Observational Study

Medicine®

OPEN

A novel approach for medical research on lymphomas

A study validation of claims-based algorithms to identify incident cases

Cécile Conte, MSci[a,b], Aurore Palmaro, MSc, PhD[a,b,c], Pascale Grosclaude, MD, PhD[a,d], Laetitia Daubisse-Marliac, MD, PhD[a,d], Fabien Despas, PharmD, PhD[a,b,c,*], Maryse Lapeyre-Mestre, MD, PhD[a,b,c]

Abstract

The use of claims database to study lymphomas in real-life conditions is a crucial issue in the future. In this way, it is essential to develop validated algorithms for the identification of lymphomas in these databases. The aim of this study was to assess the validity of diagnosis codes in the French health insurance database to identify incident cases of lymphomas according to results of a regional cancer registry, as the gold standard.

Between 2010 and 2013, incident lymphomas were identified in hospital data through 2 algorithms of selection. The results of the identification process and characteristics of incident lymphomas cases were compared with data from the Tarn Cancer Registry. Each algorithm's performance was assessed by estimating sensitivity, predictive positive value, specificity (SPE), and negative predictive value.

During the period, the registry recorded 476 incident cases of lymphomas, of which 52 were Hodgkin lymphomas and 424 non-Hodgkin lymphomas. For corresponding area and period, algorithm 1 provides a number of incident cases close to the Registry, whereas algorithm 2 overestimated the number of incident cases by approximately 30%. Both algorithms were highly specific (SPE = 99.9%) but moderately sensitive. The comparative analysis illustrates that similar distribution and characteristics are observed in both sources.

Given these findings, the use of claims database can be consider as a pertinent and powerful tool to conduct medico-economic or pharmacoepidemiological studies in lymphomas.

Abbreviations: AD = associated diagnosis, CLL/SLL = chronic lymphocytic leukemia/small lymphocytic lymphoma, DLBCL = diffuse large B cell lymphomas, FP = false positives, HL = Hodgkin lymphomas, ICD-10 = Classification of Diseases, 10th revision, ICD-O-3 = Classification of Diseases for Oncology, 3rd edition, LTD = long-term chronic diseases, MD = main diagnosis, NHL = non-Hodgkin lymphoma, NPV = negative predictive value, PMSI = Programme de Médicalisation des Systèmes d'information, PPV = predictive positive value RD = related diagnosis, SE = sensitivity, SNIIRAM = Système National d'Informations inter-Régimes de l'Assurance Maladie (National inter-scheme information system on health insurance), SPE = specificity, TP = true positives.

Keywords: administrative claims, epidemiological method/data accuracy, health care, international classification of diseases, lymphoma, registries

1. Introduction

Lymphomas are a large and heterogeneous group of lymphoid neoplasms with distinct biological and clinical features, treatment, and prognosis.[1,2] Non-Hodgkin lymphoma (NHL) is the most frequent hematologic malignancy and account for approximately 90% of lymphomas.[3,4] In the last 15 years, incidence of NHL has increased steadily, whereas the progress of pharmacological treatments improves NHL median survival time with a constant decrease of mortality.[5–8] In parallel, there is an increased incidence of Hodgkin lymphomas (HL) in adolescents and young adults with a large number of surviving patients.[5,7,9–11] Consequently, there are increased number of patients exposed to potential cancer-related consequences such as long-term adverse effects of treatment, polypharmacy and drug interactions, risk of 2nd cancer, and relapse. Moreover, oncohematology represents a fast-evolving field with continuous scientific progress, update, and changes especially in genomics and biology, diagnostic improvement, and therapeutics with targeted therapy.[12–20] Therapeutic changes are based on results of randomized controlled

Editor: Weimin Guo.

The work received support from the National Research Agency (Agence Nationale de la Recherche (ANR)) for the "investissement d'avenir" ("Investment in the Future") (ANR-11-PHUC-001: CAPTOR).

The authors have no conflicts of interest to disclose.

[a] LEASP-UMR 1027, Inserm-University of Toulouse, [b] Medical and Clinical Pharmacology Unit, [c] CIC 1436, Toulouse University Hospital, [d] Claudius Regaud Institute, IUCT-O, Tarn Cancer Registry, Toulouse, France.

[*] Correspondence: Fabien Despas, Service de Pharmacologie Clinique, Centre Hospitalier Universitaire de Toulouse, 37 Allées Jules Guesde, 31000 Toulouse, France (e-mail: fabien.despas@univ-tlse3.fr).

Medicine (2018) 97:2(e9418)

Received: 21 July 2017 / Received in final form: 30 November 2017 / Accepted: 1 December 2017

http://dx.doi.org/10.1097/MD.0000000000009418

trials, conducted on a limited number of patients with drastic selection criteria. As a consequence, these patients are nonrepresentative of patients in the real clinical practice (i.e., older, with polymorbidity, and polypharmacy) and real-life data remain scarce.[21–24] Moreover, long-term effects of new antineoplastic agents remain unknown after marketing authorization. In this context, real-life data are required to conduct pharmacoepidemiological studies, especially, safety evaluation. Multiple sources provide useful data to conduct observational study on lymphoma, as data collected in cancer registries and retrospective or prospective surveys. However, the French health insurance system database (Système National d'Informations inter-Régimes de l'Assurance Maladie, SNIIRAM) may be used as a pertinent and complementary tool for this research purpose because of several strengths that can minimize classic bias associated with other sources. First, this national database provides extensive data covering a population of more than 65 million inhabitants. The large number of patient recorded in this database permits to increase statistic power of analyses especially for studying rare disease. Moreover, the completeness of the data could minimize selection bias related to the constitution of specialized cancer center's cohorts and attrition bias related to long-term follow-up. Selection bias is an important problem giving results not always transposable to the target population. Then, it provides anonymous and individual data on patient characteristics with demographic data, long-term chronic diseases ('affections de longue durée', LTDs), and vital status. The access to ambulatory healthcare consumption (reimbursed drugs and medical acts) and the linkage with data from the national hospital database ('Programme de Médicalisation des Systèmes d'information', PMSI) gives a complete overview of lymphomas care pathway for several years all over France. The database includes also data regarding some drugs used during hospitalization such as rituximab, a cornerstone of the treatment of several types of lymphomas. Hence, this database provides extensive data on drug exposure minimizing information bias (recall bias, nonresponse bias, or reporting bias) and of great interest to conduct medico-economic study in lymphomas.[25–29] Moreover, it could be a pertinent tool for quality measurement of healthcare use in screening or treatment of lymphomas, as highlighted in other cancer.[30] In the light of the above and to improve validity of studies conducted within this database, it is essential to develop validated methods for accurate identification of specific diseases.[31,32] For lymphoma cases, it is crucial to classify with precision NHL by subtypes because of heterogeneity of diseases, treatments, and prognosis. Some identification algorithms have been validated to detect incident cancer cases but, to the best of our knowledge, there is no validated algorithm to identify incident cases of HL and NHL.[33–40]

The aims of this study were to assess the validity of hospital diagnosis codes in the PMSI database to identify incident cases of lymphomas according to results of a regional cancer registry and to compare baseline characteristics of lymphoma cases between sources.

2. Materials and method

2.1. Study design and data sources

The population source was inhabitants of the Tarn department, an administrative area of 384,474 inhabitants in southwestern France. Two algorithms were defined to detect lymphomas cases using PMSI and/or LTD data available in the SNIIRAM database. Incident lymphoma cases were identified using antecedent of hospitalization for lymphoma recorded with hospital diagnosis. An incident case must have no previous record of lymphoma diagnosis during an observation period of 24 months. The results of this identification process were compared with data from the Tarn Cancer Registry considered as the "gold standard" in this area. Complete data from the registry were available until December 31, 2013, thus, data related to hematologic malignancies were extracted from January 1, 2010 to December 31, 2013. In parallel, PMSI and LTD data were extracted from January 1, 2008 to December 31, 2013 for inhabitants of the Tarn department, allowing the reconstitution of an observation period to identify incident cases.

2.1.1. The tarn cancer registry. It is a population-based cancer registry assessed every 5 years by the "Comité d'évaluation des registres". Quality controls are carried out by the registry using tools provided by the International Agency for Research on Cancer and the data are regularly included in the "Cancer Incidence in 5 Continents" monograph series since 1982. Cancers were defined according to the International Classification of Diseases for Oncology, 3rd edition (ICD-O-3). Nominative data are collected and coded in accordance with international guidelines. Identification of potential incident cancer cases is done using several relevant data sources like oncology regional network, anatomopathology laboratories, office from specialized physicians, and LTD and PMSI data. Every case is validated after crossing these data sources and checking medical records. For all patients, the following data are available: demographic data, cancer diagnosis date, stage of the cancer, cancer topography and morphology, vital status, and so on.[41] Lymphoma cases were identified through 2 selection periods (2010–2013 and 2011–2013) on the basis of the WHO classification[1] to assess the impact of length of observation in algorithms' performance. Selection of incident Multiple myeloma (ICD-O code '9732/3'), plasmacytoma (ICD-O code '9731/3'), and extramedullary plasmacytoma (ICD-O code '9734/3') cases has been previously studied separately.[28] A complete list of codes considered to identify lymphomas cases is given in Table 1.

2.1.2. The PMSI database. In France, public and private hospital payment is based on diagnosis-related group system. For each patient hospital stay, a standard discharge summary (Résumé de Sortie Standardisé) is produced with the aim of providing a precise measure of activity which is then used for reimbursement purpose. In this context, the PMSI database contains demographic data, routinely collected medical data (diagnosis, procedures), and administrative data (date and length of stay, hospital location). Diagnoses are coded according to International Classification of Diseases, 10th revision (ICD-10). They provide the leading cause of hospital admission with main diagnosis (MD). They give accuracy on patient's management with related diagnosis (RD) and on major comorbidities and complications with associated diagnosis (AD).The coding quality of these data are regularly checked by internal controls and external audit.

Diagnoses from 'long-term conditions' scheme. LTDs are defined by severe and/or chronic diseases that require expansive or chronic treatment. There is a list established by decree that include 30 diseases, of which hematologic malignancies. After physician request, there is an exemption of copayment for care in relation with LTD. Diagnoses is coded according to *ICD-10*.

Table 1

Lymphomas diagnoses codes used for patients' selection in the registry (ICD-O-3) and PMSI/LTD data (ICD-10).

	ICD-O-3 code	ICD-10 code*
HL	9650/3; 9651/3; 9652/3; 9653/3; 9654/3; 9655/3; 9659/3; 9661/3; 9662/3; 9663/3; 9664/3; 9665/3; 9667/3	C81
NHL		
B-NHL		
FL	9690/3; 9691/3; 9695/3; 9698/3	C82
DLCBL	9678/3; 9679/3; 9680/3; 9684/3	C83.3
Other mature B-cell NHL	9590/0; 9590/3; 9591/3; 9596/3; 9597/3; 9673/3; 9687/3; 9689/3; 9699/3; 9670/0; 9670/3; 9675/3; 9688/3; 9737/3; 9738/3; 9727/3; 9728/3; 9729/3; 9826/3	C83.0; C83.1; C83.7; C83.8; C83.9; C85; C88.4
Mature T-cell NHL	9700/0; 9700/3; 9701/3; 9702/3; 9705/3; 9708/3; 9709/3; 9712/3; 9714/3; 9716/3; 9717/3; 9718/1; 9718/3; 9719/3; 9768/3; 9726/3; 9718/3; 9827/3; 9831/3; 9832/3; 9834/3	C84;C86
CLL/SLL	9670/3; 9823/3	C91.1
Chemotherapy session for neoplasm	–	Z51.1
Other chemotherapy	–	Z51.2

* Diagnosis codes include: MD, RD, and AD, coded according to ICD10. MD corresponds to the leading cause of hospital admission. RD gives accuracy on patient management and AD includes diseases or conditions coexisting with the MD (other disorder, complications, sequelae, and so on).

B-NHL = B cell non-Hodgkin's lymphoma, ICD-O-3 = International Classification of Diseases for Oncology, AD = associated diagnosis, CLL/SLL = chronic lymphocytic leukemia/small lymphocytic lymphoma, DLBCL = diffuse large B cell lymphomas, FL = follicular lymphomas, HL = Hodgkin lymphomas, MD = main diagnosis, NHL = non-Hodgkin lymphomas, RD = related diagnosis.

2.2. Algorithms of selection of incident lymphomas cases in PMSI and LTD data

For the 2 selection periods (2010–2013 and 2011–2013), inhabitants of the Tarn department with lymphoma were identified in the PMSI database through 2 algorithms:

- Algorithm 1: at least a MD of lymphoma or an MD of chemotherapy in combination with a RD or AD of lymphoma;
- Algorithm 2: at least a MD or RD or AD of lymphoma.

For each algorithm, the impact of LTD data in combination with PMSI data were explored. Then, the use of only LTD data to identify incident lymphoma cases were evaluated through algorithm 3 (at least 1 code of lymphoma in LTD data).A complete list of codes used is given in Table 1.

To be defined as incident, patients must have no record of lymphoma diagnosis code in the 24 months (selection within period 2010–2013) or 36 months (selection within period 2011 and 2013) before the 1st hospitalization date for lymphoma found in our dataset.

2.3. Matching

The linkage between the registry and PMSI and/or LTD was done using a probabilistic matching on the basis of combinations of 5 variables: family name, birth name, 1st name, date of birth, sex, place of birth ("commune", lowest administrative area in France). About 24 possible combinations were tested patients matching for at least 1 combination of these variables were considered as matched.

2.4. Analysis

Descriptive statistics were used to characterize the study population (Table 2). Qualitative variables were expressed in frequencies and percentages. Quantitative variables were

Table 2

Characteristics of lymphomas in the tarn cancer registry between 2010 and 2013, $n=476$.

	All lymphomas, $n=476$	HL, $n=52$	FL, $n=60$	DLBCL, $n=136$	Other mature B-cell NHL, $n=88$	T-NHL, $n=41$	CLL/SLL, $n=99$
Age, median IQR	69[58–81]	55[26–74]	63[53–75]	70[61–81]	71[60–82]	65[57–78]	72[65–83]
Gender, *n*, %							
Male	253 (53.2)	26 (50.0)	32 (53.3)	75 (50.0)	47 (53.4)	24 (58.5)	49 (49.5)
Female	223 (46.8)	26 (50.0)	28 (46.7)	61 (50.0)	41 (46.6)	17 (41.5)	50 (50.5)
Ann Arbor staging system, *n*, %							
I	48 (10.1)	5 (9.6)	10 (16.7)	23 (16.9)	10 (11.4)	0	–
II	51 (10.7)	22 (42.3)	8 (13.3)	15 (11.0)	5 (5.7)	1 (2.4)	–
III	47 (9.9)	11 (21.2)	11 (18.3)	19 (14.0)	4 (4.5)	2 (4.9)	–
IV	142 (29.8)	13 (25.0)	23 (38.4)	57 (41.9)	39 (44.3)	10 (24.4)	–
Missing	188 (39.5)	1 (1.9)	8 (13.3)	22 (16.2)	30 (34.1)	28 (68.3)	–
Binet staging system, *n*, %							
A	–	–	–	–	–	–	49 (49.5)
B	–	–	–	–	–	–	7 (7.1)
C	–	–	–	–	–	–	6 (6.0)
Missing	–	–	–	–	–	–	37 (37.4)

CLL/SLL = chronic lymphocytic leukemia/small lymphocytic lymphoma, DLBCL = diffuse large B cell lymphomas, FL = follicular lymphomas, HL = Hodgkin lymphomas, IQR = interquartile range, NHL = non-Hodgkin lymphomas, T-NHL = T cell non-Hodgkin's lymphoma.

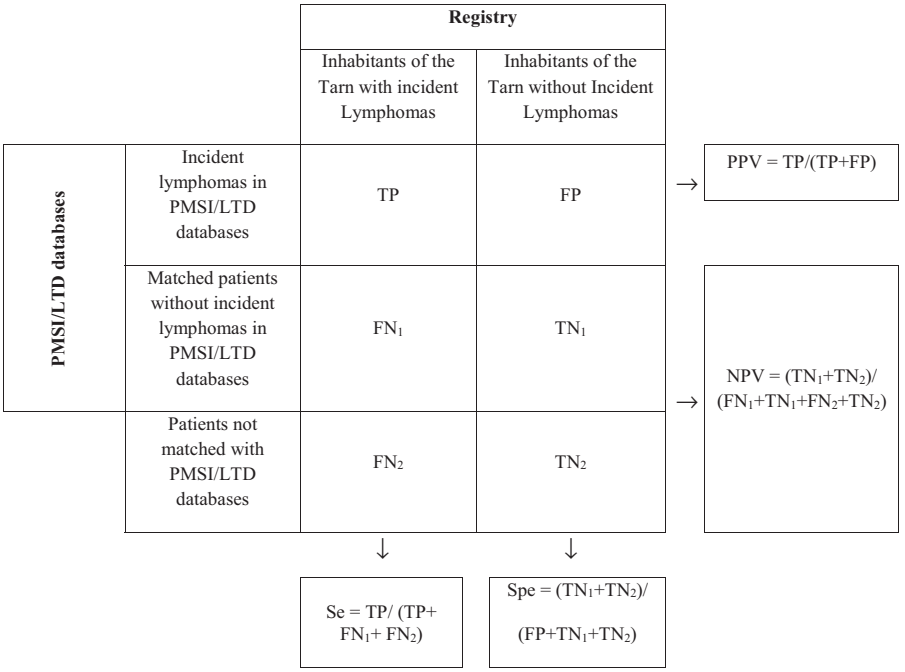

Figure 1. Estimation of algorithms performance's parameters. FN = false negatives, FP = false positives, LTD = long-term chronic diseases, NPV = negative predictive value, PMSI = Programme de Médicalisation des Systèmes d'information, PPV = predictive positive value, Se = sensitivity, Spe = specificity, TN = true negatives, TP = true positives.

expressed as median and interquartile range. The results of the identification process and characteristics of patients in PMSI/LTD databases were compared with true cases from the Tarn Cancer Registry considered as the "gold standard". Thus, each algorithm performance was assessed by estimating sensitivity (SE), predictive positive value (PPV), specificity (SPE), and negative predictive value (NPV).True positives (TP) were incident cases identified in the PMSI/LTD databases recorded in the registry as incident cases of lymphoma. False positives (FPs) were incident cases in PMSI/LTD database not recorded as incident cases in the Registry. False negatives (FNs) were incident cases recorded in the registry but not identified as incident cases in PMSI/LTD databases. Hence, FN can correspond to matched incident lymphoma in the registry not identified by the algorithm applied on the PMSI/LTD data or to incident lymphoma in the registry with no corresponding data in PMSI/LTD databases (not matched patients) PMSI/LTD databases (Fig. 1). The impact of the length of observation period and the use of LTD on algorithm performance was assessed for each algorithm (Table 3). For both algorithms, performance of detection was evaluated for each subtype of lymphoma (list of codes used in Table 1 and results in Table 4). To identify the reasons of discrepancies between the registry and the PMSI database: an exploratory analysis of FN and FP was done. For this purpose, we conducted a multivariate regression logistic to determine characteristics of incident lymphomas in the registry associated with the probability of not being identified in the PMSI database (FN) (Table 5). The FN status was used as the explanatory variable (FN = 1 for FN and FN = 0 for TP). Lymphomas characteristics included in the model were the following: age as a continuous variable, sex, type of lymphoma, and stage according to the Binet staging system or the Ann Arbor staging system. Data analyses were carried out using SAS 9.4 software (SAS Inst., Cary, NC).

2.5. Confidentiality

All data were treated confidentially and were only those already extracted for internal use of the Tarn Cancer Registry. Ethical approval has been given by the French ethical committee and Data Protection Supervisory Authority: ([Commission Nationale de l'Informatique et des Libertés (reference number: 99 80 15 (12/1998), 99 80 15 version 2 (10/2003)]).

3. Results

3.1. Population and algorithms performances for all lymphomas

Between 2010 and 2013, among the 384,474 inhabitants of the Tarn department, the registry identified 476 validated incident cases of lymphomas, of which 52 HL cases and 424 NHL cases. Among the 424 NHL patients, diffuse large B cell lymphomas (DLBCL) was the most common subtype accounting for 32.1% ($n = 136$) of patients, followed by chronic lymphocytic leukemia (CLL)/small lymphocytic lymphoma (SLL) ($N = 99$; 23.3%), other mature B-cell NHL ($n = 88$; 20.7%), follicular lymphoma ($n = 60$; 14.2%), and mature T-cell NHL ($n = 41$; 9.7%). The median age was 69 (58–81) years old with a majority of men ($n = 253$; 53.2%). Characteristics of patients are presented in Table 2.

For corresponding area and period, PMSI data were available for 15,522 patients and LTD data for 7885 patients. Among the 476 lymphomas patients, 203 (42.6%) patients were matched with LTD data and 377 (79.2%) were matched with PMSI data. When using PMSI data only, algorithm 1 provides a number of incident cases close to the Registry (475 vs. 476), whereas algorithm 2 overestimated the number of incident cases by approximately 30%. For algorithm 1, SE and PPV were closed,

Table 3

Se and PPV for both algorithms and selection period (all lymphomas).

Algorithm 1							
PMSI data							
Period	**Incident cases, *n***	**TP, *n***	**FP, *n***	**FN, *n***	**TN, *n***	**Se***	**PPV***
2010–2013	475	318	157	158	377849	66.8 [62.5–70.9]	67.0 [62.6–71.0]
2011–2013	328	222	106	132	378022	62.7 [57.6–67.6]	67.7 [62.4–72.5]
PMSI and LTD data							
Period	**Incident cases, *n***	**TP, *n***	**FP, *n***	**FN, *n***	**TN, *n***	**Se***	**PPV***
2010–2013	487	329	158	147	377848	69.1 [64.8–73.1]	67.5 [63.3–71.6]
2011–2013	340	233	121	107	378007	65.8 [60.7–70.6]	68.5 [63.4–73.2]
Algorithm 2							
PMSI data							
Period	**Incident cases, *n***	**TP, *n***	**FP, *n***	**FN, *n***	**TN, *n***	**Se***	**PPV***
2010–2013	620	361	253	119	377753	75.8 [71.8–79.5]	58.2 [54.3–62.0]
2011–2013	449	270	179	84	377949	76.2 [71.6–80.4]	60.1 [55.5–64.6]
PMSI and LTD data							
Period	**Incident cases,)**	**TP, *n***	**FP, *n***	**FN, *n***	**TN, *n***	**Se***	**PPV***
2010–2013	690	372	318	100	377688	78.1 [74.2–81.6]	54.0 [50.2–57.6]
2011–2013	501	269	232	85	377896	76.3 [71.3–80.1]	53.7 [49.3–58.0]
Algorithm 3							
LTD data							
Period	**Incident cases, *n***	**TP, *n***	**FP, *n***	**FN, *n***	**TN, *n***	**Se***	**PPV***
2010–2013	224	158	66	318	378492	33.2 [29.1–37.5]	70.5 [64.3–76.1]
2011–2013	165	117	41	237	378482	33.0 [28.4–38.1]	70.9 [63.6–77.3]

* 95% confidence interval.

FN = false negatives, FP = false positives, LTD = long-term diseases, PMSI = Programme de Médicalisation des Systèmes d'information, PPV = predictive positive value, Se = sensitivity, TN = true negatives, TP = true positives.

respectively 66.8% (95% confidence interval (CI) [62.5–70.9]) and 67.0% (95% CI [62.6–71]). For algorithm 2, SE was increased by up to 10%, whereas PPV was decreased by up to 9% because of a decrease number of FNs counterbalanced by an increase number of FPs. Both algorithms presented high SPE and NPV (99.9%). The results of SE and PPV calculation for all lymphomas are presented in Table 2. For each algorithm, there was no impact of length of observation in algorithm performance. The use of LTD data alone for identifying lymphomas in claims database resulted in poor performances with SE around 33% and PPV around 70%. The use of LTD data in combination with PMSI data had no impact in algorithm performance to detect incident cases of lymphomas. Characteristics of lymphomas were similar when using the 3 sources. The diagnosis date in the Registry was closed to the 1st hospitalization date identified with a median delay of 0[−1; 21] days.

Table 4

Se and PPV for both algorithms by subtype of lymphomas.

Algorithm 1								
Subtype of lymphoma	**Registry incident cases, *n***	**PMSI incident cases, *n***	**TP, *n***	**FP, *n***	**FN, *n***	**TN, *n***	**Se***	**PPV***
HL	52	66	49	17	3	378413	94.2 [84.4–98.0]	74.2 [62.6–83.2]
NHL								
B-NHL	296	342	221	121	75	378065	74.6 [69.4–79.3]	64.6 [59.4–69.5]
T-NHL	37	23	18	5	19	378440	48.6 [33.4–64.1]	78.3 [58.1–90.4]
LLC/SLL	100	84	19	65	81	378317	19.0 [12.5–27.8]	22.6 [15.0–32.6]
Algorithm 2								
Subtype of lymphoma	**Registry incident cases, *n***	**PMSI incident cases, *n***	**TP, *n***	**FP, *n***	**FN, *n***	**TN, *n***	**Se***	**PPV***
HL	52	75	49	26	3	378404	94.2 [84.4–98.0]	65.3 [54.0–75.1]
NHL								
B-NHL	296	406	240	166	56	378020	81.1 [76.2–85.1]	59.1 [54.3–63.8]
T-NHL	37	34	22	12	15	378433	59.4 [43.5–73.6]	64.7 [47.9–78.5]
LLC/SLL	100	201	47	154	53	378228	47 [37.5–56.7]	23.4 [18.1–29.7]

* 95% confidence interval.

B-NHL = B cell non-Hodgkin's lymphoma, CLL/SLL = chronic lymphocytic leukemia/small lymphocytic lymphoma, HL = Hodgkin lymphomas, NHL = non-Hodgkin lymphomas, PMSI = Programme de Médicalisation des Systèmes d'information, PPV = predictive positive value, Se = sensitivity, T-NHL = T cell non-Hodgkin's lymphoma.

Table 5

Characteristics of incident lymphomas in the registry not identified through the PMSI[a], $n=476$.

	Crude OR, 95% CI	*P*	Adjusted OR, 95% CI	*P*
Age, year	1.02 [1.01–1.03]	0.0005	1.01 [1.00–1.03]	0.0280
Gender				
Women	–	0.4378	–	
Men	0.86 [0.59–1.26]			
Type of lymphomas		0.0002		0.0077
NHL	–			
HL	0.11 [0.03–0.35]		0.18 [0.053–0.64]	
Stage[b]		<.0001		
I or A	–		–	<.0001
II or B	0.20 [0.10–0.43]		0.30 [0.14–0.66]	
III/IV or C	0.14 [0.08–0.25]		0.14 [0.08–0.25]	
Missing	0.88 [0.51–1.52]		0.77 [0.44–1.34]	

[a] False negatives: patients not identified by the algorithm applied on the PMSI data or not found in the PMSI database.
[b] Stage defined according to the Binet staging system or the Ann Arbor staging system; 95% confidence interval.
HL = Hodgkin lymphomas, NHL = non-Hodgkin lymphomas, OR = odds ratio, PMSI = Programme de Médicalisation des Systèmes d'information.

3.2. Algorithms performances by subtypes of lymphomas

Performances of detection of incident cases by both algorithms differ according to lymphomas subtypes for SE and PPV. However, values of SPE and NPV remains maximal (99.9%) for each lymphoma subtype. The results of SE and PPV calculation by lymphomas subtypes are presented in Table 3.

3.2.1. HL. Among the 52 HL identified in the registry between 2010 and 2013, 49 were selected by the 2 algorithms leading to very high SE of 94.2% (95% CI [84.4–98.0]). However, PPV was 10% higher for algorithm 1 than for algorithm 2.

3.2.2. B-NHL. Among the 296 B cell non-Hodgkin's lymphoma (B-NHL) identified in the registry between 2010 and 2013, 221 were selected by algorithm 1 leading to a SE of 74.6% (95% CI [69.4–79.3]) and a PPV of 64.6% (95% CI [59.4–69.5]). For corresponding period, 240 B-NHL were selected by algorithm 2 leading to higher SE (81.1% (95% CI [76.2–85.1])) and a slight decrease in PPV around 5%.

3.2.3. T-NHL. For T cell non-Hodgkin's lymphoma (T-NHL) patients, SE dropped to low values of 48.6% (95% CI [33.4–64.1]) for algorithm 1 and 59.4% (95% CI [43.5–73.6]) for algorithm 2. PPV values were similar (78.3% vs. 64.7%) for each algorithm when considering the width of the CIs.

3.2.4. CLL/SLL. The use of algorithm 2 to identify new CLL patients resulted in better performances with a SE of 47% (95% CI [37.5–56.7]) against a SE of 19.0% (95% CI [12.5–27.8]) for algorithm 1. PPV values were similar for each algorithm.

3.3. Exploratory analysis of FN

Among the 158 FN, 59 were found in the PMSI database, whereas 99 were not found. For matched FN, reasons of misclassification were:

a) Exclusion of patients by algorithm 1: considered as prevalent ($n=2$), patients with only an AD or RD of lymphoma ($n=31$), or missing value for type of diagnosis ($n=12$).
b) Coding error ($n=7$): lymphomas were coded as other hematologic malignancies ($n=5$) such as Waldenström macroglobulinamia, other malignant immunoproliferative diseases, and leukemia or only lymphoma's localization or procedures was coded ($n=2$).
c) No corresponding data in the PMSI database for corresponding period for patients with cutaneous lymphoma, low-grade follicular lymphoma, or CLL Binet stage A ($n=7$).

The results of the univariate and multivariate logistic regression are given in Table 5. After adjustment, characteristics of incident lymphomas associated with an increased probability of being a FN were: older age, type of lymphoma (NHL patients), and localized stage of lymphoma.

3.4. Exploratory analysis of FP

Among the 157 FP, only 10 patients were matched with the registry. These patients were identified in the registry with other hematologic malignancies as follows: chronic myeloid leukemia, lymphoproliferative disorder, refractory anemia with excess blasts, and interdigitating dendritic cell sarcoma. Among the FP with no record in the registry, we identified in PMSI data 45 (30.6%) CLL, 28 (19.0%) DLBCL, 10 (6.8%) HL, 15 (10.2%) follicular lymphoma, 43 (29.2%) other mature B-cell NHL, and 6 (4.1%) mature T-cell NHL.

4. Discussion

4.1. Main findings

The proposed algorithms are extremely specific and consequently diagnosis codes in the PMSI database allow an accurate identification of new lymphomas cases. By contrast, these algorithms are moderately sensitive. Algorithm 1 based on diagnosis and procedure codes seem to be more accurate with optimal performance parameters and incidence close to the registry. The length of the observation period and the combination of LTD with PMSI data do not improve performances. Algorithms exhibited very different performances according to lymphomas subtype, ranging to very poor performance for CLL to very acceptable parameters for HL. The implications of these findings suggest that the use of the PMSI database alone is not enough sensitive to conduct epidemiological studies. Indeed, the incidence provided by PMSI data is close to the registry because FN and FP have similar frequencies and counterbalanced each other.

4.2. Strengths and limitations

Our study presents some limitations. First, this study was conducted in a specific geographic area. Hence, we cannot exclude a lack of representativeness of the algorithms' performance at the national level to detect incident lymphomas cases. Even if coding practice are standardized at the national level and are improving over time, we cannot exclude some discrepancies between hospitals, according to their interpretation of national coding rules. Finally, the performance of algorithm may be underestimated because of a potential failure of linkage between the registry and the PMSI database leading to an increased number of FNs and FPs.

Our study presents several strengths. First, our study provides for the 1st time a validated algorithm to detect incident lymphoma in the French SNIIRAM, but also suitable for other healthcare database using ICD-10th classification medico-administrative database. Some selection algorithms have been validated in cancer but the literature related to hematological diseases is very poor with only 1 systematic review of validated method to identify lymphoma in administrative data. This review identified only 1 publication with a validated algorithm defined with ICD-9 code. The results of this validation study were concordant with our results.[42] Moreover, validation study using ICD-10 are lacking for European and Nordic database, in which ICD-10 is more frequent. Then, our results demonstrate that this approach is of great interest to conduct pharmacoepidemiological or medico-economic studies in lymphomas because of several strengths. First, SPE of each algorithm is maximal allowing an accurate identification of cases. Then, the French health insurance database provides the exhaustiveness of healthcare consumption data at the national level. Finally, our analysis revealed that incident lymphomas not detected as incident or identified in the PMSI database are more likely to be old, with localized stage of lymphoma and concern more NHL patients. According to these findings, FN may concern patients never hospitalized for their lymphoma because of different disease management and/or a gap between diagnosis and treatment. These results suggest that it would have been of interest to conduct analyses of SE including only treated lymphoma patients but this information was lacking in the registry database. However, when regarding algorithm performances by lymphomas subtype, the results directly reflects the heterogeneity of lymphoma care pathway and questioned on the relevance of the use of PMSI data to select new cases in certain lymphomas subtypes. In fact, the very low SE for CLL identification can be explained because a majority of CLL is nonprogressive at diagnosis and does not require active treatment.[43] As a corollary, algorithm 2 results in better performances in CLL because CLL or chemotherapy for CLL is not necessarily the leading cause of hospitalization for these patients. The same reason can be cited for T-NHL. Apart from the majority of FNs corresponded to cutaneous lymphomas which do not require hospitalization and are nondetectable by PMSI data.[44] By contrast, algorithms revealed very high SPE and SE to detect HL patients. These results can be explained because HL always requires inpatient treatment and variability in ICD-10 code is minor.[45] Given the low incidence of this disease and the completeness of SNIIRAM data at the national level, the SNIIRAM database could be used as a relevant and powerful tool to conduct pharmacoepidemiological studies with exhaustive real-life data in HL. Finally, our results illustrate that PMSI data can be used to describe with accuracy lymphomas and that the date of diagnosis can be estimated by the 1st hospitalization for lymphoma found in the dataset. However, the use of ICD-10 to classify NHL by subtypes lacks precision because of the multiplicity of code to register 1 subtype of lymphomas. For that matter, the classification system used impact directly data produced on lymphomas. As depicted by Adzersen et al,[46] the choice of the classification system leads to differences on incidence rate estimates from data coming from a same registry dataset. Differences were stronger for B-NHL. In our study, differences between registry and hospital data may directly result from these discrepancies between ICD-O-3 and ICD-10.

4.3. International initiatives

These considerations and examples highlight that the relevance of the use of claims database for research purpose must be based on a case by case reflection process. In this way, several aspect must be consider to improve validity of the results of future studies conducted on these databases like intrinsic features of diseases and management, type, design, and aims of study conducted. The development of validated tool and the use of standardized method are crucial for the validity of future active surveillance study in lymphomas. In this way, several initiative and project are conducted with the aims to harmonize detection of medical event in claims database in the United States and in Europe (Mini Sentinel program, Observational Medical Outcomes Partnership, Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium).[31,32,47] This validation study follows this quality approach and demonstrates that claims database, and the French SNIIRAM specifically can be a useful and powerful tool for postmarketing studies or medico-economic context for proper research purpose.

References

[1] Campo E, Swerdlow SH, Harris NL, et al. The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. Blood 2011;117:5019–32.
[2] Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood 2016;127:2375–90.
[3] Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. Eur J Cancer 2013;49:1374–403.
[4] Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359–86.
[5] Smith A, Crouch S, Lax S, et al. Lymphoma incidence, survival and prevalence 2004-2014: sub-type analyses from the UK's Haematological Malignancy Research Network. Br J Cancer 2015;112:1575–84.
[6] Coiffier B, Lepage E, Briere J, et al. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. N Engl J Med 2002;346:235–42.
[7] Monnereau A, Troussard X, Belot A, et al. French Network of Cancer Registries (FRANCIM)Unbiased estimates of long-term net survival of hematological malignancy patients detailed by major subtypes in France. Int J Cancer 2013;132:2378–87.
[8] Howlader N, Morton LM, Feuer EJ, et al. Contributions of subtypes of non-Hodgkin lymphoma to mortality trends. Cancer Epidemiol Biomark Prev 2016;25:174–9.
[9] Desandes E, Lacour B, Belot A, et al. Cancer incidence and survival in adolescents and young adults in France, 2000-2008. Pediatr Hematol Oncol 2013;30:291–306.
[10] Dandoit M, Mounier M, Guy J, et al. The heterogeneity of changes in incidence and survival among lymphoid malignancies in a 30-year French population-based registry. Leuk Lymphoma 2015;56:1050–7.
[11] Le Guyader-Peyrou S, Belot A, Maynadié M, et al. Cancer incidence in France over the 1980-2012 period: hematological malignancies. Rev Epidemiol Sante Publique 2016;64:103–12.

[12] Goldman JM, Melo JV. Chronic myeloid leukemia: advances in biology and new approaches to treatment. N Engl J Med 2003;349:1451–64.
[13] Kumar SK, Rajkumar SV, Dispenzieri A, et al. Improved survival in multiple myeloma and the impact of novel therapies. Blood 2008;111: 2516–20.
[14] Palanca-Wessels MC, Press OW. Advances in the treatment of hematologic malignancies using immunoconjugates. Blood 2014;123: 2293–301.
[15] Hamilton A, Gallipoli P, Nicholson E, et al. Targeted therapy in haematological malignancies. J Pathol 2010;220:404–18.
[16] Gifford GK, Gill AJ, Stevenson WS. Molecular subtyping of diffuse large B-cell lymphoma: update on biology, diagnosis and emerging platforms for practising pathologists. Pathology 2016;48:5–16.
[17] Seiler T, Hutter G, Dreyling M. The emerging role of PI3K inhibitors in the treatment of hematological malignancies: preclinical data and clinical progress to date. Drugs 2016;76:639–46.
[18] Piggin A, Bayly E, Tam CS. Novel agents versus chemotherapy as frontline treatment of CLL. Leuk Lymphoma 2017;58:1320–4.
[19] Dunleavy K, Roschewski M, Wilson WH. Precision treatment of distinct molecular subtypes of diffuse large B-cell lymphoma: ascribing treatment based on the molecular phenotype. Clin Cancer Res 2014;20:5182–93.
[20] Puvvada S, Kendrick S, Rimsza L. Molecular classification, pathway addiction, and therapeutic targeting in diffuse large B cell lymphoma. Cancer Genet 2013;206:257–65.
[21] Freemantle N, Marston L, Walters K, et al. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. BMJ 2013;347:f6409.
[22] Al-Refaie WB, Vickers SM, Zhong W, et al. Cancer trials versus the real world in the United States. Ann Surg 2011;254:438–42.
[23] Kwiatkowski K, Coe K, Bailar JC, et al. Inclusion of minorities and women in cancer clinical trials, a decade later: have we improved? Cancer 2013;119:2956–63.
[24] Penberthy LT, Dahman BA, Petkov VI, et al. Effort required in eligibility screening for clinical trials. J Oncol Pract 2012;8:365–70.
[25] Ajrouche A, Estellat C, De Rycke Y, et al. Evaluation of algorithms to identify incident cancer cases by using French health administrative databases. Pharmacoepidemiol Drug Saf 2017;26:935–44.
[26] Moulis G, Germain J, Adoue D, et al. Validation of immune thrombocytopenia diagnosis code in the French hospital electronic database. Eur J Intern Med 2016;32:e21–2.
[27] Moulis G, Palmaro A, Montastruc JL, et al. Epidemiology of incident immune thrombocytopenia: a nationwide population-based study in France. Blood 2014;124:3308–15.
[28] Palmaro A, Gauthier M, Conte C, et al. Identifying multiple myeloma patients using data from the French health insurance databases: validation using a cancer registry. Medicine (Baltimore) 2017;96:e6189.
[29] Palmaro A, Gauthier M, Despas F, et al. Identifying cancer drug regimens in French health insurance database: An application in multiple myeloma patients. Pharmacoepidemiol Drug Saf 2017;26:1492–9.
[30] Fenton JJ, Onega T, Zhu W, et al. Validation of a medicare claims-based algorithm for identifying breast cancers detected at screening mammography. Med Care 2016;54:e15–22.
[31] Avillach P, Coloma PM, Gini R, et al. EU-ADR consortiumHarmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. J Am Med Inform Assoc 2013;20:184–92.
[32] Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiol Drug Saf 2012;21(suppl 1):82–9.
[33] Quantin C, Benzenine E, Hägi M, et al. Estimation of national colorectal-cancer incidence using claims databases. J Cancer Epidemiol 2012; 2012:298369.
[34] Ganry O, Taleb A, Peng J, et al. Evaluation of an algorithm to identify incident breast cancer cases using DRGs data. Eur J Cancer Prev 2003;12:295–9.
[35] Couris CM, Seigneurin A, Bouzbid S, et al. French claims data as a source of information to describe cancer incidence: predictive values of two identification methods of incident prostate cancers. J Med Syst 2006; 30:459–63.
[36] Remontet L, Mitton N, Couris CM, et al. Is it possible to estimate the incidence of breast cancer from medico-administrative databases? Eur J Epidemiol 2008;23:681–8.
[37] Hafdi-Nejjari Z, Couris CM, Schott AM, et al. Role of hospital claims databases from care units for estimating thyroid cancer incidence in the Rhône-Alpes region of France. Rev Dépidémiologie Santé Publique 2006;54:391–8.
[38] Carré N, Uhry Z, Velten M, et al. Predictive value and sensibility of hospital discharge system (PMSI) compared to cancer registries for thyroïd cancer (1999-2000). Rev Dépidémiologie Santé Publique 2006;54:367–76.
[39] Coureau G, Baldi I, Saves M, et al. Performance evaluation of hospital claims database for the identification of incident central nervous system tumors compared with a cancer registry in Gironde, France, 2004. Rev DÉpidémiologie Santé Publique 2012;60:295–304.
[40] Setoguchi S, Solomon DH, Glynn RJ, et al. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data. Cancer Causes Control 2007;18:561–9.
[41] Registre des cancers du Tarn (registre qualifié 2010-2013) /Portail Epidemiologie - France | Health Databases [Internet]. [cited Mar 24, 2017]. Available from: https://epidemiologie-france.aviesan.fr/epidemiologie-france/fiches/tarn-cancer-registry-certified-registry-2010-2013. Accessed June 21, 2017.
[42] Herman RA, Gilchrist B, Link BK, et al. A systematic review of validated methods for identifying lymphoma using administrative data. Pharmacoepidemiol Drug Saf 2012;21(suppl 1):203–12.
[43] Hallek M, Cheson BD, Catovsky D, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. Blood 2008;111:5446–56.
[44] Willemze R, Hodak E, Zinzani PL, et al. Primary cutaneous lymphomas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2013;24(suppl 6):vi149–54.
[45] Eichenauer DA, Engert A, André M, et al. Hodgkin's lymphoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2014;25(suppl 3):iii70–5.
[46] Adzersen K-H, Friedrich S, Becker N. Are epidemiological data on lymphoma incidence comparable? Results from an application of the coding recommendations of WHO, InterLymph, ENCR and SEER to a cancer registry dataset. J Cancer Res Clin Oncol 2016;142:167–75.
[47] Ehrenstein V, Petersen I, Smeeth L, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. Clin Epidemiol 2016;8:49–51.