Characterizing Metastatic Non-Small Cell Lung Cancer Patients Across a Federated Network of Observational Data

# Table of Contents

[Table of Contents 2](#_Toc186548091)

[Abbreviations 3](#_Toc186548092)

[Responsible Parties 4](#_Toc186548093)

[Investigators 4](#_Toc186548094)

[Abstract 5](#_Toc186548095)

[Background and significance 5](#_Toc186548096)

[Study aims 5](#_Toc186548097)

[Study description 5](#_Toc186548098)

[Population 5](#_Toc186548099)

[Outcomes 5](#_Toc186548100)

[Design 6](#_Toc186548101)

[Amendments and Updates 7](#_Toc186548102)

[Background and Rationale 8](#_Toc186548103)

[Research Question and Objectives 8](#_Toc186548104)

[Research Methods 9](#_Toc186548105)

[Study Design 9](#_Toc186548106)

[Study Setting and Data Sources 9](#_Toc186548107)

[Study Period 10](#_Toc186548108)

[Study Population 10](#_Toc186548109)

[Target Cohorts 10](#_Toc186548110)

[NSCLC 10](#_Toc186548111)

[Metastatic NSCLC (mNSCLC) 11](#_Toc186548112)

[Patients with systemic treatment and mNSCLC 11](#_Toc186548113)

[Follow up 12](#_Toc186548114)

[Variables 12](#_Toc186548115)

[Treatment-related information 12](#_Toc186548116)

[Outcomes 13](#_Toc186548117)

[Covariates 13](#_Toc186548118)

[Stratifications 14](#_Toc186548119)

[Study size 15](#_Toc186548120)

[Data Management 15](#_Toc186548121)

[Analysis 15](#_Toc186548122)

[General 15](#_Toc186548123)

[Baseline Demographics and Comorbidities 15](#_Toc186548124)

[Treatment Patterns 16](#_Toc186548125)

[Time to Event Outcomes 16](#_Toc186548126)

[Quality Assurance 16](#_Toc186548127)

[General Database Quality Control 16](#_Toc186548128)

[Study-specific Quality Control 17](#_Toc186548129)

[Software Quality Control 17](#_Toc186548130)

[Strength and Limitations 17](#_Toc186548131)

[Protection of Human Subjects 17](#_Toc186548132)

[Management and Reporting of Adverse Events/Reactions 18](#_Toc186548133)

[Dissemination and Communication of Study Results 18](#_Toc186548134)

[References 18](#_Toc186548135)

# Abbreviations

|  |  |
| --- | --- |
| BMI | Body Mass Index |
| CCI | Charlson Comorbidity Index |
| DQD | Data Quality Dashboard |
| DT | Diagnosis to Treatment Interval |
| ECOG | Eastern Cooperative Oncology Group |
| EMA | European Medicines Agency |
| ENCePP | European Network of Centres for Pharmacoepidemiology and Pharmacovigilance |
| ESMO | European Society for Medical Oncology |
| ICH | International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use |
| IQR | Interquartile Range |
| ICI | Immune Checkpoint Inhibitors |
| KPS | Karnofsky Performance Status |
| LoT | Line of Treatment |
| mNSCLC | Metastatic Non-Small Cell Lung Cancer |
| NSCLC | Non-Small Cell Lung Cancer |
| OHDSI | Observational Health Data Sciences and Informatics |
| OS | Overall Survival |
| PD1 | Programmed Cell Death Protein 1 |
| PD-L1 | Programmed Cell Death Ligand 1 |
| PS | Performance Status |
| mAb | Monoclonal Antibody |
| RWD | Real World Data |
| SD | Standard Deviation |
| TFI | Treatment-Free Interval |
| TTNT | Time to Next Treatment |
| TTD | Time to Treatment Discontinuation |
| TKI | Tyrosine Kinase Inhibitors |
| VEGF-A | Vascular Endothelial Growth Factor A |

# Responsible Parties

## Investigators

|  |  |
| --- | --- |
| Investigator | Affiliation |
| Kimmo Porkka | Helsinki University Hospital Comprehensive Cancer Center |
| Paula Kauppi | Helsinki University Hospital Comprehensive Cancer Center |
| Eric Fey | Helsinki University Hospital Comprehensive Cancer Center |
| Valtteri Nieminen | Helsinki University Hospital Comprehensive Cancer Center |
| Alexey Ryzhenkov | Helsinki University Hospital Comprehensive Cancer Center |
| Åslaug Helland | Oslo University Hospital Comprehensive Cancer Centre |
| Elin Hallan Naderi | Oslo University Hospital Comprehensive Cancer Centre |
| Åsa Öjlert | Oslo University Hospital Comprehensive Cancer Centre |
| Elisabeth Ross | Oslo University Hospital Comprehensive Cancer Centre |
| Olivier Bouissou | Oslo University Hospital Comprehensive Cancer Centre |
| Annelies Verbiest | Antwerp University Hospital, University of Antwerp |
| Zarah Van Schoor | Antwerp University Hospital |
| Hans Prenen | Antwerp University Hospital, University of Antwerp |
| Elyne Scheurwegs | Antwerp University Hospital |
| Annelies Janssens | Antwerp University Hospital, University of Antwerp |
| Stelios Theophanous | Leeds Teaching Hospitals NHS Trust |
| Geoff Hall | Leeds Teaching Hospitals NHS Trust |
| Elad Sharon | Dana Farber Cancer Institute |
| John Methot | Dana Farber Cancer Institute |
| Roshanthi Weerasinghe | Providence St. Joseph Health |
| Andrew Nute | Providence St. Joseph Health |
| Martin Wermke | University Hospital Dresden, National Cancer Center |
| Koch, Martin | University Hospital Dresden, National Cancer Center |
| Yuan Peng | University Hospital Dresden, Center for Medical Informatics |
| Ines Reinecke | University Hospital Dresden, Center for Medical Informatics |
| Katja Hoffmann | University Hospital Dresden, Center for Medical Informatics |
| Jasmin Carus | University Medical Center Hamburg-Eppendorf (UKE) |
| Michael Franz | Berlin Institute of Health @ Charité |
| Henry Morgan Stewart | Nemesis Health, OHDSI Oncology WG |
| Christian Reich | Nemesis Health, OHDSI Oncology WG |
| Asieh Golozar | Nemesis Health, OHDSI Oncology WG |

# Abstract

### Background and significance

Metastatic non-small cell lung cancer (mNSCLC) represents a significant global health burden, characterized by poor prognoses and high mortality rates. The introduction of immune checkpoint inhibitors (ICIs) has revolutionized treatment for mNSCLC, offering improved survival outcomes in clinical trials for select patient groups. However, evidence of the uptake, treatment and effectiveness patterns of ICIs in real-world settings, particularly across diverse healthcare systems and populations, remains scarce. Real-world data (RWD) is vital to complement clinical trial findings, yet barriers such as data standardization, especially across borders, as well as privacy concerns hinder comprehensive global assessments. Federated data analysis frameworks, such as those facilitated by the Observational Health Data Sciences and Informatics (OHDSI) initiative addresses these challenges by enabling standardized, large-scale analyses while maintaining patient privacy.

### Study aims

This study seeks to evaluate the real-world treatment landscape of mNSCLC patients across a global network of observational healthcare data sources, with a particular focus on the introduction and impact of ICIs. Key objectives include:

* Characterizing demographics and clinical characteristics of mNSCLC patients.
* Describing treatment pathways and clinical outcomes, including time to treatment discontinuation, time to next line of treatment and overall survival.
* Investigating geographic and temporal trends in ICI uptake and the resulting shifts in treatment outcomes.

### Study description

This retrospective cohort study uses a federated network of observational healthcare data, standardized to the OMOP CDM.

### Population

#### Adult patients with mNSCLC patients (≥18 years) diagnosed from January 1, 2015, to the most recent available data. Specific cohorts include patients with NSCLC:

* Transitioning to metastatic disease.
* Initiating systemic antineoplastic therapies for metastatic disease.

Subgroups are defined by demographic and clinical factors, including programmed cell death ligand 1 (PD-L1) status, tumor histology, comorbidities and age.

### Outcomes

Key outcomes are time to treatment discontinuation, to next line of treatment and overall survival.

### Design

This study is an observational retrospective cohort study. It generates robust real-world evidence base on treatment patterns and outcomes of mNSCLC patients, particularly in the era of ICI therapies. Findings will inform clinical decision-making, support regulatory and policy development and highlight disparities in global treatment practices. The federated model ensures secure, reproducible and scalable insights, setting a precedent for future collaborative oncology research.

# Amendments and Updates

# Background and Rationale

Non-small cell lung cancer (NSCLC), particularly in its metastatic form (mNSCLC), remains a significant global health challenge. Over the past decade, advancements in our understanding of NSCLC’s molecular mechanisms have led to the development of novel therapies. Among these, immune checkpoint inhibitors (ICIs) have emerged as a groundbreaking treatment modality, demonstrating improved overall survival rates for certain patient subgroups (1-3).

Despite the promise shown by ICIs in clinical trials, there remains a critical gap in our understanding of their performance in real-world settings across different countries and healthcare systems. This knowledge gap is particularly pronounced when considering diverse patient populations, varying healthcare policies and differing medical practices. Differences in regulatory approvals, reimbursement strategies and treatment guidelines affect ICI availability and uptake across regions (4-7). Furthermore, variations in treatment sequencing, combination therapies and management of adverse events can impact overall outcomes (7, 8).

Current evidence from clinical trials, while crucial, may not fully represent the heterogeneity of real-world patient populations or account for the complexities of diverse healthcare systems. There is a lack of understanding of how ICI availability, uptake and effectiveness vary across different regions and healthcare settings. Real-world data are increasingly recognized as essential for complementing clinical trial results, informing treatment decisions and guiding regulatory policies (9). However, collecting and analyzing such data on a global scale presents significant challenges, including availability of the relevant data, data privacy concerns, varying or lack of data standards and the need for large-scale collaboration.

To address these challenges and gain a comprehensive, global perspective on the real-world impact of new therapies in mNSCLC, there is a pressing need for large-scale international collaborative efforts. Federated data analysis approaches, such as those employed by the OHDSI research network, offer a powerful solution to this problem. By utilizing common data models like the OMOP CDM, these approaches enable the harmonization and analysis of data from diverse sources while maintaining patient privacy and adhering to local data governance regulations (10, 11).

This study aims to bridge the current gaps in our understanding of the current treatment landscape of mNSCLC across different geographies. The findings from this study have the potential of valuable insights to inform treatment decisions, guide regulatory policies and improve patient outcomes on a global scale.

# Research Question and Objectives

The overarching aim of this study is to characterize patients with metastatic NSCLC and to assess the shift in treatment patterns and outcomes in patients with mNSCLC after the introduction of ICI.

1. To describe demographics and clinical characteristics of patients with NSCLC, mNSCLC and mNSCLC who initiated systemic antineoplastic treatments.
2. To characterize treatment patterns and outcomes of patients with mNSCLC across an international network of observational data. Specifically, the following outcomes are described for all lines of therapy (LoT), up to a maximum of three lines, depending on data availability:
   * Distribution of treatment regimens per LoT
   * Treatment flow across LoT (treatment pathways)
   * Diagnosis of metastatic disease to first treatment time interval
   * Treatment free intervals between LoTs
   * Time to treatment discontinuation
   * Time to next treatment
3. To estimate time between the first encounter with a diagnosis of NSCLC and progression to metastatic disease.
4. To estimate overall survival (OS) of patients with mNSCLC and the subset of mNSCLC patients who initiated systemic anti-neoplastic treatment by any of the various treatment regimens (see below).
5. To quantify the uptake of ICI treatment across health care setting and geography.
6. To explore the shift in treatment patterns and outcomes with the introduction of ICI.

All objectives are assessed in the entire populations and for the subgroups specified in the protocol.

# Research Methods

## Study Design

This is a retrospective cohort study of patients diagnosed with NSCLC across a network of observational healthcare databases, all standardized to OMOP CDM.

Databases are owned, managed and standardized locally by the different isntitutions (Data Partners). The study code is developed for all Data Partners to execute locally and the results are aggregated and analyzed.

## Study Setting and Data Sources

The selection of Data Partners for this study is based on their availability of patients with mNSCLC as well as cancer treatments and date of death. All databases are mapped to the OMOP CDM ([CDM Specification GitHub link](https://github.com/OHDSI/CommonDataModel)). A description of the participating data sources is provided in Table 1.

**Table 1:** Databases participating in this study contingent of passing the quality test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data source** | **Country** | **Population size** | **Data capture process and short description** | **Data capture timeframe** |
| Helsinki University Hospital | Finland | ~2.000.000 | Structural data pulled from operational EHR systems. Covers secondary and tertiary care. | 2012-present |
| FinOMOP | Finland |  |  |  |
| UZA | Belgium | >2.100.000 | Routinely collected structured data of patients visiting UZA tertiary hospital for in- or outpatient care | 2006-present (depending on data source) |
| Oslo University Hospital Comprehensive Cancer Centre | Norway |  |  |  |
| Leeds Teaching Hospitals NHS Trust | United Kingdom | 85,143 | Data has been sourced directly from the hospital's EHR system. The LTHT OMOP database includes data from all patients diagnosed with cancer at LTHT from 2010 onwards. | 2010-August 2024 |
| Dana Farber Cancer Institute | United States | 359,000 | In- and outpatients EHR linked with cancer registry | 2015-present |
| Providence | United States | 16,500,000 | In- and outpatients EHR linked with cancer registry | 2015-present |
| Dresden | Germany | ~ 80,000 | Data is sourced form EHR and on site tumor documentation system from all cancer patients in in- and outpatient care. | 2015-present |
| University Medical Center Hamburg-Eppendorf | Germany | 80,000 | Cancer Registry patients  linked with EHR | 2010-present |
| Charité | Germany | 215,000 | inpatient EHR |  |

## Study Period

The study period starts from 1 January 2015 and ends at the latest available date for all data sources. The identification period is from the date of database inception to six months prior to the latest available data in each database. This six-month buffer allows for a potential of 6 months of follow-up data for the last person included in the study.

## Study Population

All patients in a data source meeting cohort inclusion criteria are considered subjects. Broadly, these cohorts consist of adult patients with a diagnosis code for NSCLC who have at least 365 days of prior observation (unless defined otherwise, see below). The specific definitions for cohorts are provided in the following sections.

## Target Cohorts

### NSCLC

There are two cohorts to allow participation of general healthcare institutions with longitudinal patient data capture to specialized cancer centers, which typically only see the patient during diagnosis and treatment decision making/induction.

**Cohort 1A:** Adult patients (³18 years) with a diagnosis of NSCLC with at least 365 days of a prior observation period are identified. The index date is set at the date of the first NSCLC diagnosis record. Patients with a diagnosis of any other malignancy, except non-melanoma skin cancer, or with any NSCLC-related treatment (see below) prior to the index date are excluded.

A diagram of a patient's diagnosis

Description automatically generated

**Figure 1.** Schematic representation of cohort 1A

**Cohort 1B:** Same adult patients (³18 years) with a diagnosis of NSCLC, but with no prior observation period requirement. The index date is also set at the date of the first NSCLC diagnosis record. Patients with a diagnosis of any other malignancy, except non-melanoma skin cancer, or with any NSCLC-related treatment (see below) prior to the index date are excluded.

### Metastatic NSCLC (mNSCLC)

Metastases are a hallmark of malignant diseases and represent their natural progression. However, their detection, and therefore their capture in databases, may not follow this sequence. Instead, the metastatic spread might be detected prior to the originating primary malignant disease. Therefore, a cohort definition for mNSCLC therefore must allow for some timing flexibility.

**Cohort 2.** Adult patients (³18 years) with metastatic disease evidenced either through a metastasis or stage IV disease record with at least 365 days of prior observation are identified. The index date is set to the metastatic disease or stage IV record, whichever comes earlier. All patients must have at least one diagnosis record of NSCLC within a period of 180 days before to 30 days after the index date. In addition, patients with the diagnosis of any other malignancy, except non-melanoma skin cancer, any time prior to and up to 30 days after index date are excluded. Figure 2 depicts the schematic representation of the definition for cohort 2.

A screenshot of a medical test

Description automatically generated

**Figure 2.** Schematic representation of cohort 2.

### Patients with systemic treatment and mNSCLC

**Cohort 3:** Adult patients (³18 years) initiated mNSCLC specific systemic antineoplastic treatment regimen with at least 365 days of prior observation are identified. The index date is set at the date of initiation of treatment regimen for mNSCLC. To be included in the cohort, patients must have their first evidence of metastatic disease, either a record of metastasis or stage IV disease, within the period from 30 days prior to regimen start date (index) and before the regimen end date. In addition, patients must have a diagnosis of NSCLC 180 days before to 30 days after the index date. Patients are excluded if they have a diagnosis of other malignancies, except for non-melanoma skin cancers, prior to and up to 30 days after the index date. Figure 3 depicts the schematic representation of the treatment mNSCLC cohort definition.

A screenshot of a medical test

Description automatically generated

**Figure 2.** Schematic representation of cohort 3.

NSCLC treatment regimens can be one of the following groups:

* EGFR tyrosine kinase inhibitors (TKI) [Erlotinib, Gefitinib, Afatinib, Dacomitinib, Osimertinib]
* Other TKIs [Crizotinib, Ceritinib, Brigatinib, Alectinib, Lorlatinib, Entrectinib, Capmatinib, Tepotinib, Selpercatinib, Pralsetinib, Vandetanib, Cabozantinib, Lenvatinib]
* Immune checkpoint inhibitors (anti-PD1/L1, anti-CTLA-4 or both)
* Immune checkpoint inhibitors (anti-PD1/L1) and platinum doublet chemotherapy with or without anti-VEGF monoclonal antibody (mAb)
* Dual immune checkpoint inhibitors (anti-PD1 and anti-CTLA-4) and platinum doublet chemotherapy
* Platinum doublet chemotherapy with or without anti-VEGF mAb
* Single agent chemotherapy with or without anti-VEGF mAb [Pemetrexed with or without Bevacizumab, Docetaxel with or without Ramucirumab]

Separate cohorts will be developed for each group.

## Follow up

Patients are followed until death, diagnosis of another malignancy except non-melanoma skin cancers or end of study period, whichever occurs first.

## Variables

### Treatment-related information

mNSCLC regimens in each database are inferred by the ARTEMIS (Advanced Regimen deTection EMploying Temporal Smith-Waterman) package, an advanced regimen detection algorithm developed by the OHDSI Oncology Workgroup (12). ARTEMIS enables the systematic identification of oncology treatment regimens, defined through HemOnc, from real-world data (RWD) standardized to the OMOP CDM (13, 14). Using the Temporal Smith-Waterman (TSW) algorithm, ARTEMIS identifies treatment regimens as temporal sequences of drug exposure events within longitudinal health records, enabling precise stratification of patients based on their chemotherapy protocols rather than individual drugs. This approach supports large-scale, reproducible research by addressing challenges in oncology data, such as variability in coding practices and dynamic treatment pathways, ultimately facilitating the generation of reliable evidence for oncological research and real-world applications.

Each identified treatment regimen will be considered a distinct LoT. The date a patient discontinues a LoT is considered the end date of this LoT. Discontinuation is defined as starting a different systemic anti-neoplastic regimen, having a gap following the last administration of more than 120 days with no systemic anti-neoplastic therapy, or having a date of death while on the regimen, whichever occurs first.

### Outcomes

The following outcomes are assessed in the study:

* Time to Progression to Metastatic Disease: the number of days from date of first visit with a NSCLC diagnosis to first metastasis.
* Diagnosis to Treatment Interval (DT): the number of days from date of first encounter with a diagnosis of mNSCLC to the initiation of the first LoT.
* Treatment-free Interval (TFI): Time from discontinuation of each LoT to initiation of the subsequent LoT, or date of death occurring before any subsequent LoT.
* Time to Next Treatment (TTNT): Time from the initiation of each LoT to the date the patient starts the next LoT or to the date of death if death occurs prior to having another systemic anti-neoplastic treatment regimen.
* Time to Treatment Discontinuation (TTD): Length of time from the initiation of each LoT to the date of its discontinuation, or date of death occurring treatment discontinuation.
* Overall Survival (OS): Length of time from the index date for each cohort to the date of death, or loss to follow up or end of the study. Patients are censored at last recorded clinical activity within the database or end of follow up.

### Covariates

The following characteristics are determined **prior to or at index date**.

#### Demographics

* Age
  + age groups (≤65 and >65years old)
  + age described as a continuous variable
* Biological sex
* Smoking status
* Weight, height, body mass index (BMI)
* Performance status (PS), e.g., Eastern Cooperative Oncology Group (ECOG) PS or Karnofsky PS (KPS) at index. In case of KPS, it will be converted to ECOG PS using the following logic: **KPS of 100–90 corresponds to ECOG 0, KPS of 80–70 corresponds to ECOG 1, KPS of 60–50 corresponds to ECOG 2, and KPS ≤40 corresponds to ECOG ≥3.** PS will then be categorized into three groups: ECOG=0, ECOG =1 and ECOG ≥2

#### Baseline comorbidities

* Concept-based
  + Condition groups (SNOMED + descendants), >=1 occurrence during the interval
  + Drug era groups (ATC/RxNorm + descendants), >=1 day during the interval which overlaps with at least 1 drug era
* Cohort based
  + Charlson comorbidity index (CCI)
  + Type 2 diabetes mellitus
  + Hypertension
  + Hyperlipidemia
  + Cardiovascular disease
  + Stroke
  + Venous thromboembolic events (pulmonary embolism and deep vein thrombosis)
  + Chronic obstructive pulmonary disease
  + Parenchymal lung disease (pulmonary fibrosis, sarcoidosis, pneumoconiosis)
  + Liver disease
  + Renal disease
  + Dementia
  + Autoimmune disease
  + Histology: squamous and non-squamous
  + Presence of programmed cell death 1 ligand 1 (PD-L1): <1%, 1%-49% and ³50%
* Others
  + Index year

## Stratifications

Each target cohort is analyzed in full as well as stratified by the following pre-index characteristics:

* Liver metastasis
* Bone metastasis
* Index year
* Age groups: ≤65 and >65
* Sex
* Histology: squamous and non-squamous

# Study size

This study uses routinely collected data, all patients meeting the eligibility criteria above are included. No formal sample size and power calculation is conducted.

# Data Management

Source data at the participating institutions are converted to the OMOP CDM through an Extract, Transform, Load (ETL) process. It includes source code to standardized concept mapping while maintaining quality and context of information (15). The data are refreshed periodically responding to updates by the data source. There are six key standardized domains: condition, drug, measurement, procedure, visit and observation. Data Partners are responsible for their own validation to ensure that the OMOP database instance matches the information provided in the original source. Each Partner is responsible for following their own local data permits, processing and disclosure publication regulations and standard operating procedures.

This study follows relevant ENCePP guidelines and International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) guidelines for data management (ENCePP) (5). Standardization of the databases to the OMOP CDM enables the use of standardized analytics and tools across the network since the structure of the data and the terminology system is harmonized. The OMOP CDM is developed and maintained by OHDSI and is described in detail on the wiki page of the CDM ([CDM Specification GitHub link](https://github.com/OHDSI/CommonDataModel)) and in The [Book of OHDSI](http://book.ohdsi.org) (16).

# Analysis

## General

Data partners execute the analytics R package against their OMOP CDM instance and review and approve the aggregated results before sharing them with the study team. No patient level data are shared.

Data quality is evaluated using data quality checks described in the next section. A diagnostic package, built on the OHDSI Cohort Diagnostics library ([CohortDiagnostics GitHub link](https://github.com/OHDSI/CohortDiagnostics)), is used to assess fitness of use and evaluate measurement error in the phenotype development and evaluation process. The full study package is only executed on databases that pass data quality assessment and cohort diagnostics.

All analyses are reported by database, overall and stratified by age, sex and index year, unless the resulting patient size falls below a given threshold as set by the Data Partner (minimal cell size).

## Baseline Demographics and Comorbidities

Demographic and clinical characteristics are obtained for all target cohorts and the specified strata. For categorical variables, frequencies and percentages are drawn for each level. Continuous variables are summarized as mean (standard deviation (SD)), minimum, maximum, median and interquartile range (IQR) within each database. Standardized mean differences (SMD) are calculated when comparing characteristics of different study cohorts.

## Treatment Patterns

For mNSCLC and treatment-initiated mNSCLC cohorts, the number and percentage of patients receiving each treatment regimen in each LoT are described. Distribution of treatment regimens by treatment groups are summarized for each LoT. The number and proportion of mNSCLC patients who did not receive any treatment are described for the mNSCLC cohort. Treatment flow from each LoT to the other are presented in Sankey plots.

## Time to Event Outcomes

Time to progression to metastatic disease is reported as median (IQR) for the NSCLC cohort. DTI is reported as median (IQR) for the mNSCLC cohort. TFI, TTD, TTNT are evaluated descriptively by Kaplan-Meier curves along with two-sided 95% confidence interval (CI) for the median time to event estimates for the treatment-initiated mNSCLC cohort. OS is estimated for all cohorts using the Kaplan-Meier method and results are reported as plots of the estimated survival curves as well as the overall and median (95% CI) estimated probability of survival at years 1, 2 and 3.

Pending data availability, results are further stratified by index year to evaluate the uptake of ICI for the treatment of mNSCLC and the shift in treatment patterns and outcomes of patients with mNSCLC with the introduction of ICI.

# Quality Assurance

## General Database Quality Control

A number of open-source quality control mechanisms for the OMOP CDM have been developed (17). In particular, the Achilles tool, which systematically characterizes the data and presents it in a dashboard format for inspection. The generated data characteristics such as age distribution, condition prevalence per year, data density, measurement value distribution can be compared against expectations for the data. Achilles tool is used to systematically characterize the data. The generated data characteristics such as age distribution, condition prevalence per year, data density, measurement value distribution can be compared against expectations for the data. Additionally, the data quality dashboard (DQD) ([DataQualityDashboard GitHub link](https://github.com/OHDSI/DataQualityDashboard)). This tool provides a number of checks relating to the conformance, completeness and plausibility of the mapped data. Conformance focuses on checks that describe the compliance of the representation of data against internal or external formatting, relational, or computational definitions, completeness in the sense of data quality is solely focused on quantifying missingness or absence of data, while plausibility seeks to determine the believability or truthfulness of data values. Each of these categories has one or more subcategories and are evaluated in two contexts: validation and verification. Validation relates to how well data align with external benchmarks with expectations derived from known true standards, while verification relates to how well data conform to local knowledge, metadata descriptions and system assumptions.

To ensure each participating database has the relevant oncology specific data, each participating Data Partner executes an oncology specific quality control query (available here: [Oncology Data Quality Assessment](https://ohdsiorg.sharepoint.com/:f:/s/Workgroup-Oncology/EquH2_GVLPVFloRTLOZmbIYBtWfKo9cEMDZk7_Mj2iZ68w?e=i1z6vn)), testing conformance with domain, standard concept and choice of vocabulary rules as well as mapping validity.

## Study-specific Quality Control

The OHDSI CohortDiagnostics package ([CohortDiagnostics GitHub link](https://github.com/OHDSI/CohortDiagnostics)) is used to assess the relevance of each participating database. This package evaluates phenotype algorithms within OMOP CDM databases, offering a standardized set of analytics to understand patient capture, including data generation processes. It provides detailed insights into cohort characteristics, record counts and potential index event misclassification. By offering a consistent methodology for evaluating cohort definitions and phenotype algorithms across diverse observational databases, CohortDiagnostics ensures reproducibility and comparability in observational research.

## Software Quality Control

The analytic package for this study is developed using open-source software from OHDSI under the HADES analytical suite. All HADES packages follow developer guidelines including unit testing to ensure software validity. The study package is tested and developed on HUS data. Any bugs or issues are documented on the Github repository. Github is used for version control to provide an audit trail of the study package and ensure that changes properly address reported issues.

# Strength and Limitations

The common data model and standardized vocabularies ensures interoperability and portability of phenotypes utilized in this analysis. The federated study model ensures there is no movement of patient-level data from institutions participating in this analysis. This is critically important for the protection of patient privacy in the secondary use of routinely collected patient data. Data partners remain in control of the analysis run on these data and conduct their own validation processes to evaluate case reports against public health reporting.

This study is carried out using data recorded in a collection of EHR and reimbursement claims systems. Misclassification may occur in identification of patients, outcomes and covariates. As with any healthcare database used for secondary data analysis, the patient records can be incomplete in many respects and may have erroneous entries. Events occurring prior to the observation period (enrollment, beginning of care provision to the patient) such as diagnosis of metastatic disease, site of metastasis, chemotherapy regimens or baseline covariates within the database may not be available or may be recorded with the wrong timing. Data on treatment provided outside each participating institution at referring institutions are not included. Finally, comorbidities potentially impacting treatment at the Data Partner but that were themselves treated outside of the oncology setting might be under-reported.

# Protection of Human Subjects

This study is conducted in accordance with the International Society of Pharmacoepidemiology (ISPE) Guidelines for Good Pharmacoepidemiology Practices (GPP)and EMA, European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology (18, 19). The use of the OMOP CDM and OHDSI tools enable the federated analysis of these different databases without accessing patient-level data outside the participating institutions. The study is conducted in compliance with all applicable data protection, security and privacy laws, rules and regulations with respect to the collection, production, use, processing, storage, transfer, modification, deletion and disclosure of any information related to this study. Each data partner is required to provide a statement about IRB approval or exemption to participate.

Confidentiality of patient records is maintained at all times. Data partners remain in full control of executing the analysis and packaging results. There are no transmission of patient-level data at any time during these analyses. Only aggregate statistics are shared. Study packages contain minimum cell count parameters to obscure any cells which fall below allowable reportable limits.

# Management and Reporting of Adverse Events/Reactions

According to the new guidelines for good pharmacovigilance practice (EMA/813938/2011 Rev 3), there is no requirement for expedited reporting of adverse drug reactions from studies with secondary use of data (19).

# Dissemination and Communication of Study Results

Results of this study will be published following guidelines, including those for authorship, established by the International Committee of Medical Journal Editors (20). When reporting results of this study, the appropriate Strengthening the Reporting of Observational (STROBE) Studies in Epidemiology checklist and ESMO Guidance for Reporting Oncology real-world evidence studies, will be followed (21, 22).

# References

1. Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, et al. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. N Engl J Med. 2020;383(7):640-9.

2. Mamdani H, Matosevic S, Khalid AB, Durm G, Jalal SI. Immunotherapy in Lung Cancer: Current Landscape and Future Directions. Front Immunol. 2022;13:823618.

3. Tian T, Yu M, Yu Y, Wang K, Tian P, Luo Z, et al. Immune checkpoint inhibitor (ICI)-based treatment beyond progression with prior immunotherapy in patients with stage IV non-small cell lung cancer: a retrospective study. Transl Lung Cancer Res. 2022;11(6):1027-37.

4. Cherny N, Sullivan R, Torode J, Saar M, Eniu A. ESMO European Consortium Study on the availability, out-of-pocket costs and accessibility of antineoplastic medicines in Europe. Ann Oncol. 2016;27(8):1423-43.

5. Ferrario A. Time to Entry for New Cancer Medicines: From European Union-Wide Marketing Authorization to Patient Access in Belgium, Estonia, Scotland and Sweden. Value Health. 2018;21(7):809-21.

6. Martinalbo J, Bowen D, Camarero J, Chapelin M, Demolis P, Foggi P, et al. Early market access of cancer drugs in the EU. Ann Oncol. 2016;27(1):96-105.

7. Slowley A, Phiri K, Multani JK, Casey V, Mpima S, Yasuda M, et al. Real-world treatment patterns and clinical outcomes after introduction of immune checkpoint inhibitors: Results from a retrospective chart review of patients with advanced/metastatic non-small cell lung cancer in the EU5. Thorac Cancer. 2023;14(28):2846-58.

8. Wu Y, Yu G, Jin K, Qian J. Advancing non-small cell lung cancer treatment: the power of combination immunotherapies. Front Immunol. 2024;15:1349502.

9. Miksad RA, Calip GS. Future of Cancer Treatment Guidelines: Integrating Real-World Insights for Equitable Cancer Care. JCO Clin Cancer Inform. 2024;8:e2400081.

10. Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clin Cancer Inform. 2021;5:12-20.

11. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.

12. Golozar A, Lawrence-Archer L, Zack T, Warner J, Reich C. Introducing ARTEMIS: Advanced Regimen Detection Using an Adapted Smith-Waterman Algorithm. OHDSI 2023 Global Symposium2023.

13. Syed H, Das AK. Temporal Needleman-Wunsch. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)2015. p. 1-9.

14. Warner JL, Dymshyts D, Reich CG, Gurley MJ, Hochheiser H, Moldwin ZH, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform. 2019;96:103239.

15. Blacketer C, Voss E. Extract Transform Load. The Book of OHDSI 2021.

16. Blacketer C. The Common Data Model. The Book of OHDSI 2021.

17. Schuemie M, Huser V, Blacketer C. Data Quality The Book of OHDSI2021.

18. Guidelines for Good Pharmacoepidemiology Practices (GPP), Revision 3. International Society for Pharmacoepidemiology; 2015.

19. Guideline on Good Pharmacovigilance Practices (GVP) - Module VIII – Post-Authorisation Safety Studies (Rev. 3). European Medicines Agency; 2017. Report No.: EMA/813938/2011 Rev 3.

20. ICMJE. Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals. . International Committee of Medical Journal Editors; 2024.

21. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.

22. Guide on Methodological Standards in Pharmacoepidemiology (Revision 11). The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP); 2010. Report No.: EMA/95098/2010.