294	21.1	130	8.8	
Unknown**	19	1.4	13	0.9	

*14 subjects aged 20–21 in the PHIS-assembled cohort are excluded from all comparisons, as the NCI SEER program does not include this specific age range.

**Other/unknown categories were excluded from this comparison

https://doi.org/10.1371/journal.pone.0186960.t003

chemotherapy review, the majority (85.7%) of the false positive cases were patients with alternate diagnoses, largely T-cell lymphomas and Hodgkin lymphoma. Smaller subsets consisted of patients with relapsed disease (1.4%), patients who began their treatment at a non-PHIS institution and were transferred for second opinions or continuation of therapy (10.0%), and patients who had no evidence of malignancy (2.9%). Following manual chemotherapy review, the five false positive cases included one patient with relapsed disease and four patients with alternate diagnoses treated on B-NHL regimens. The nine false negatives were comprised of five patients who were not identified in the PHIS database ICD-9 code query and four patients who received altered chemotherapy regimens due to comorbidities. There were statistically significant differences in demographic characteristics of patients identified in PHIS as compared to those from the National Cancer Institute's SEER Registry for years 2004–2013 (Table 3).

Discussion

This study outlines the derivation and validation process of assembling a cohort of children with B-NHL within a large administrative database. Our group has previously reported that using ICD-9 diagnosis codes alone confers a poor positive predictive value for accurate patient identification in pediatric malignancies [2–4]. Our low initial PPV based on ICD-9 diagnosis codes alone further underscores the importance of combining chemotherapy review with ICD-9 diagnosis codes in order to accurately identify patients with B-NHL.

Aside from the many potential advantages of this cohort, there are some important limitations. First, demographic parameters within our cohort demonstrate statistically significant differences from those reported by the SEER program across all categories, thus threatening the generalizability of this cohort. One potential explanation for this discrepancy is that older teenagers and young adults may be treated at either pediatric or adult institutions, only the former of which is captured by PHIS. This hypothesis is supported by secondary comparisons of our cohort to SEER data, with the age range restricted to patients <15 years old. The supplemental analysis negated this difference for both age and race categories, with a persistent statistically significant difference noted in the proportions of male and female patients (77.5% male patients in our cohort compared to 71.7% in SEER data; p = 0.004). This finding suggests that the generalizability of our cohort is likely to be preserved for patients treated at children's hospitals, but may not extend to those treated at adult institutions.

Second, we *a priori* excluded certain mature B cell non-Hodgkin lymphomas, including follicular lymphoma, mantle zone lymphoma, and other similar malignancies. These malignancies were excluded due to their rarity among pediatric populations as well as their treatment heterogeneity. While this exclusion also has the potential to limit the generalizability of our cohort, we expect this effect to be minimal due to the infrequency of these conditions in children.

Third, the nature of our two-step patient identification process creates the potential to erroneously exclude patients from the cohort (false negatives). Specifically, this process relies on accurate assignment of ICD-9 diagnosis codes, as well as comprehensive chemotherapy billing and reporting. It is possible that institutions have different processes for assignment of ICD-9 diagnosis codes as well as different treatment and pharmacy billing practices, which could result in misclassification. However, our validation cohort showed a robust sensitivity, suggesting that the risk of falsely excluding patients was likely small. Furthermore, this approach yielded a high PPV, indicating that the patients included in the final cohort truly had B-NHL. Notably, the validation was only conducted at a single center and thus there is the potential for altered sensitivity and PPV calculations if the validation were conducted at multiple sites. However, the demographic similarities of our cohort to SEER data suggest that the risk of false negatives is likely small.

Lastly, the use of ICD-9 codes for initial patient identification may make this process difficult to repeat in the era of ICD-10 coding. As detailed in Table 1, all ICD-9 codes used in this process have comparable ICD-10 codes, most with higher level of detail. Given these similarities, we anticipate that utilization of ICD-10 codes will generate a similarly representative patient cohort.

Despite these limitations, this cohort represents a unique opportunity to perform large-scale comparative effectiveness and epidemiologic research studies on an uncommon collection of pediatric malignancies.

Author Contributions

Conceptualization: Rebecca Citrin, Kelly D. Getz, Richard Aplenc.

Data curation: Yimei Li, Yuan-Shung Huang.

Formal analysis: Rebecca Citrin, Yimei Li.

Investigation: Joseph P. Horowitz.

Methodology: Rebecca Citrin, Anne F. Reilly, Yimei Li, Kelly D. Getz, Alix E. Seif, Brian T. Fisher, Richard Aplenc.

Resources: Richard Aplenc.

Supervision: Anne F. Reilly, Yimei Li, Kelly D. Getz, Richard Aplenc.

Writing – original draft: Rebecca Citrin.

Writing – review & editing: Kelly D. Getz, Alix E. Seif, Brian T. Fisher, Richard Aplenc.

References

1. PDQ® Pediatric Treatment Editorial Board. PDQ Childhood Non-Hodgkin Lymphoma Treatment. Bethesda, MD: National Cancer Institute. http://www.cancer.gov/types/lymphoma/hp/child-nhl-treatment-pdq Updated 11/30/2016. Accessed 12/22/2016.
2. Desai AV, Kavcic M, Huang YS, Herbst N, Fisher BT, Seif AE, et al. Establishing a high-risk neuroblastoma cohort using the Pediatric Health Information System Database. Pediatr Blood Cancer. 2014; 61 (6):1129–31. https://doi.org/10.1002/pbc.24930 PMID: 24616331
3. Fisher BT, Harris T, Torp K, Seif AE, Shah A, Huang YV, et al. Establishment of an 11-year cohort of 8733 pediatric patients hospitalized at United States free-standing Children's Hospitals with de novo acute lymphoblastic leukemia from heath care administrative data. Medical Care. 2014; 52(1):e1–e6. https://doi.org/10.1097/MLR.0b013e31824deff9 PMID: 22410405
4. Kavcic M, Fisher BT, Torp K, Li Y, Huang YS, Seif AE, et al. Assembly of a cohort of children treated for acute myeloid leukemia at free-standing children's hospitals in the United States using an administrative database. Pediatr Blood Cancer. 2013; 60(3):508–11. https://doi.org/10.1002/pbc.24402 PMID: 23192853
5. Children's Oncology Group, National Cancer Institute. Combination chemotherapy with or without rituximab in treating younder patients with stage III-IV non-Hodgkin lymphoma or B-cell acute leukemia. In: ClinicalTrials.gov [Internet]. Available from: Bethesda (MD): National Library of Medicine (US). 2000. https://clinicaltrials.gov/ct2/show/NCT01595048 Accessed 4/15/16.
6. Children's Oncology Group, National Cancer Institute. Rituximab, rasburicase, and combination chemotherapy in treating young patients with newly diagnosed advanced B-cell leukemia or lymphoma. In: ClinicalTrials.gov [Internet]: Bethesda (MD): National Library of Medicine (US). 2000- [cited 2016 April 15]. Available from: Bethesda (MD): National Library of Medicine (US). 2000. https://clinicaltrials.gov/ct2/show/NCT00057811 Accessed 4/15/16.
7. Gustave Roussy. Treatment of mature B-cell lymphoma/leukemia. In: ClinicalTrials.gov [Internet]. Available from: Bethesda (MD): National Library of Medicine (US). 2000. https://clinicaltrials.gov/ct2/show/NCT00162656 Accessed 4/15/16.
8. Woessmann W. The impact of the methotrexate administration schedule and dose in the treatment of children and adolescents with B-cell neoplasms: a report of the BFM Group Study NHL-BFM95. Blood. 2004; 105(3):948–58. https://doi.org/10.1182/blood-2004-03-0973 PMID: 15486066
9. Surveillance Research Program, National Cancer Institute. SEER*Stat software (seer.cancer.gov.seerstat) version.