GOOD SCIENCE
BETTER MEDICINE
BEST PRACTICE

REAL WORLD
DATA AND DIGITAL
ONCOLOGY

ORIGINAL ARTICLE

Data analytics for real-world data integration in TKI-treated NSCLC patients using electronic health records

L. Mazzeo[1,2†], F. Corso[1†], P. Baili[3†], F. Scotti[4], V. Torri[5], M. Ganzinelli[1], V. Mišković[1,2], R. Leporati[1], L. Provenzano[1], A. Spagnoletti[1], C. Silvestri[1], C. Giani[1], C. Cavalli[1], R. M. di Mauro[1], M. Meazza Prina[1], C. Proto[1], M. Brambilla[1], M. Occhipinti[1], S. Manglaviti[6], T. Beninato[1], D. Miliziano[1], A. D. Dumitrascu[1], G. Di Liberti[1], T. S. Cassano[1], F. G. M. de Braud[1], G. L. Russo[1], A. Cappozzo[7], A. M. Paganoni[5], F. Ieva[5,8‡] & A. Prelaj[1,2*‡]

[1]Department of Medical Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan; [2]Department of Electronic, Information and Bioengineering, Politecnico di Milano, Milan; [3]Department of Epidemiology and Data Science, Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan; [4]Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milan; [5]MOX — Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Milan; [6]Oncology Unit, ASST Ospedale Maggiore di Crema, Crema; [7]Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milan; [8]Health Data Science Centre, Human Technopole, Milan, Italy

Available online xxx

Background: Real-world data (RWD) are routinely collected in clinical practice during therapeutic interventions. Data warehouses (DWHs) represent the primary source of RWD in which electronic health records (EHRs) can be rapidly analyzed via natural language processing. This study illustrates an analytic framework that systematically exploits RWD and methods to generate real-world evidence (RWE) about innovative cancer drugs. The framework has been applied to investigate real-world treatment patterns and clinical outcomes of patients with advanced non-small-cell lung cancer (aNSCLC) treated with tyrosine kinase inhibitors (TKIs).

Materials and methods: Data from a cohort of 190 epidermal growth factor receptor-positive mutation (EGFRm) patients with aNSCLC were retrospectively collected in an Italian cancer institute between 2014 and 2022. Patients were treated in first-line (1L) with osimertinib or other TKIs (non-osimertinib). A text-mining algorithm was implemented to retrieve RWD from EHRs. Survival endpoints were median time to treatment discontinuation (mTTD) and median overall survival (mOS) estimated with Kaplan—Meier curves. Time-dependent multivariate Cox analysis was carried out to overcome immortal time bias.

Results: Approximately 38% of patients received 1L osimertinib, while the remaining 62% received previous-generation TKIs. Longer mTTD [15 months; 95% confidence interval (CI) 11.9-26.4 months] was found for patients treated with 1L osimertinib compared with non-osimertinib (10 months; 95% CI 7.9-13.1 months). In multivariate analysis, osimertinib was an independent protective factor regardless of bone and brain metastases and local radiotherapy. mOS was 27 months (95% CI 21.4-39.5 months) for osimertinib versus 20.2 months (95% CI 17.6-23.1 months) for non-osimertinib.

Conclusions: Data analytics frameworks are useful tools to integrate RWE in cancer research and data-driven models are suitable to process large amounts of RWD. This study demonstrates that real-world treatment patterns and outcomes of TKIs are comparable with those found in both clinical trials and other real-world studies. RWE studies can support clinicians in investigating the best treatment strategy and decision makers to drive new health policies.

Key words: non-small cell lung cancer, tyrosine kinase inhibitors, osimertinib, real-world data, text-mining, survival analysis

INTRODUCTION

Real-world data (RWD), collected routinely by hospitals' data systems, serve as a complementary information source to randomized controlled trials (RCTs). RCTs still represent the gold standard for determining drug efficacy and securing regulatory approvals. However, limited by high costs, strict eligibility criteria, and controlled conditions they do not always reflect real-world settings.

*Correspondence to: Dr Arsela Prelaj, Fondazione IRCCS Istituto Nazionale dei Tumori, 20133 Milan, Italy. Tel: +02 2390 3647
E-mail: arsela.prelaj@istitutotumori.mi.it (A. Prelaj).

†These authors contributed equally to this work as co-first authors.

‡These authors contributed equally to this work as co-last authors.

RWD generate real-world evidence (RWE), which provides insights into the real-world benefits and risks of medical interventions, supporting clinical decision making.

RWD primarily originate from electronic health records (EHRs) in hospital data warehouses (DWHs), where 80% of information is in unstructured text.[1] However, EHRs pose challenges due to their high dimensionality, inconsistency, and bias. Consequently, health care systems need data analytics frameworks to systematically collect, organize, and analyze RWD, bridging the gap between RCTs and RWE. These frameworks integrate data from diverse hospital departments, though data extraction, linkage, and quality assessments remain challenging.

Recent advancements in data collection, information sharing, and data transfer technologies have greatly enhanced the ability to implement analytical frameworks based on RWD in health care. Additionally, the growing availability of statistical and machine learning algorithms enables efficient processing of large datasets, making it easier to generate valuable insights and new knowledge in the health care sector.

For instance, natural language processing (NLP) uses text-mining algorithms to identify, extract, and analyze relevant information from unstructured data formulated in human language.[2] Although NLP algorithms require data quality assessment to evaluate the accuracy of information retrieval, they offer more efficient and scalable methods for generating RWD than alternatives based on manual extractions.

This study applies an analytic framework for RWD on advanced non-small-cell lung cancer (aNSCLC), a leading global cancer with high mortality.[3,4]

Tyrosine kinase inhibitors (TKIs) constitute the standard of care (SoC) in the first-line (1L) treatment of aNSCLC for patients with epidermal growth factor receptor-positive mutation (EGFRm), offering superior outcomes in both quality of life and response rates compared with chemotherapy (CTx).[5,6]

Since 2016, osimertinib, a third-generation TKI, has been initially approved as a new treatment for patients with aNSCLC and EGFRm who have progressed after EGFR first/second-generation TKIs (1st/2nd-gen TKIs). Subsequently, osimertinib has been extended to 1L treatment, showing better clinical management than 1st/2nd-gen TKIs in the phase III FLAURA trial.[7]

To date, osimertinib remains the SoC offering extended progression-free survival (PFS) and overall survival (OS) compared with earlier TKIs.[8]

Given that real-world treatment involves more frequent therapy changes than clinical trials, this study complements extant literature through a data analytics framework devised to monitor the last decade of real-world patterns and treatment outcomes of EGFRm patients with aNSCLC patients treated in 1L with erlotinib/gefinitib (1st-gen TKI), afatinib (2nd-gen TKI), and osimertinib (3rd-gen TKI).

This study advances robust analytical methods for RWE studies, using time to treatment discontinuation (TTD) as a real-world endpoint that reflects treatment dynamics and disease management in routine care. Unlike PFS, which relies on standardized disease assessments, and may report inconsistencies in real-world dataset, TTD is coherently recorded in EHRs. Therefore, it provides a realistic view of cancer drug safety and efficacy in the real world aligning with factors such as toxicity, patient preference and the common practice of treatment beyond RECIST-defined progression.[9,10]

Furthermore, multivariate and time-dependent survival analyses are introduced in this analytic framework to address confounding effects and immortal time bias, commonly occurring in observational cohort studies when participants cannot experience the study outcome during a certain period of follow-up.[11,12]

To the best of our knowledge, the current Italian research landscape still lacks a data analytics framework setting out the collection, linkage, and analysis of all the routine health data. Specifically, this work supports the adoption of NLP to extract RWD from EHRs stored in hospitals' DWHs in line with the European Society for Medical Oncology Guidance for Reporting Oncology real-World evidence (ESMO-GROW) recently introduced by the ESMO.[11]

MATERIALS AND METHODS

Data analytics framework

Figure 1 presents a schematic of the data analytics framework, consisting of three layers: study design, data collection, and data analysis.

These steps establish a sequential framework for conducting RWE studies, especially useful for researchers who are less familiar with this type of analysis. This systematic approach ensures a robust methodology when analyzing RWD and is adaptable to a broad spectrum of RWE questions.

The first layer in designing a real-world study involves conducting a literature review to establish the study context and identify existing research gaps. This analysis informs the formulation of research questions, which define the study's aims and objectives. Broadly, research questions may explore adherence to therapeutic guidelines in real clinical settings or investigate new epidemiological insights about a particular disease. Additionally, this phase includes setting inclusion and exclusion criteria to define the study cohort.

The data layer focuses on identifying suitable data sources, primarily DWHs, which house structured data and unstructured data. Data may come from hospital databases and other referral systems, with variability in the temporal patterns across sources. This step often requires manual review to finalize case selection for analysis.

The third layer involves selecting appropriate analytical methodologies to address the research questions. RWE research employs various approaches, including NLP for text mining, descriptive analysis for clinical characteristics, survival analysis for clinical effectiveness, and health economic evaluations to assess cost-benefit measures. Data-driven models are especially effective in analyzing longitudinal data for RWE studies.

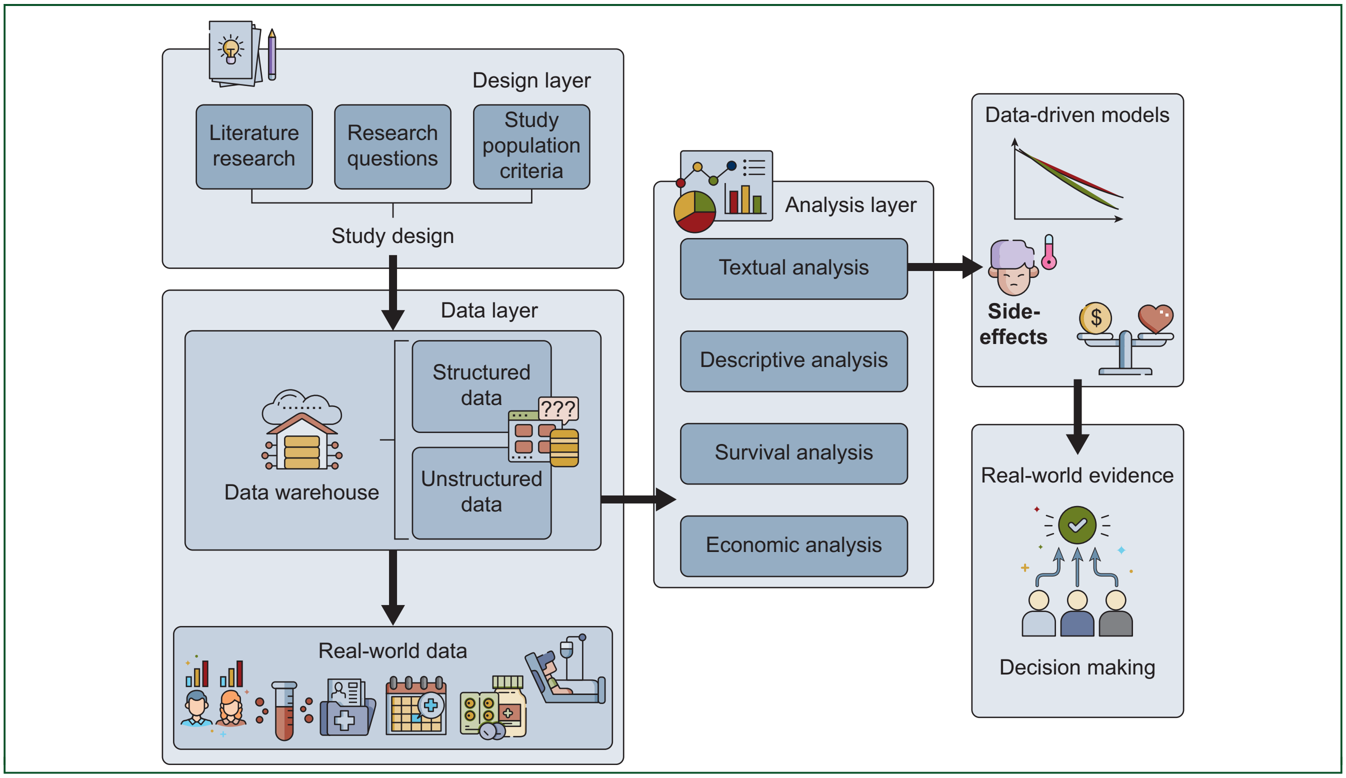


Figure 1. Data analytics framework. Graphical illustration of the three layers comprising the data analytics framework.

Study design and population

This is a real-world study conducted in an Italian cancer institute. The study was approved by the hospital ethics committee. Data were retrospectively collected for patients who fulfilled the eligibility criteria. The following inclusion criteria were applied to identify the real-world population: age ≥18 years; Eastern Cooperative Oncology Group performance status (ECOG PS) 0-4; histologically or cytologically diagnosed with stage IIIB/IV NSCLC treated in 1L with first-, second-, or third-generation of EGFR-TKIs. Conversely, the exclusion criteria were administration of 1L treatment different from EGFR-TKIs; any EGFR-TKI initiation was suggested in our hospital but carried out in other institutions. The study index date was defined as the date of initiation of the 1L treatment between 2014 and 2022. Patients were observed until death, loss to follow-up, or study cut-off date (27 July 2023), whichever occurred first.

Real-world data collection and extraction

INT Data Warehouse. The INT DWH is a real-world patient-oriented database that exploits heterogeneous data sources including the information contained in EHRs and combines them into a single and unified system. Among the main features of a DWH are integration, consistency, and representation of temporal evolution. In particular, integration and consistency are fundamental aspects to ensure the correct management of several data sources, which come from different hospital departments or external information systems. Unlike standard hospital EHR systems that focus on admissions and outpatient data, the INT DWH centralizes information from multiple clinical applications (e.g. pathology, radiology, admissions, surgery, chemotherapy, radiotherapy), enabling researchers to query data across these systems for research purposes. At first access in the institute, each patient is assigned to a unique and unchanging INT patient identification code, with which the hospital centrally manages the patient's records. Neither image nor omics data are currently stored in INT DWH. More details about DWH governance, ownership, and accessibility are reported in Supplementary Material, Section 1.1, available at https://doi.org/10.1016/j.esmorw.2024.100109. Patient extraction from DWH is described in detail in Figure 2.

Data extraction via SQL and text mining. Clinical information was extracted from the INT DWH by querying patient health records for baseline data. Data sources included the Cartella Clinica Elettronica (CCE) and Rete Oncologica Lombarda (ROL), both stored in the DWH (see Supplementary Material, Section 1.2, Supplementary Figure S1, and Supplementary Table S1, available at https://doi.org/10.1016/j.esmorw.2024.100109). Two methods were used: Structured Query Language (SQL) queries addressed anatomical names and clinical acronyms, while a rule-based named entity recognition (NER) technique extracted complex medical entities. Details on the algorithm are provided in Supplementary Material, Section 2.1, available at https://doi.org/10.1016/j.esmorw.2024.100109, and the extraction process is shown in Figure 3.

The following clinical variables were extracted with SQL: smoking habits, programmed death-ligand 1 (PD-L1) gene

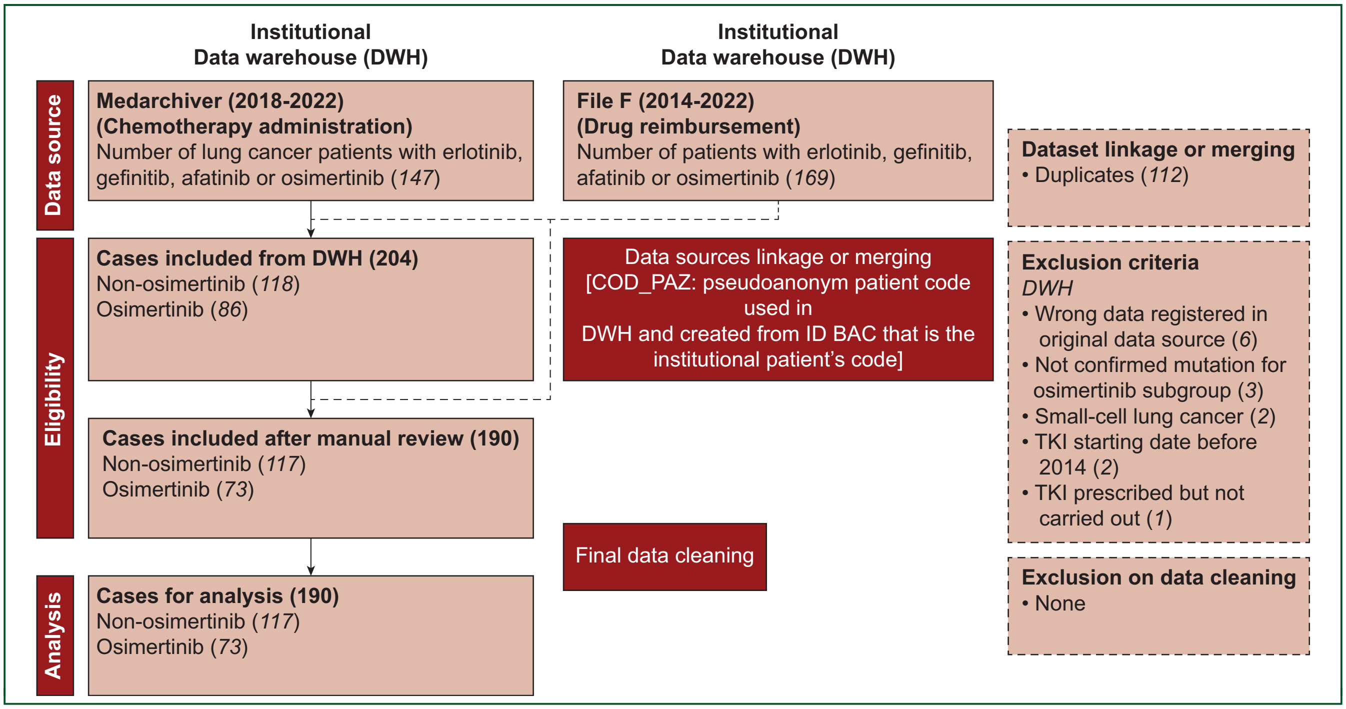

Institutional
Data warehouse (DWH)
Institutional
Data warehouse (DWH)
Data source
Medarchiver (2018-2022)
(Chemotherapy administration)
Number of lung cancer patients with erlotinib, gefinitib, afatinib or osimertinib (147)
File F (2014-2022)
(Drug reimbursement)
Number of patients with erlotinib, gefinitib, afatinib or osimertinib (169)
Dataset linkage or merging
• Duplicates (112)
Eligibility
Cases included from DWH (204)
Non-osimertinib (118)
Osimertinib (86)
Data sources linkage or merging [COD_PAZ: pseudoanonym patient code used in DWH and created from ID BAC that is the institutional patient's code]
Exclusion criteria
DWH
• Wrong data registered in original data source (6)
• Not confirmed mutation for osimertinib subgroup (3)
• Small-cell lung cancer (2)
• TKI starting date before 2014 (2)
• TKI prescribed but not carried out (1)
Cases included after manual review (190)
Non-osimertinib (117)
Osimertinib (73)
Final data cleaning
Exclusion on data cleaning
• None
Analysis
Cases for analysis (190)
Non-osimertinib (117)
Osimertinib (73)

Figure 2. ESMO-GROW flowchart. Process of patient extraction from the INT DWH.
TKI, tyrosine kinase inhibitor.

expression, histology, sites of metastases, and staging. ECOG PS was the only clinical field in CCE and ROL recorded as structured data. Other clinical variables, like EGFR mutation type, oligoprogressive diseases (OPDs), and Response Evaluation Criteria in Solid Tumors (RECIST)-based treatment response, were manually annotated. Survival endpoints, such as treatment discontinuation and death status, were also manually curated from clinician notes. A rule-based NER algorithm extracted drug toxicity data by identifying toxicity episodes through specific clinical terms and associating them with treatment initiation (erlotinib, gefitinib, afatinib, or osimertinib), ensuring toxicity was linked to the treatment. This approach captures toxicity events occurring closest in time to the start of therapy thus reflecting the variable as a single observation per patient at baseline. No formal split training/validation was created, but rules were manually crafted based on domain knowledge. Finally, extractions were validated against a manually labeled ground truth of 20 patients, with 70% accuracy ($n =$ 14 correctly labeled, 30% incorrect/missing). The algorithm was only applied to records matched to each patient's follow-up start date.

The rule-based NER algorithm was implemented with the 're' library available in Python version 3.[13]

Statistical analysis

Time variables, such as mTTD and mOS, were reported as medians with 95% confidence interval (CI) obtained with the Kaplan—Meier (KM) method.

Demographics and clinical-pathological characteristics were described using frequencies and percentages for categorical data, and medians with interquartile range (IQR) for continuous data. Group differences were tested with chi-Square and Wilcoxon rank sum tests for categorical and continuous variables, respectively.

Follow-up was defined from the first TKI administration (1L therapy) to last contact or death. TTD was defined from 1L start to treatment end for any reason (e.g. progression, toxicity, patient choice, or death), with patients censored if no discontinuation was observed by the cut-off date. OS was from 1L start to death, with censored cases for those alive at last follow-up or cut-off.

Median TTD and OS (mTTD and mOS) were reported with 95% CI using Kaplan—Meier (KM) estimates. Cox proportional hazard (PH) models identified survival prognostic factors, adjusting for confounders through hazard ratios (HRs). To avoid immortal time bias, time-dependent analysis, for OS estimation, was implemented by specifying the treatment as a longitudinal variable.[11,14] Moreover, clinical outcomes were presented as adjusted survival curves from the Cox model, accounting for explanatory variables.

PH assumptions were verified with Schoenfeld's residual test, and P values $<$ 0.05 were considered statistically significant.

Statistical analyses were conducted with R software (version 4.1.3)[15] using survival[16,17] and survminer[18] packages.

RESULTS

Patient baseline characteristics

A cohort of 190 patients treated with 1L EGFR-TKIs was enrolled, with 38% (n = 73) receiving osimertinib (defined as the 'osimertinib group' or 3rd-gen TKI) and 62% (n = 117) receiving 1st/2nd-gen TKIs (the 'non-osimertinib group'). Demographic and clinical-pathological characteristics are summarized in Table 1. Age, gender, ECOG PS, and smoking

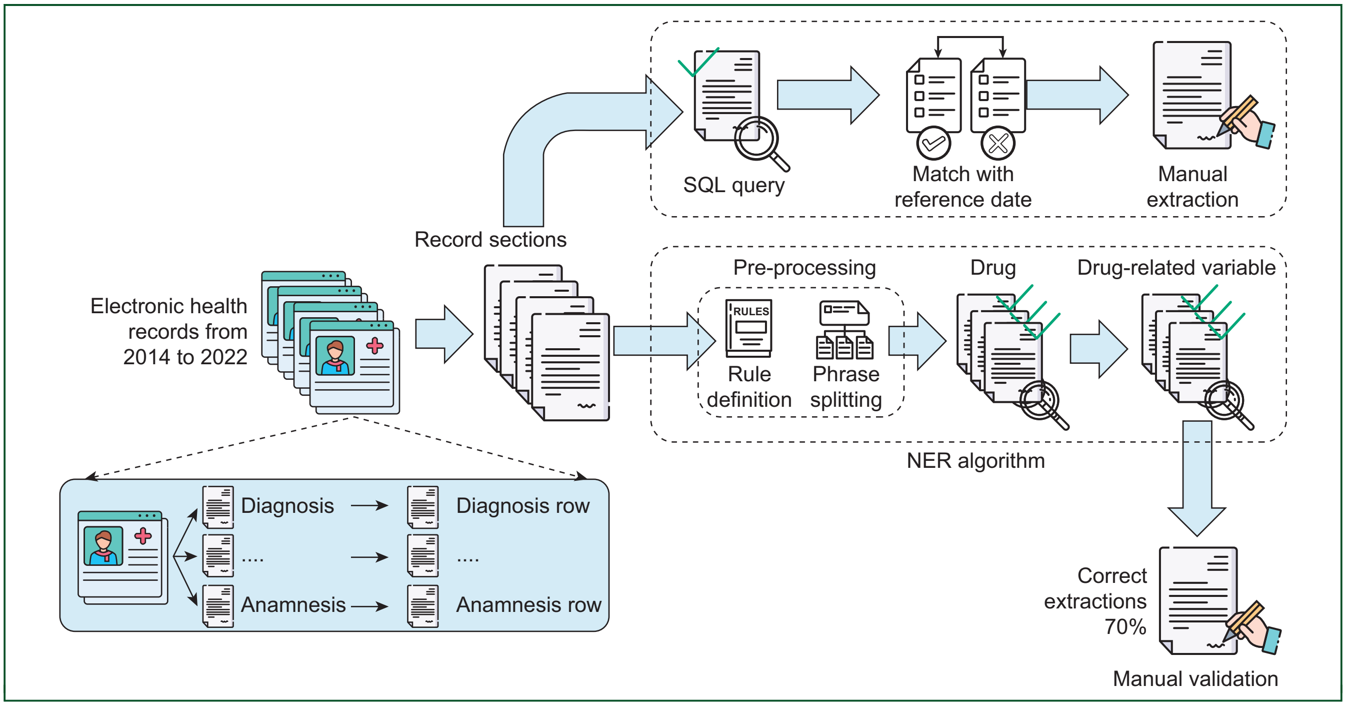


Figure 3. Data extraction. Clinical data extraction process using SQL and NLP. Each medical record contains multiple sections detailing the patient's clinical history, including oncological diagnosis and remote anamnesis. SQL and NLP are then applied to these sections as outlined in the figure.
NLP, natural language processing; SQL, Structured Query Language.

habits were similar between the two groups. NSCLC adenocarcinoma was the most common subtype, with PD-L1 <1% in most patients. The non-osimertinib group showed a higher number of metastatic sites ($P = 0.048$), including brain metastases ($P = 0.06$). Metastatic sites were assessed exclusively at baseline, representing the status at the initiation of treatment. Moreover, the non-osimertinib group received more subsequent chemotherapy lines ($P = 0.022$) compared with the osimertinib group.

No significant differences in toxicity related to TKI administration were found. The osimertinib group was also characterized by EGFR mutation type, OPDs (defined as limited metastatic areas progressing amenable to locoregional treatment together with an ongoing therapy continuation), and treatment response per RECIST criteria [RECIST criteria was applied only subsequently to data extraction, to provide an unbiased evaluation of the information extracted about disease progressions. In particular, progressions were defined as the appearance of one or more new lesions or an increase in pre-existing lesions (so that a change in therapy and/or locoregional treatment was needed)].[19]

Real-world treatment patterns

Within the osimertinib and non-osimertinib groups, 38.3% and 38.4% of patients received a single line of treatment before death (i.e. only 1L osimertinib or non-osimertinib before death without any other therapy change), respectively (see Supplementary Figure S2, available at https://doi.org/10.1016/j.esmorw.2024.100109).

The remaining 61.6% in the osimertinib group included 38.3% of patients continuing osimertinib and 23.2% moving to 2L therapy. In the non-osimertinib group, 2.5% stayed on 1L 1st/2nd-gen TKIs, while 58.9% moved to 2L. Among osimertinib patients receiving 2L ($n = 17$), all received CTx. In the non-osimertinib group's 2L ($n = 69$), 33.3% received osimertinib alone, 20.2% received osimertinib followed by third-line (3L) CTx, 30.4% received CTx only, and 15.9% received CTx followed by 3L osimertinib. Sequential therapy use was common; notably, 25.2% ($n = 48$) of all patients switched from 1L 1st/2nd-gen TKIs to 2L osimertinib, similar to cross-over in RCTs.

Real-world treatment efficacy

In this section, treatment effect analysis is represented as KM curves and multivariate Cox models.

Time to treatment discontinuation. At data cut-off, median follow-up was 18.42 months (IQR 11.96-23.27 months) for the osimertinib group and 24.09 months (IQR 10.59-43.50 months) for the non-osimertinib group. According to guidelines, additional EGFR-TKI lines are permitted for OPDs with possible local-regional treatments.

Among patients with treatment discontinuation ($n = 166$), 53.6% ($n = 89$) continued beyond progression, with 30.3% ($n = 27$) remaining on therapy for 3 months or more after progression, 70.3% ($n = 19$) of whom were in the osimertinib group. Conversely, 24% ($n = 13$) discontinued early (>3 months before progression).

mTTD was 15 months (95% CI 11.97-26.40 months) in the osimertinib group versus 10.1 months (95% CI 7.93-13.5 months) in the non-osimertinib group, with a HR of 0.58 (95% CI 0.41-0.81) (see Figure 4 A).

All covariates, including toxicity associated with treatment initiation, bone metastases, and brain metastases,

Table 1. Demographic features and clinical-pathological characteristics of the study population

Variable	*n*	Non-osimertinib, *n* = 117	Osimertinib, *n* = 73	*P* value
Age	190	68 (58-73)	63 (58-73)	>0.9
Gender	190			0.4
Female		73 (62)	50 (68)	
Male		44 (38)	23 (32)	
ECOG PS, *n* (%)	129			0.7
0		31 (38)	18 (38)	
1		35 (43)	20 (43)	
2		14 (17)	6 (13)	
3		2 (2)	2 (4)	
4		0 (0)	1 (2)	
Unknown		35	26	
Smoking habits	178			>0.9
Never		49 (46)	32 (44)	
Current		10 (10)	8 (12)	
Former Unknown		47 (44)	32 (44)	
		11	1	
Histology	190			0.5
Adenocarcinoma		113 (97%)	68 (93)	
Sarcomatoid carcinoma		1 (0.9)	1 (1.4)	
Squamous carcinoma		3 (2.1)	4 (5.6)	
Staging	190			0.6
III		2 (2)	2 (3)	
IV		115 (98)	71 (97)	
PD-L1	112			0.071
<1%		21 (42)	29 (47)	
1%-49%		13 (26)	24 (38)	
>50%		16 (32)	9 (15)	
Unknown		67	11	
EGFR mutation	71			—
del19		—	37 (52)	
ex18		—	5 (7)	
ex21		—	28 (40)	
Other		—	1 (1)	
Unknown		—	2	
Metastases sites	166			0.048
1		39 (40)	35 (52)	
2		25 (25)	20 (30)	
+3		35 (35)	12 (18)	
Unknown		18	6	
Brain metastases	166	39 (39)	17 (25)	0.061
Unknown		18	6	
Bone metastases	166	48 (48)	40 (60)	0.2
Unknown		18	6	
Chemotherapy	190	46 (39)	17 (23)	0.022
Radiotherapy	190	73 (62)	43 (59)	0.6
Toxicity	125	21 (30)	12 (21)	0.3
Unknown		48	17	
OPD	71			—
No		—	55 (77.5)	
Yes unknown		—	16 (22.5)	
		—	2	
RECIST response	67			—
PR		—	47 (70)	
PD		—	3 (4.5)	
SD		—	16 (24)	
CR		—	1 (1.4)	
Unknown		—	6	

Median (IQR) for continuous variables; *n* (%) for categorical variables.
CR, complete response; ECOG, Eastern Cooperative Oncology Group; EGFR, epidermal growth factor receptor; IQR, interquartile range; OPD, oligoprogressive disease; PD, progressive disease; PD-L1, programmed death-ligand 1; PR, partial response; PS, performance status; SD, stable disease.

were treated as baseline covariates in this analysis. No covariates were treated as time-varying in the TTD analysis.

Adjusted models showed longer mTTD for osimertinib (see Supplementary Figure S3, available at https://doi.org/10.1016/j.esmorw.2024.100109) in both the bone metastases group (13.04 versus 8.89 months) and brain metastases group (11.97 versus 7.84 months) (see Supplementary Figure S4, available at https://doi.org/10.1016/j.esmorw.2024.100109).

The protective effect of osimertinib on treatment discontinuation was also supported by the multivariate Cox model (see Supplementary Material, Section 4.1 and Supplementary Figure S4, available at https://doi.org/10.1016/j.esmorw.2024.100109), where radiotherapy emerged as a protective factor (HR 0.66, 95% CI 0.46-0.94), while bone (HR 1.45) and brain metastases (HR 1.51) were risk factors, adjusted for age and gender. No significant effect was observed for toxicity. The PH assumption for the TTD model was confirmed (Schoenfeld's test $P = 0.4$).

Overall survival. To evaluate the real-world impact of osimertinib on OS, osimertinib treatment was codified as a time-varying variable. This approach ensures that patients switching to osimertinib in later lines contribute to the risk set for osimertinib only after they start receiving it, avoiding immortal time bias. Therefore, this analysis reflects the clinical question whether osimertinib, administered at any point during the treatment course, might improve OS.

All covariates were considered at baseline except for treatment that was codified as a dichotomous time-varying variable, with osimertinib as the treatment of interest.

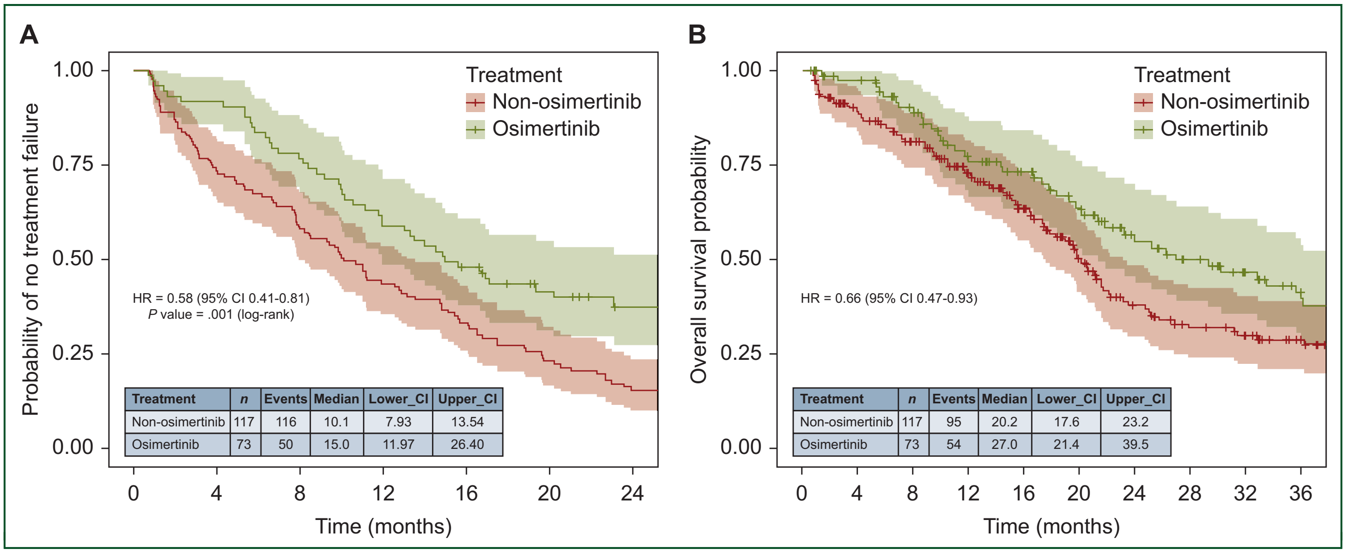
The mOS (see Figure 4B) for the osimertinib group was 27 months (95% CI 21.4-39.5 months) versus 20.2 months (95% CI 17.6-23.2 months) in the non-osimertinib group (HR 0.66, 95% CI 0.47-0.93).

Adjusted OS curves were also reported for the two risk groups accounting for bone and brain metastases (see Supplementary Figure S6, available at https://doi.org/10.1016/j.esmorw.2024.100109).

As for TTD, a multivariate time-dependent Cox model for OS was employed (see Supplementary Figure S7, available at https://doi.org/10.1016/j.esmorw.2024.100109). Osimertinib treatment remained a significant protective factor for OS (HR 0.56, 95% CI 0.39-0.82), even if adjusting for gender, age, and bone (HR 1.43) and brain metastases (HR 1.55) which are conversely significant risk factors for survival. However, toxicity did not result as a significant risk factor for patients' survival.

DISCUSSION

This study introduces a data analytics framework to integrate RWE into cancer research, focusing on a systematic approach to design, collection, and analysis of RWD.



Treatment	*n*	Events	Median	Lower_CI	Upper_CI
Non-osimertinib	117	116	10.1	7.93	13.54
Osimertinib	73	50	15.0	11.97	26.40

Treatment	*n*	Events	Median	Lower_CI	Upper_CI
Non-osimertinib	117	95	20.2	17.6	23.2
Osimertinib	73	54	27.0	21.4	39.5

Figure 4. Treatment effect analysis. Kaplan−Meier curves for (A) TTD and (B) OS of first-line treatments osimertinib and non-osimertinib. For median OS estimation, treatment is modeled as a time-dependent variable. No log-rank test is available for this model.
CI, confidence interval; HR, hazard ratio from the univariate Cox model.

Emphasis was placed on extracting and analyzing RWD from DWHs, with an application to a cohort of patients with aNSCLC with EGFR mutations. The main research questions were to analyze treatment patterns and outcomes of EGFR-TKIs, particularly examining osimertinib compared with first- and second-generation TKIs.

This work exploited a rule-based NER algorithm to accelerate data collection and improve the quality preventing erroneous and incomplete data. EHRs were queried via SQL and text-mining algorithms. In particular, a rule-based NER algorithm was used to retrieve information about drug toxicity.

A formal data quality assessment was carried out only for the NER algorithm, which achieved 70% accuracy on manually labeled texts. For other clinical variables, the SQL-based pipeline was semi-automatic, with a final manual extraction step. Automating extraction significantly accelerated data collection and minimized manual review in both experiments.

The cohort analyzed reflects the treatment landscape for EGFR mutation-positive patients over the past decade.

For patients with aNSCLC, the acquisition of the T790M mutation represents the condition to switch to osimertinib after prior 1L 1st/2nd-gen TKIs. According to recent literature, a percentage of 50% is expected to acquire this mutation after previous-generation TKIs.[20]

In this cohort, data revealed that 53.6% of patients initially receiving 1st/2nd-gen TKIs switched to osimertinib in 2L therapy or subsequent treatments (15.9%). Among patients in the osimertinib group, only 23.2% transitioned to subsequent therapies, with chemotherapy as the main option. These treatment patterns are coherent with those described by Ramalingam et al. in the FLAURA trial.[8]

For studying clinical outcomes, two endpoints, namely TTD and OS, are investigated with different statistical methodologies, including Cox regression model and time-dependent analysis. At this time, EGFR-TKIs are occasionally maintained beyond RECIST-defined progressions.

Many studies have demonstrated that carrying out EGFR-TKI maintenance, along with local radiotherapy, produces clinical benefits in patients with OPDs.[21-23] Therefore, TTD can be reported as a potential proxy of clinical outcomes for RWE studies, showing a high correlation with PFS found in clinical trials.[9] Moreover, outcomes such as TTD usually do not require a review-like process of clinical charts; thus a larger amount of data can be easily collected, increasing the study sample size and thus lowering study costs.[24] The importance of using TTD as a measure of real-world benefit was also stated in several systematic reviews.[9,25] For the estimation of TTD, a Cox analysis with time-fixed covariates is adopted.

Furthermore, the long follow-up of this study has also allowed the estimation of OS. In case of long-term effects estimated via RWD, immortal time bias can significantly affect the results. In this cohort, a significant percentage of patients in the 1L non-osimertinib group crossed over the osimertinib group, at the time of T790M mutation acquisition. However, unlike intention-to-treat analyses used in clinical trials, this study assessed the impact of osimertinib administered at any time during the treatment course. By treating osimertinib as a time-dependent variable in OS analysis, we correctly accounted for real-world treatment switching in later lines of therapy. In fact, with osimertinib being the exposition factor of interest, time-fixed analysis would have assigned to the unexposed patients (non-osimertinib patients who switched to osimertinib later) a spurious survival advantage as if they had been assigned to the exposed group (osimertinib) since study initiation. This bias that happens very frequently in cohort studies using EHRs was also studied in pharmaco-epidemiology literature.[14,26,27]

To obtain an unbiased estimation of mOS, this work adopted a time-dependent Cox model.[9,11] Specifically, the treatment constituted the only time-dependent covariate. All the other covariates were considered as baseline features, since no measurements at different time points were available. Finally, multivariate analyses and adjusted survival curves were also used to minimize confounding effects.

A significantly longer mTTD equal to 15 months was found for patients treated with osimertinib compared with 10 months for 1st/2nd-gen TKIs in line with other real-world studies.[28-30] The beneficial effect of osimertinib in terms of treatment discontinuation was also demonstrated by multivariate Cox analysis, in which the risk of treatment failure for osimertinib was reduced by half compared with other TKIs. Moreover, this model established the favorable role of osimertinib regardless of local radiotherapy. This is a result of relevant clinical interest for treatment choice in clinical practice.

Finally, TTD was analyzed in comparison to PFS, demonstrating that TTD represents a significant and practical endpoint in real-world studies. Specifically, for osimertinib, the analysis showed a median PFS of 13 months (Supplementary Material, Section 4.2 and Supplementary Figure S5, available at https://doi.org/10.1016/j.esmorw.2024.100109), with TTD aligning closely as an upper-bound measure. These results support the utility of TTD as a reliable and pragmatic outcome in real-world studies, offering a complementary perspective on treatment effectiveness. While TTD is not yet widely adopted, its advantages make it a valuable tool for RWE generation.[31-37]

Median OS estimated by time-dependent Cox model was 27 and 20.2 months for the osimertinib group and non-osimertinib groups, respectively. This result confirms the recent study proposed by Wells et al.[38] in which outcomes of patients not eligible for the FLAURA trial have been analyzed. Multivariate analysis proved that the risk of death for osimertinib was significantly reduced even when adjusting for gender, age, and bone and brain metastases which were conversely significant risk factors for survival.

The data analytics framework introduced in this study demonstrates scalability, making it applicable to a range of RWE studies beyond the oncology field. This approach aligns with ESMO-GROW guidelines for RWE reporting[11] and enables systematic monitoring of real-world drug efficacy and safety. By applying robust analytic methods, including NER and statistical modeling, this framework effectively integrates diverse data sources to inform clinical practice and health policy.

While the framework has shown promising results, there are some limitations. This is a retrospective study with a single-institution dataset which may limit generalizability of the results. Notably, this cohort included a significant proportion of frail patients with ECOG PS of 2-3, who would not typically meet clinical trial criteria, yet their inclusion offers valuable insight into real-world treatment responses.

Moreover, clinical variables were treated mainly as baseline covariates in the survival models. In particular, due to the design of the NER algorithm, toxicity was analyzed as a baseline covariate capturing adverse events only at the initiation of first-line treatment. While this approach provides insights into the impact of early toxicity on treatment outcomes, it does not account for the potential dynamic effects of toxicity over time.

A further limitation in the current framework regards the manual review to supplement the extraction of clinical events from electronic records such as treatment discontinuation. Further attempts could be made also to improve the methodology, especially in terms of statistical analysis and text-mining techniques. As the application of NER algorithms is challenging for health records due to the doctor's writing styles, different forms of medical terms, and ambiguity in abbreviations, alternative more advanced machine learning algorithms could be compared to improve the performance in terms of automatic clinical event retrieval, missing values, and percentage of wrong/incomplete extractions.

In conclusion, this data analytics framework highlights the potential of RWE for guiding treatment decisions and policy development in oncology. RWD-informed models, capable of processing large volumes of routinely collected health data, are essential for assessing treatment strategies aimed at enhancing life expectancy and reducing health care costs. Despite ongoing challenges in implementing RWE frameworks, such tools hold promise as practical decision support systems in health care.

DISCLOSURE

LM: Novartis, MSD: honoraria. Sanofi, Daiichi Sankyo: conference grants outside of the submitted work; LP: Novartis, Pfizer, Gilead: honoraria outside the submitted work; AS: Novartis, BMS, Immunocore: honoraria outside the submitted work; MB: Eli Lilly: conference grant outside the submitted work; MO: AstraZeneca, BMS, MSD: honoraria. AstraZeneca, BMS, MSD Advisory Role. Eli Lilly: Conference Grants outside the submitted work; TB: MSD, Sanofi, Pfizer, E Lilly: conference grants. MSD: honoraria outside the submitted work; SM: travel accommodation MSD, Sanofi, personal fees Italfarmaco AstraZeneca, outside the submitted work; CP: personal fees from BMS and MSD, outside the submitted work; FGMdB: reports a patent for PCT/IB2020/055956 pending and a patent for IT201900009954 pending; honoraria from, or consultant role for, Roche, EMD Serono, NMS Nerviano Medical Science, Sanofi, MSD, Novartis, Incyte, BMS, Menarini Healthcare Research & Pharmacoepidemiology, Merck Group, Pfizer, Servier, AMGEN, Incyte, outside the submitted work; GLR: consultant role for Roche, Novartis, BMS, MSD, AstraZeneca, Takeda, Amgen, Sanofi, Italfarmaco, Pfizer; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Roche, Novartis, BMS, MSD, AstraZeneca, Takeda, Amgen, Sanofi; support for attending meetings and/or travel from Roche, BMS, MSD; data safety monitoring board or advisory board for Roche, Novartis, BMS, MSD, AstraZeneca, Sanofi; has acted as principal investigator in sponsored clinical trials for Roche, Novartis, BMS, MSD, AstraZeneca, GSK, Amgen, Sanofi, outside the submitted work; AP: consulting or advisory role for BMS, AstraZeneca, Novartis; travel, accommodations, or other expenses paid or reimbursed by Roche, Italfarmaco; principal investigator of Spectrum Pharmaceuticals. personal fees from Roche, AstraZeneca and BMS, outside the submitted work. All other authors have declared no conflicts of interest.

ACKNOWLEDGEMENTS

We thank all the patients and the participating centers who accepted to actively participate in the Apollo 11 study with protocol number INT 128/22.

FUNDING

This work was supported by Fondazione IRCCS ‘Istituto Nazionale dei Tumori’, 5 per 1000 Ministry of Health funds under the project ‘DWH 2.0: maintenance, integration, harmonization, and diffusion’; and 5 per 1000 Ministry of University and Research funds and Fondazione IRCSS Istituto Nazionale dei Tumori, through its call for the Valorisation of Institutional Research Program BRI 2021 (no grant number). The research has also received non-conditioning support from AstraZeneca.

DATA SHARING

The datasets presented in this article are not readily available because of patients’ privacy protection. Requests to access the datasets should be directed to the corresponding author.

REFERENCES

1. Benedum CM, Sondhi A, Fidyk E, et al. Replication of real-world evidence in oncology using electronic health record data extracted by machine learning. *Cancers*. 2023;15(6):1853.
2. Friedman C. Towards a comprehensive medical language processing system: methods and issues. In: Proceedings of the AMIA Annual Fall Symposium. *American Medical Informatics Association*. 1997. p. 595.
3. Ferlay J, Colombet M, Soerjomataram I, et al. Cancer statistics for the year 2020: an overview. *Int J Cancer*. 2021;149(4):778-789.
4. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209-249.
5. Mitsudomi T, Morita S, Yatabe Y, et al. Cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *Lancet Oncol*. 2010;11:121-128.
6. Rosell R, Carcereny E, Gervais R, et al. Spanish Lung Cancer Group in collaboration with Groupe Fran,cais de Pneumo-Canc´erologie and Associazione Italiana Oncologia Toracica. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol*. 2012;13(3):239-246.
7. Soria J-C, Ohe Y, Vansteenkiste J, et al. Osimertinib in untreated EGFR-mutated advanced non—small-cell lung cancer. *N Engl J Med*. 2018;378(2):113-125.
8. Ramalingam SS, Vansteenkiste J, Planchard D, et al. Overall survival with osimertinib in untreated, EGFR-mutated advanced NSCLC. *N Engl J Med*. 2020;382(1):41-50.
9. Blumenthal G, Gong Y, Kehl K, et al. Analysis of time-to-treatment discontinuation of targeted therapy, immunotherapy, and chemotherapy in clinical trials of patients with non-small-cell lung cancer. *Ann Oncol*. 2019;30(5):830-838.
10. Gaitonde P, Chirikov V, Kelkar S, Liljas B. Considerations for the utility of real-world evidence beyond trial data in advanced NSCLC: the case of frontline tyrosine kinase inhibitors. *Cancer Manag Res*. 2022;14:3421-3435.
11. Castelo-Branco L, Pellat A, Martins-Branco D, et al. ESMO Guidance for Reporting Oncology real-World evidence (GROW). *Ann Oncol*. 2023;34(12):1097-1112.
12. Tyrer F, Bhaskaran K, Rutherford MJ. Immortal time bias for life-long conditions in retrospective observational studies using electronic health records. *BMC Med Res Methodol*. 2022;22(1):86.
13. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.
14. Giobbie-Hurder A, Gelber RD, Regan MM. Challenges of guarantee-time bias. *J Clin Oncol*. 2013;31(23):2963.
15. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2023. Available at https://www.R-project.org/. Accessed October 1, 2023.
16. Therneau TM. A Package for Survival Analysis in R. R package version 3. 5-5. 2023. Available at https://CRAN.R-project.org/package=survival. Accessed October 1, 2023.
17. Therneau TM, Grambsch PM. The cox model. In: *Modeling Survival Data: Extending the Cox Model*. Springer; 2000. p. 39-77.
18. Kassambara A, Kosinski M, Biecek P. Survminer: Drawing Survival Curves using ’’ggplot2’. R package version 0.4.9. 2021. Available at https://CRAN.R-project.org/package=survminer. Accessed October 1, 2023.
19. Schwartz LH, Liti`ere S, De Vries E, et al. RECIST 1.1.-Update and clarification: from the RECIST committee. *Eur J Cancer*. 2016;62:132-137.
20. Hsieh P-C, Wu Y-K, Huang C-Y, et al. Comparison of T790M acquisition after treatment with first- and second-generation tyrosine-kinase inhibitors: a systematic review and network meta-analysis. *Front Oncol*. 2022;12:869390.
21. Helena AY, Sima CS, Huang J, et al. Local therapy with continued EGFR tyrosine kinase inhibitor therapy as a treatment strategy in EGFR-mutant advanced lung cancers that have developed acquired resistance to EGFR tyrosine kinase inhibitors. *J Thorac Oncol*. 2013;8(3):346-351.
22. Schmid S, Klingbiel D, Aeppli S, et al. Patterns of progression on osimertinib in EGFR T790M positive NSCLC: a Swiss cohort study. *Lung Cancer*. 2019;130:149-155.
23. Hu C, Wu S, Deng R, et al. Radiotherapy with continued EGFR-TKIs for oligoprogressive disease in EGFR-mutated non-small cell lung cancer: a real-world study. *Cancer Med*. 2023;12(1):266-273.
24. Walker B, Boyd M, Aguilar K, et al. Comparisons of real-world time-to-event end points in oncology research. *JCO Clin Cancer Inform*. 2021;5:45-46.
25. Lasala R, Zovi A, Isgr`o V, Romagnoli A, Musicco F, Santoleri F. Time to treatment discontinuation in first-line non-small cell lung carcinoma: an overview. *Curr Med Res Opin*. 2023;39(12):1603-1612.
26. Suissa S. Immortal time bias in pharmacoepidemiology. *Am J Epidemiol*. 2008;167(4):492-499.
27. Hern´an MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med*. 2017;377(14):1391-1398.
28. Lorenzi M, Ferro A, Cecere F, et al. First-line osimertinib in patients with EGFR-mutant advanced non-small cell lung cancer: outcome and safety in the real world: FLOWER study. *Oncologist*. 2022;27(2):87-e115.
29. Griesinger F, Popat S, Okhuoya P, et al. 372P Global real-world (rw) study of patients (pts) with epidermal growth factor receptor (EGFR) mutated advanced non-small cell lung cancer (NSCLC) treated with first-line (1L) osimertinib: interim analysis of an rw pt registry in Germany. *Ann Oncol*. 2022;33:S1587.
30. Nieva J, Karia P, Okhuoya P, et al. 1344P A real-world (rw) observational study of long-term survival (LTS) and treatment patterns after first-line (1L) osimertinib in patients (pts) with epidermal growth factor receptor (EGFR) mutation-positive (m) advanced non-small cell lung cancer (NSCLC). *Ann Oncol*. 2023;34:S774.

31. Imamura F, Kimura M, Yano Y, et al. Real-world osimertinib for *EGFR* mutation-positive non-small-cell lung cancer with acquired T790M mutation. *Future Oncol*. 2020;16(21):1537-1547.
32. Peng D, Shan D, Dai C, et al. Real-world data on osimertinib in Chinese patients with pretreated, EGFR T790M mutation positive, advanced non-small cell lung cancer: a retrospective study. *Cancer Manag Res*. 2021;13:2033-2039.
33. Mu Y, Xing P, Hao X, Wang Y, Li J. Real-world data of osimertinib in patients with pretreated non-small cell lung cancer: a retrospective study. *Cancer Manag Res*. 2019;11:9243-9251.
34. Agulnik JS, Kasymjanova G, Pepe C, et al. Real-world pattern of treatment and clinical outcomes of EGFR-mutant non-small cell lung cancer in a single academic centre in Quebec. *Curr Oncol*. 2021;28(6):5179-5191.
35. Provencio M, Terrasa J, Garrido P, et al. Osimertinib in advanced EGFR-T790M mutation-positive non-small cell lung cancer patients treated within the Special Use Medication Program in Spain: OSIREX-Spanish Lung Cancer Group. *BMC Cancer*. 2021;21(1):230.
36. Watanabe K, Yoh K, Hosomi Y, et al. Efficacy and safety of first-line osimertinib treatment and postprogression patterns of care in patients with epidermal growth factor receptor activating mutation-positive advanced non-small cell lung cancer (Reiwa study): study protocol of a multicentre, real-world observational study. *BMJ Open*. 2022;12(1):e046451.
37. Cao Y, Qiu X, Xiao G, Hu H, Lin T. Effectiveness and safety of osimertinib in patients with metastatic EGFR T790M-positive NSCLC: an observational realworld study. *PLoS One*. 2019;14(8):e0221575.
38. Wells JC, Mullin MM, Ho C, et al. Outcomes of patients with advanced epithelial growth factor receptor mutant lung cancer treated with first-line osimertinib who would not have met the eligibility criteria for the FLAURA clinical trial. *Lung Cancer*. 2024;190:107529.