Drugs - Real World Outcomes (2021) 8:577–587
https://doi.org/10.1007/s40801-021-00269-0

ORIGINAL RESEARCH ARTICLE

Comparison Between Health Insurance Claims and Electronic Health Records (EHRs) for Metastatic Non-small-cell Lung Cancer (NSCLC) Patient Characteristics and Treatment Patterns: A Retrospective Cohort Study

Yookyung Christy Choi[1,2] · **Dongmu Zhang**[1] · **Jerzy E. Tyczynski**[1]

Accepted: 30 June 2021 / Published online: 28 August 2021

Abstract

Background The clinical landscape in non-small-cell lung cancer (NSCLC) treatment has rapidly evolved in recent years. Real-world data (RWD) can provide insights into current clinical practice.

Objective This study examined the patient characteristics and treatment patterns of patients with metastatic NSCLC using RWD sources.

Methods This was a retrospective cohort study using health insurance claims and electronic health records (EHRs). Adult patients treated for metastatic NSCLC during the period 2017 to September 2020 were followed from the earliest treatment date until a censoring event.

Results The claims cohort included 7917 patients with a mean age of 70 years and a mean follow-up period of 373 days. The EHR cohort included 7087 patients with a mean age of 67 years and a mean follow-up period of 362 days. The five most common first-line therapies (LoT1) were the same for both cohorts: carboplatin + paclitaxel, pembrolizumab, carboplatin + pemetrexed + pembrolizumab, cisplatin + pemetrexed, and nivolumab. Mean LoT1 durations were 146 and 147 days in the claims and EHR cohorts, respectively. For patients who received a second LoT (LoT2), the five most common LoT2 were also the same in both cohorts: durvalumab, nivolumab, pembrolizumab, carboplatin + pembrolizumab + pemetrexed, and carboplatin + pemetrexed. Mean LoT2 durations were 157 and 158 days in the claims and EHR cohorts, respectively.

Conclusions LoTs between the claims and EHR cohorts were comparable and showed similar treatment patterns. Traditional platinum-containing chemotherapy was most common in LoT1, whereas programmed cell death protein-1 inhibitors became the most common choices in LoT2. Our findings suggest that RWD can reliably provide up-to-date insight into current treatment modalities and indicate that new clinical evidence is rapidly adopted in patients with NSCLC.

1 Introduction

Lung cancer is the second most common type of cancer for both men and women and has been a leading cause of cancer-related deaths for many years, both globally and in the USA, regardless of sex [1]. In the USA, new lung cancer cases in 2021 were estimated at 235,760, representing

Key Points

Patient characteristics and treatment patterns (regimen choice, progression to subsequent therapy, and therapy duration) were comparable between the health insurance claims cohort and the electronic health record (EHR) cohort of patients with metastatic non-small-cell lung cancer (NSCLC) despite differences in data sources.

The findings suggest that real-world data (claims and EHRs) can reliably provide up-to-date insight into current treatment modalities in patients with NSCLC.

✉ Dongmu Zhang
dongmu.zhang@abbvie.com

[1] Global Epidemiology, AbbVie, Inc., North Chicago, IL, USA

[2] Pharmaceutical Care & Health Systems, College of Pharmacy, University of Minnesota, Minneapolis, MN, USA

roughly 12.4% of all new cancer cases [2]. As the leading cause of cancer-related death in the USA, lung cancer represents 21.7% of all cancer deaths, with an estimated 5-year relative survival probability of 21.7% in the USA based on the 2011–2017 estimates. Over half of patients newly diagnosed with lung cancer present with metastases at the time of initial diagnosis [3, 4]. The 5-year survival probability of patients with metastatic lung cancer is estimated at 6.3% [2].

The prognosis and clinical management of lung cancer differs by histology and tumor characteristics. Broadly, lung cancer can be classified as non-small-cell lung cancer (NSCLC), which accounts for approximately 85% of all lung cancer cases, and small-cell lung cancer (SCLC), which accounts for the remaining 15% [5]. Within NSCLC, the most common histological type is adenocarcinoma, which accounts for roughly half of all NSCLC, followed by squamous cell carcinoma, representing approximately one-quarter of all NSCLC cases [3].

Improved understanding of the molecular pathogenesis of lung cancer has led to the development of specific targeted treatment, which has dramatically improved treatment outcomes for individuals presenting with specific oncogene mutations in their tumor [6–8]. NSCLC is heterogenous in its molecular complexities, and multiple targetable oncogene mutations have been identified, for example, epidermal growth factor receptor mutations present in approximately 15% of NSCLC adenocarcinomas [6]; anaplastic lymphoma kinase (ALK) gene rearrangements present in 3–5% of patients with NSCLC [9]; and programmed death ligand-1 (PD-L1) with tumor proportion score $\geq$ 50% presents in approximately 23–28% of NSCLCs [10, 11].

With new treatment options as a contributor, age-adjusted death rates of patients with lung cancer have been steadily declining in recent years [2]. However, the prognosis for metastatic lung cancer remains poor. It is important to examine how new evidence from clinical trials translates into practice for patients with metastatic NSCLC.

As the NSCLC clinical landscape is changing rapidly, this study examined the utility of two different sets of real-world data (RWD) to provide insights into current clinical practice in the USA. This study used both health insurance claims data and US nationwide electronic health records (EHRs). In general, health insurance claims data are records of utilized healthcare services and products for reimbursement by health insurance; EHRs are patient care records from healthcare providers within a healthcare system. This study aimed to examine patient characteristics and treatment patterns among adult patients newly diagnosed with metastatic NSCLC through two RWD sources: health insurance claims data and US nationwide EHRs.

2 Methods

2.1 Study Design and Data Sources

This was a retrospective observational cohort study using the Optum Clinformatics® Data Mart (CDM) database and Optum EHRs. Optum CDM is a de-identified administrative health claims database from commercial and Medicare Advantage health plans, geographically representing the members from all 50 states in the USA. The insurance claims data include member eligibility and all medical (outpatient, emergency department, inpatient) and pharmacy claims submitted for reimbursement on behalf of health plan members. Optum's longitudinal EHR repository is derived from dozens of healthcare provider organizations in the USA, which includes more than 700 hospitals and 7000 clinics. Optum® EHRs include demographics, medications prescribed and administered, immunizations, allergies, laboratory results (including microbiology), vital signs and other observable measurements, clinical and inpatient stay administrative data, and coded diagnoses and procedures. Optum used a generalized natural language processing (NLP) system to extract and organize concepts from free text into semistructured data fields. Optum's NLP system was developed using vocabularies from the unified medical language system, which includes multiple medical dictionaries. The performance of the NLP system is verified by a team of medical terminologists and clinicians via manual review of sample EHR notes. Both data sources contain de-identified and anonymized patient records. Data management follows statistical de-identification rules and customer data use agreements that adhere to Health Insurance Portability and Accountability Act requirements. This study was exempt from institutional review board approval as it was a secondary database analysis.

2.2 Study Cohorts

NSCLC-specific *International Classification of Diseases, ninth* or *tenth edition* (ICD-9, ICD-10) codes are lacking, so we adapted a treatment-based algorithm published by Duh et al. [12] and validated by Turner et al. [13] to identify patients with NSCLC. All chemotherapeutic agents and regimens defined in this study follow the 2015 American Cancer Society and 2021 National Comprehensive Cancer Network guidelines. Patients with any record of SCLC-specific treatments were excluded. The remaining patients were considered to have NSCLC. In detail, patients who met the following criteria were included in this study: (1) lung cancer diagnosis (ICD-9: 162.xx; ICD-10: C34.xx) between 1

January 2017 and 31 December 2019 and metastasis diagnosis (ICD-9: 196.xx-198.xx; ICD-10: C77.xx-C79.xx) after lung cancer diagnosis; (2) no prior cancer/metastasis diagnosis (except non-melanoma skin cancer) (ICD-9: 140.xx-209.3x, except 173.xx; ICD-10: C00.xx-C96.xx, except C44.xx) before the earliest metastasis diagnosis; (3) with at least one lung cancer treatment after the earliest metastasis diagnosis (the earliest treatment date was defined as the index date) see Table 6 (Appendix) Lung cancer drugs included in the study (adapted from Turner et al. [13], with inclusion of newly approved drugs and relevant HCPCS codes); (4) with continuous health plan enrollment or active EHR records for at least 6 months before and at least 1 month after the systemic treatment initiation (patients who were deceased within 1 month from systemic treatment initiation were kept in the cohorts); (5) with age ≥ 18 years and known sex at the time of systemic treatment initiation; (6) without any record of receiving SCLC-specific treatments (topotecan, cyclophosphamide, doxorubicin, vincristine, temozolomide, ifosfamide, bendamustine at any time, or platinum + topoisomerase inhibitor combination [cisplatin/carboplatin + etoposide/irinotecan] as the first-line regimen). Patients were followed from the earliest systemic treatment date until a censoring event (i.e., end of continuous health plan enrollment or active EHR, death, or end of data availability [30 September 2020]). In the EHRs, NLP tables extracted from the clinical notes were further used to identify patients with metastatic/advanced NSCLC. For metastatic stage, any attribute terms including 'metasta' or 'advanc' were used. SCLC-specific terms were used to exclude patients with SCLC. Figure 1 provides further description of the study design.

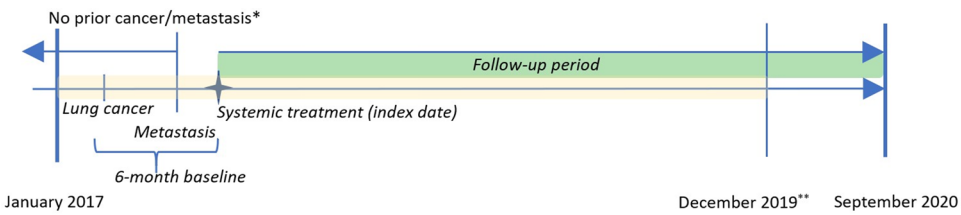
2.3 Study Variables

Demographic characteristics included patient age at the index date (18–54, 55–64, 65–74, and ≥ 75 years), sex (female, male), geographic region (midwest, northeast, south, and west), and Charlson Comorbidity Index (CCI) based on comorbidities [15, 16] during the 6 months before the index date. Because this cohort required a diagnosis of metastatic NSCLC, any malignancy-related CCI scoring was not included. Additional patient characteristics were explored within the EHR cohorts through NLP tables, such as tumor histology types and smoking status if such records were available. Within the claims data, when capturing treatments from prescription claims, prescriptions with American Hospital Formulary Service class other than antineoplastic agents were excluded. With EHRs, when capturing treatments from administered medication records, for medications available with multiple routes of administration, non-chemotherapeutic uses (e.g., non-intravenous bevacizumab) were excluded. If multiple prescribed medication records within the same drug class (e.g., ALK inhibitors) were captured within 28 days, the subsequent prescription records were examined to determine the appropriate regimen.

First line of therapy (LoT1) was defined as all systemic therapy a patient received during the first 28 days from the earliest treatment date. If a patient initiated a different drug or had a > 90-day gap after the LoT1, the initiation date of that new drug or any drug after the gap was defined as the initiation date of the second LoT (LoT2), and the LoT2 regimen was defined as all systemic therapy drugs the patient received within the 28-day window from the newly established initiation date. LoT duration was defined as number of days from the LoT initiation date to the last treatment date within that LoT + 28 days or to a censoring event.

2.4 Statistical Analysis

Descriptive statistics, including frequencies and percentages for categorical variables and mean ± standard deviation (SD) for continuous variables, were used to describe the characteristics of the patients. The top ten regimens (frequencies and percentages) for the first three LoTs were reported. Differences in characteristics and treatment regimens between the claims cohort and the EHR cohort were assessed for significance using standardized difference scores [25]. A threshold of 0.2, 0.5, or 0.8 suggested small, medium, or large effect sizes, respectively, between the two cohorts [25]. All statistical analyses were performed using SAS Studio version 3.6 (SAS Institute, Cary, NC, USA). Sankey diagrams were constructed to examine treatment



*Except lung cancer and non-melanoma skin cancer
** Lung cancer diagnosis made between January 2017 and December 2019

Fig. 1 Study design

pathways/trajectories using R Studio version 3.6.0 (R Studio, Boston, MA, USA).

3 Results

3.1 Patient Characteristics

Table 1 shows the cohort attrition for both the Optum claims and the Optum EHRs. The Optum claims cohort included a total of 7917 patients with NSCLC, 4010 (51%) of whom were female. Approximately 77% of the patients were aged ≥ 65 years, with a mean ± SD age of 70 ± 9 years at the time of systemic treatment initiation and a mean follow-up period of 373 days. The mean ± SD CCI score was 2.3 ± 1.9. The EHR cohort included 7087 patients, of whom 3549 (50%) were female. Approximately 58% of the patients were aged ≥ 65 years, with a mean ± SD age of 67 ± 10 years at the time of systemic treatment initiation and a mean follow-up period of 362 days. The mean ± SD CCI score was 1.6 ± 1.7. In both claims and EHR cohorts, the most represented geographical regions were midwest and south; in the claims cohort, 25% of the patients were from the midwest and 46% were from the south, whereas in the EHR cohort, 54% were from the midwest and 16% were from the south. Detailed patient characteristics of the two cohorts are presented in Table 2.

Within the EHRs, further patient characteristics were explored through NLP-derived terms. Of the 7087 patients, 6132 (87%) had clinical notes available to generate NLP-derived terms. Approximately 67 and 16% of the EHR cohort presented with non-squamous and squamous NSCLC, respectively. Such information was not available for the other 16% of patients. At the time of treatment initiation, roughly 24% of the patients were current smokers, 63% had a history of smoking, and approximately 13% had never smoked.

3.2 Treatment Patterns

Platinum-containing (carboplatin/cisplatin) chemotherapies were the most common LoT1 type, with 3374 (43%) and 2662 (38%) patients in the claims and EHR cohorts, respectively, receiving these treatments. Mean LoT1 duration was 146 ± 162 and 147 ± 173 days, with a median duration of 91 and 84 days in the claims and EHR cohorts, respectively. In both claims and EHR cohorts, the five most common LoT1 were the same: carboplatin + paclitaxel, pembrolizumab, carboplatin + pemetrexed + pembrolizumab, carboplatin + pemetrexed, and nivolumab. The standardized difference of the LoT1 distribution between the two cohorts was 0.25. Detailed LoT1 distribution by cohorts is presented in Table 3.

In the claims cohort, 2580 (33%) patients died after initiating their LoT1, and 2965 (37%) went on to their second line of therapy (LoT2), with a mean duration of 157 ± 157 days and a median of 98 days. In the EHR cohort, 2573 (36%) died after their LoT1, and 2229 (33%) went on to an LoT2, with a mean duration of 158 ± 158 days and a median of 99 days. The five most common LoT2 were also the same in both claims and EHR cohorts: durvalumab, nivolumab, pembrolizumab, carboplatin + pembrolizumab + pemetrexed, and carboplatin + pemetrexed. The standardized difference of LoT2 distribution between the two cohorts was 0.20. Immune checkpoint inhibitor monotherapy was the most common LoT2 type, with 1449 (49%) and 956 (43%) in the claims and EHR cohorts, respectively. Detailed LoT2 distribution by cohort is presented in Table 4.

Subsequently, 837 of the 2965 patients (28%) and 580 of the 2229 patients (26%) went on to receive a third LoT

Table 1 Cohort attrition

Inclusion/exclusion criteria	Claims			EHRs		
	n	% of overall	% of prior step	*n*	% of overall	% of prior step
1. One or more diagnosis of lung cancer with metastasis during the period January 2017 to December 2019	63420	100		76119	100	
2. No other cancer (primary or secondary) (except NMSC and lung cancer) prior to metastasis	25600	40	40	37766	50	50
3. Systemic treatment after metastasis (index on the first systemic treatment date)	13133	21	51	11142	15	30
4. With continuous health plan enrollment or active EHRs ≥ 6 months before treatment initiation date and ≥ 1 month after treatment initiation date	9723	15	74	9088	12	82
5. Age ≥ 18 years at index date	9723	15	100	9087	12	100
6. With known sex	9722	15	100	9084	12	100
7. Exclude SCLC-specific treatments	7917	12	81	7087	9	78

EHRs electronic health records, *NMSC* non-melanoma skin cancer, *SCLC* small-cell lung cancer

Table 2 Patient characteristics in the claims cohort and the electronic health records cohort

	Claims cohort ($N = 7917$)	EHR cohort ($N = 7087$)	Standardized difference[a]
Age (years)	70 ± 9; 71	67 ± 10; 67	0.35
Age group			
18–54	399 (5)	744 (10)	0.39
55–64	1440 (18)	2214 (31)	
65–74	3405 (43)	2349 (33)	
≥ 75	2673 (34)	1780 (25)	
Sex			
Female	4010 (51)	3557 (50)	− 0.01
Male	3907 (49)	3551 (50)	
Regions			
Midwest	1991 (25)	3835 (55)	0.80
Northeast	1134 (14)	1406 (20)	
South	3613 (46)	1157 (17)	
West	1158 (15)	532 (8)	
Unknown	21	178	
Race			
White	4945 (81)	5881 (87)	0.14
Black	913 (15)	716 (11)	
Asian	239 (4)	196 (3)	
Unknown	1820	315	
Follow-up duration (days)	373 ± 299; 295	362 ± 303; 280	0.04
CCI[b]	2.3 ± 1.9; 2	1.6 ± 1.7; 1	0.36
CCI group			
0	1070 (14)	1944 (27)	0.39
1	2263 (29)	2257 (32)	
2	1772 (22)	1254 (18)	
3–4	1741 (22)	1077 (15)	
≥ 5	1071 (14)	555 (8)	
First treatment initiation year (cohort entry year)			
2017	1931 (24)	2073 (29)	0.15
2018	2486 (31)	2226 (31)	
2019	2803 (35)	2309 (33)	
2020[c]	697 (9)	479 (7)	

Data are presented as *n* (%) or mean ± standard deviation; median

CCI Charlson Comorbidity Index, *EHRs* electronic health records

[a]In interpretation of standardized difference, thresholds of 0.2, 0.5, and 0.8 suggest small, medium, and large effect sizes, respectively, between the two cohorts[25]

[b]CCI scoring excluded malignancy

[c]Initiation year 2020 includes data until September 2020

(LoT3), with a mean and median treatment duration of 121 ± 124 and 78 days, and 138 ± 153 and 85 days in the claims and EHR cohorts, respectively. The five most common LoT3 were docetaxel, gemcitabine, nivolumab, pembrolizumab, and docetaxel + ramucirumab in both cohorts.

Table 3 The ten most common first lines of therapy in the claims cohort and the electronic health records cohort

Claims cohort		EHR cohort	
LoT1	*n* (%)	LoT1	*n* (%)
Patients with LoT1	7917 (100)	Patients with LoT1	7108 (100)
1. Carboplatin, paclitaxel	1794 (23)	1. Carboplatin, paclitaxel	1306 (18)
2. Pembrolizumab	1248 (16)	2. Pembrolizumab	1169 (16)
3. Carboplatin, pembrolizumab, pemetrexed	1093 (14)	3. Carboplatin, pembrolizumab, pemetrexed	881 (12)
4. Carboplatin, pemetrexed	720 (9)	4. Carboplatin, pemetrexed	598 (8)
5. Nivolumab	375 (5)	5. Nivolumab	400 (5)
6. Osimertinib	280 (4)	6. Osimertinib	368 (5)
7. Cisplatin, pemetrexed	279 (4)	7. Cisplatin, pemetrexed	295 (4)
8. Carboplatin, paclitaxel, pembrolizumab	195 (2)	8. Erlotinib	193 (3)
9. Carboplatin, gemcitabine	128 (2)	9. Carboplatin, paclitaxel, pembrolizumab	155 (3)
10. Carboplatin, paclitaxel protein-bound	123 (2)	10. Durvalumab	141 (2)
Standardized difference: 0.25			

EHRs electronic health records, *LoT1* first line of therapy

Table 4 The ten most common second lines of therapy in the claims cohort and the electronic health records cohort

Claims cohort		EHR cohort	
LoT2	*n* (%)	LoT2	*n* (%)
Patients with LoT2	2965 (100)	Patients with LoT2	2229 (100)
1. Durvalumab	671 (23)	1. Durvalumab	437 (20)
2. Nivolumab	353 (12)	2. Nivolumab	260 (12)
3. Pembrolizumab	316 (11)	3. Pembrolizumab	206 (9)
4. Carboplatin, pembrolizumab, pemetrexed	149 (5)	4. Carboplatin, pembrolizumab, pemetrexed	146 (7)
5. Carboplatin, pemetrexed	131 (4)	5. Carboplatin, pemetrexed	110 (5)
6. Carboplatin, paclitaxel	112 (4)	6. Osimertinib	93 (4)
7. Docetaxel	109 (4)	7. Carboplatin, paclitaxel	85 (4)
8. Osimertinib	106 (4)	8. Docetaxel, ramucirumab	84 (4)
9. Docetaxel, ramucirumab	101 (3)	9. Docetaxel	72 (3)
10. Atezolizumab	91 (3)	10. Gemcitabine	58 (3)
Standardized difference: 0.20			

EHRs electronic health records, *LoT2* second line of therapy

In LoT3, monotherapy regimens were predominant. A new combination regimen (docetaxel + ramucirumab) was observed as one of the common LoT3. The standardized difference of the LoT3 distribution between the two cohorts was 0.38. See Table 5 for details.

Figure 2 shows Sankey diagrams of metastatic NSCLC treatments by LoT. From both the Optum claims and the Optum EHR cohorts, the main pathways from LoT1 to PD-L1 agents as LoT2 were through platinum-based doublet chemotherapies: carboplatin/cisplatin + paclitaxel and carboplatin/cisplatin + pemetrexed.

Over treatment initiation years, platinum-containing chemotherapy as LoT1 declined. From 2017 to 2019, its use was 52%, 46%, and 37% of LoT1 in the claims cohort and 46%, 39%, and 31% of LoT1 in the EHR cohort. Immune checkpoint inhibitor monotherapy (e.g., pembrolizumab, nivolumab) as LoT1 was relatively stable over index years, with 22 and 25% in the claims and EHR cohorts, respectively.

4 Discussion

This study explored two US nationwide healthcare utilization data sources to examine the current treatment landscapes for patients with metastatic NSCLC. The claims cohort was slightly older than the EHR cohort (mean age 70 vs. 67 years; standardized difference 0.35). Ages of both cohorts were within the range of the US lung cancer statistics. In NSCLC, treatment decisions are primarily guided by stage and tumor characteristics (e.g., histology, tumor molecular testing) and patient's performance status [17]. As mean and median ages of both cohorts aligned with general US statistics for NSCLC (mostly diagnosed at age 65–74 years, with median age at diagnosis of 71 years [2]), we did not consider differences in age distribution would significantly affect treatment choices. A similar rationale was taken for region, another demographic variable that differed between the cohorts. The southern regions were overrepresented in claims data, and the midwest regions were overrepresented in the EHR data when compared with US statistics [2].

The claims cohort was slightly sicker than the EHR cohort (mean CCI 2.3 vs. 1.6; standardized difference 0.36). Although the Optum EHRs are not limited to oncologists, they still may not include a complete comorbidity profile for patients with NSCLC as they are an open data source. However, both cohorts were comparable in sex, race, treatment initiation year, and follow-up period and showed similar treatment patterns in terms of choice of regimen, progression to subsequent therapy, and duration of therapy. This implies relatively uniform NSCLC treatment approaches regardless of age and region in the USA. However, the literature

Table 5 The ten most common third lines of therapy in the claims cohort and the electronic health records cohort

Claims cohort		EHR cohort	
LoT3	*n* (%)	LoT3	*n* (%)
Patients with LoT3	837 (100)	Patients with LoT3	580 (100)
1. Gemcitabine	81 (10)	1. Docetaxel	64 (11)
2. Pembrolizumab	68 (8)	2. Gemcitabine	62 (11)
3. Docetaxel	58 (7)	3. Docetaxel, ramucirumab	54 (9)
4. Nivolumab	57 (7)	4. Nivolumab	45 (8)
5. Durvalumab	48 (6)	5. Pembrolizumab	31 (5)
5. Docetaxel, ramucirumab	48 (6)	6. Carboplatin, pemetrexed	27 (5)
7. Carboplatin, paclitaxel	42 (5)	7. Carboplatin, pembrolizumab, pemetrexed	26 (4)
8. Osimertinib	39 (5)	8. Durvalumab	24 (4)
9. Carboplatin, pembrolizumab, pemetrexed	33 (4)	9. Pemetrexed	21 (4)
10. Pemetrexed	30 (4)	10. Paclitaxel	17 (3)
Standardized difference: 0.38			

EHRs electronic health records, *LoT3* third line of treatment

suggests that age may influence whether patients receive treatments [18, 19]. We discuss this aspect further in Sect. 4.

Different data features between the two data sources should be noted. First, the claims data provided more comprehensive healthcare resource utilization information for each patient, whereas the EHR data provided more comprehensive clinical information available within the institution. EHRs provide the patient's clinical records regardless of health insurance coverage. They also provide more detailed and comprehensive clinical information by capturing data from clinical notes and laboratory results.

In terms of treatment patterns, the claims data provide more definite treatment patterns, as a record of processed claims indicates the actual medication received. On the other hand, EHRs provide administered and prescribed medication records. Prescribed medication records do not provide any information on the actual patient's drug acquisition and/or drug utilization. Prescription abandonment (i.e., patients not picking up prescribed medications at pharmacies) is a known issue in the USA [20]. This is a significant disadvantage of EHRs when examining treatment patterns. However, it may not have a significant impact on this study for two reasons. First, cancer drugs are a protected drug class for prescription drug benefits under the Medicare program. Given that the majority of the cohort was aged > 65 years (77% in claims, 58% in EHRs), the eligible age for Medicare, we assumed that the drug coverage by health plans did not play a significant role in determining the treatment choices in this study. Second, most NSCLC treatments are administered via infusion. Within-hospital or clinic uses are well captured through administered medication records from EHRs. Additionally, the similarity of oral chemotherapeutic agent uptake (e.g., 4 and 5% osimertinib) as LoT1 in the claims and EHR cohorts, respectively, further supported this position. The Sankey diagram of treatment pathways between the two cohorts suggested similar regimen choices by lines of therapy as well as similar general subsequent regimen trajectories. This suggests that general treatment approaches in NSCLC are relatively standardized in the USA, and new clinical evidence/guideline recommendations are rapidly adopted into practice.

One strength of our study is that we examined the most up-to-date treatment patterns in patients with NSCLC. In the USA, the first immune checkpoint inhibitor was approved for NSCLC treatment in late 2015, and some of the targeted therapy drugs were approved within the last 5 years. In addition to identifying commonly used regimens, newer treatment options may also influence the overall patient characteristics of patients with metastatic NSCLC who initiated treatment.

Traditionally, platinum-containing chemotherapy (i.e., without combination with newer treatment options such as immune checkpoint inhibitors or targeted therapy) was a

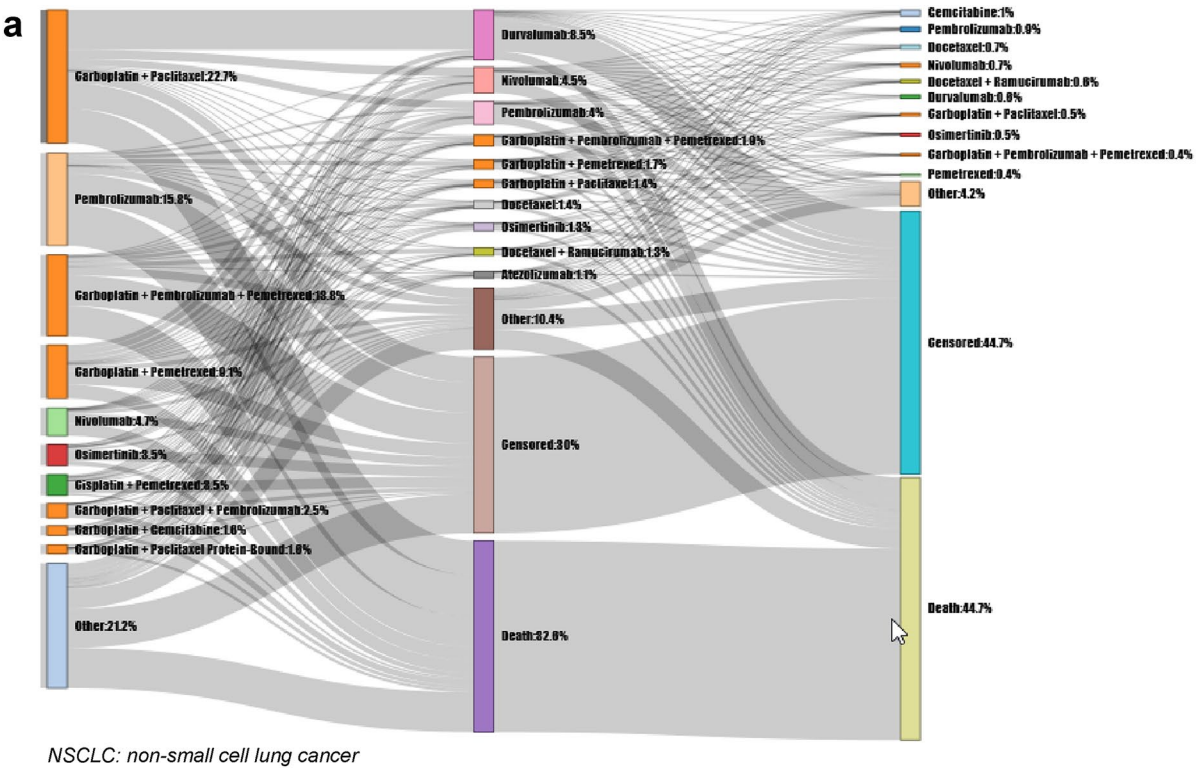


Fig. 2 Sankey diagram of metastatic non-small-cell lung cancer treatments by line of therapy in **a** the claims cohort ($N = 7917$) and **b** the electronic health records cohort ($N = 7087$)

standard of care in patients with metastatic NSCLC. In this study, platinum-containing chemotherapy regimens were the most common LoT1. However, its use declined each year, as noted in Sect. 3.2. It is worth noting that LoT1 that consist of platinum-containing chemotherapy in combination with newer treatment options increased between 2017 and 2019, with 13%, 20%, and 28% of LoT1 in the claims cohort and 9%, 18%, and 26% of LoT1 in the EHR cohort.

We explored the potential additional utility of EHRs. Approximately 12% of the metastatic stage was additionally captured by NLP terms, compared with using only ICD codes in the EHR cohort. After patients receiving SCLC treatments were excluded, a further 3% of patients were excluded because NLP terms indicated they had SCLC. This may indicate that the NSCLC cohort algorithm adapted from the literature [13, 14] with updated treatments reliably captured the NSCLC cohort through RWD. After the cohort was defined, we explored smoking status and histology, information that is useful for NSCLC management and a potential advantage of EHRs.

We were unable to estimate the overlap of patients between the two study cohorts as we could only analyze de-identified patient data. According to Optum, about 26% of the patients with an ICD code for lung cancer in the EHRs can be linked with the Optum claims. In both cohorts, the criterion that lost most patients was the inclusion of patients with systemic treatments after the metastasis date, excluding roughly 50 and 70% from the prior step in claims and EHRs, respectively. David et al. [17] found that 22–26% of patients with advanced/metastatic NSCLC received no treatment in 1998–2012. Kehl et al. [18] reported that approximately one-half of elderly patients (aged $\geq$ 66 years) received a systemic treatment within 1 year of the metastatic NSCLC diagnosis between 2012 and 2015. Although the estimates varied by study population, studies generally identified that certain patient demographics/characteristics were associated with not receiving any systemic treatment, including older age [18, 19], low socioeconomic status [18, 24], and poorer prognosis [19, 24]. Given that this study's cohorts were patients with metastatic disease and therefore a worse prognosis, and considering the mean ages of the cohorts, it may be conceivable that a relatively large proportion of patients did not receive treatments. Excluded patients would also have included patients who were lost to follow-up before treatment initiation because of discontinued health plan enrollment or no patient activity found in the EHR system. The larger number of patients excluded in the EHR cohort might be influenced by insurance status and the possibility of receiving treatment somewhere else after diagnosis. From the step that excluded SCLC-specific treatments in the attrition table, both cohorts lost roughly 18–20% from the eligible total of patients with metastatic lung cancer, which seems reasonable as roughly 15% of lung cancers are SCLC [5].

A few limitations need to be noted before interpreting the study findings. First, both data sources were primarily collected for non-research uses. Therefore, the databases may be subject to coding errors and missing values. Second, the claims cohort only included commercially insured patients under one payer (Medicare Advantage patients), which represents a smaller proportion of all Medicare enrollees. The EHR cohort may overrepresent the practice patterns of large healthcare systems as > 80% of the cohort patients were from an integrated delivery system. Third, the NLP tables in EHRs may be subject to missingness and potential extraction errors. Finally, as there were no ICD codes for NSCLC, and we used treatments as a proxy, this study only included patients with NSCLC receiving treatment.

However, the similarities in patient demographics and treatment patterns between the two cohorts suggest that this study can provide relevant insights into current clinical practices in NSCLC treatments despite the mentioned limitations. Further studies are needed to explore the comparison between claims and EHRs to understand treatment patterns in other therapeutic areas.

5 Conclusion

This study used two different RWD sources (health insurance claims and EHRs) to capture patients with metastatic NSCLC, demographic and clinical characteristics, and treatment patterns. In both cohorts, commonly used medication regimens were comparable for all LoT1, LoT2, and LoT3. For the LoT1, traditional chemotherapy was more common, followed by immune checkpoint inhibitor monotherapy (e.g., pembrolizumab, nivolumab). With subsequent lines of therapy, new treatment modalities were commonly adopted in patients with metastatic NSCLC, especially immune checkpoint inhibitor monotherapy. Despite differences in data features, the findings suggest that the two cohorts utilizing RWD provided up-to-date insight into the current treatment modalities in patients with metastatic NSCLC.

Appendix

Table 6 Lung cancer drugs included in the study

Generic names	Brand names (US)	HCPCS codes
Afatinib	Gilotrif	
Alectinib	Alecensa	
Atezolizumab	Tecentriq	J9022, C9483
Bendamustine	Bendeka, Treanda, Belrapzo	J9033, J9034, J9036
Bevacizumab*	Avastin, Mvasi, Zirabev	J9035, Q5107, Q5118
Brigatinib	Alunbrig	
Capmatinib	Tabrecta	
Carboplatin	Paraplatin	J9045
Cemiplimab	Libtayo	J9119, C9044
Ceritinib	Zykadia	
Cisplatin	Platinol	J9060
Crizotinib	Xalkori	
Cyclophosphamide	Cytoxan, Neosar	J9070, J8530
Dabrafenib	Tafinlar	
Dacomitinib	Vizimpro	
Docetaxel	Taxotere	J9171
Doxorubicin	Adriamycin, Rubex	Q2049, Q2050
Durvalumab	Imfinzi	J9173, C9492
Entrectinib	Rozlytrek	
Erlotinib	Tarceva	
Etoposide	Toposar, Etopophos	J8560, J9181, J9182, C9414, C9425
Gefitinib	Iressa	J8565
Gemcitabine	Gemzar	J9201
Ifosfamide	Ifex	J9208
Ipilimumab	Yervoy	J9228, C9284
Irinotecan	Camptosar	J9206
Larotrectinib	VitrakviI	
Lorlatinib	Lorbrena	
Necitumumab	Portrazza	J9295, C9475
Nivolumab	Opdivo	J9299, C9453
Osimertinib	Tagrisso	
Paclitaxel	Taxol	J9267
Paclitaxel, protein-bound	Abraxane	J9264
Pembrolizumab	Keytruda	J9271, C9027
Pemetrexed	Alimta	J9305
Pralsetinib	Gavreto	
Ramucirumab	Cyramza	J9308, C9025
Selpercatinib	Retevmo	
Temozolomide	Temodar	J8700, J9328
Tepotinib	Tepmetko	
Topotecan	Hycamtin	J9351, J8705
Trametinib	Mekinist	
Vinblastine	Velban	J9360
Vincristine	Oncovin, Vincasar, Vincrex	J9370, J9371
Vinorelbine	Navelbine	J9390

HCPCS: Healthcare Common Procedure Coding System

*Bevacizumab includes biosimilars

Acknowledgements The authors thank Robert Diegidio (AbbVie, Inc.) for his assistance in running the SAS programs to manipulate the health insurance claims data and thank Xiaomeng Yue (AbbVie, Inc.) for her help in running the R programs for the Sankey diagrams.

Declarations

Funding This study was sponsored by AbbVie, Inc.

Conflict of interest DZ and JT are employees of AbbVie, Inc. and owned AbbVie stocks or stock options at the time of the study. YC was a research fellow of AbbVie, Inc. at the time of the study.

Ethical approval Not applicable.

Consent Not applicable.

Availability of data and material Data for this study are not publicly available but may be obtained from the third party.

Code availability Not applicable.

Author contributions JT and DZ contributed to the conception and design of this study. Data analyses were performed by YC and DZ. The manuscript was drafted by YC and DZ. All authors critically reviewed all drafts of the manuscript and consented to the publication of the manuscript in its present form.

References

1. Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Ann Glob Health. 2019;85(1):8. https://doi.org/10.5334/aogh.2419.
2. Howlader N. et al. SEER cancer statistics review 1975–2018. Bethesda, MD: National Cancer Institute; 2021.
3. Molina JR, et al. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc. 2008;83(5):584–94.
4. Schnabel PA, et al. Influence of histology and biomarkers on first-line treatment of advanced non-small cell lung cancer in routine care setting: baseline results of an observational study (FRAME). Lung Cancer. 2012;78(3):263–9.
5. Rosell R, et al. Screening for epidermal growth factor receptor mutations in lung cancer. N Engl J Med. 2009;361(10):958–67.
6. Rikova K, et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell. 2007;131(6):1190–203.
7. Ding L, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455(7216):1069–75.
8. Solomon B, Varella-Garcia M, Camidge DR. ALK gene rearrangements: a new therapeutic target in a molecularly defined subset of non-small cell lung cancer. J Thorac Oncol. 2009;4(12):1450–4.
9. Herbst RS, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. The Lancet. 2016;387(10027):1540–50.
10. Aggarwal C, et al. Prevalence of PD-L1 expression in patients with non-small cell lung cancer screened for enrollment in KEYNOTE-001, -010, and -024. Ann Oncol. 2016;27:vi363.
11. Peters BJM, et al. Trends in prescribing systemic treatment and overall survival for non-small cell lung cancer stage IIIB/IV in the Netherlands: 2008–2012. Cancer Epidemiol. 2017;51:1–6.
12. Sheng Duh M, et al. Costs associated with intravenous chemotherapy administration in patients with small cell lung cancer: a retrospective claims database analysis. Curr Med Res Opin. 2008;24(4):967–74.
13. Turner RM, Chen Y-W, Fernandes AW. Validation of a case-finding algorithm for identifying patients with non-small cell lung cancer (NSCLC) in administrative claims databases. Front Pharmacol. 2017;8:883–883.
14. Glasheen WP, et al. Charlson comorbidity index: ICD-9 update and ICD-10 translation. Am Health Drug Benefits. 2019;12(4):188.
15. Thygesen SK, et al. The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients. BMC Med Res Methodol. 2011;11(1):1–6.
16. Ettinger DS, et al. NCCN guidelines insights: non-small cell lung cancer, version featured updates to the NCCN guidelines. J Natl Compr Cancer Netw. 2021;19(3):254–66.
17. David EA, et al. Increasing rates of no treatment in advanced-stage non–small cell lung cancer patients: a propensity-matched analysis. J Thorac Oncol. 2017;12(3):437–45.
18. Kehl KL, Hassett MJ, Schrag D. Patterns of care for older patients with stage IV non-small cell lung cancer in the immunotherapy era. Cancer Med. 2020;9(6):2019–29.
19. Small AC, et al. Prevalence and characteristics of patients with metastatic cancer who receive no anticancer therapy. Cancer. 2012;118(23):5947–54.
20. Doshi JA, et al. Association of patient out-of-pocket costs with prescription abandonment and delay in fills of novel oral anticancer agents. J Clin Oncol. 2018;36(5):476–82.
21. Cassidy RJ, et al. Health care disparities among octogenarians and nonagenarians with stage III lung cancer. Cancer. 2018;124(4):775–84.
22. Tufman ALH, et al. Preselection based on clinical characteristics in german non–small-cell lung cancer patients screened for EML4-ALK translocation. J Thorac Oncol. 2014;9(1):109–13.
23. Shaw AT, et al. Clinical features and outcome of patients with non-small-cell lung cancer who harbor EML4-ALK. J Clin Oncol Off J Am Soc Clin Oncol. 2009;27(26):4247–53.
24. Rodig SJ, et al. Unique clinicopathologic features characterize ALK-rearranged lung adenocarcinoma in the western population. Clin Cancer Res. 2009;15(16):5216–23.
25. Yang D, Dalton J. A unified approach to measuring the effect size between two groups using SAS; 2012. Available at: https://support.sas.com/resources/papers/proceedings12/335-2012.pdf. Accessed 26 July 2021.