LUCC lung – Large and Unified Cancer Cohort

First published: 20/10/2026

Last updated: 20/10/2026

Data source | Human | Hospital discharge records | Hospital inpatient records | Hospital outpatient visit records | Primary care medical records

Administrative details

Administrative details

Data source ID

1000001144

Data source acronym

LUCC lung

Data holder

Gustave Roussy

Data source type

Hospital discharge records

Hospital inpatient records

Hospital outpatient visit records

Primary care medical records

Main financial support

Funding by own institution

Funding from industry or contract research

National, regional, or municipal public funding

Care setting

Hospital inpatient care

Hospital outpatient care

Primary care – GP, community pharmacist level

Primary care – specialist level (e.g. paediatricians)

Secondary care – specialist level (ambulatory)

Data source qualification

If the data source has successfully undergone a formal qualification process (e.g., from the EMA, ISO or other certifications), this should be described.

Yes

Description of the qualification

undefinedHTA (HAS) : listed among the databases and real-world studies that may be used to respond to HAS requests for additional data in the assessment of health products and technologies. ISO : 27001 certification obtained. HDS-certified (health data hosting certification from French Governement).

Data source website

LUCC – Lung Cancer Cohort

Contact details

Antoine Meffre antoine.meffre@lifen.fr

Main

antoine.meffre@lifen.fr

contact contact contact@lifen.fr

Alternate

contact@lifen.fr

Data source regions and languages

Data source countries

France

Data source languages

English

French

Data source establishment

Data source established

01/01/2024

Data source time span

First collection: 01/01/2024

The date when data started to be collected or extracted.

Publications

Data source publications

Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions – Annals of Oncology, online 15/12/2025.

Studies

No information provided.

Data elements collected

The data source contains the following information

Disease information

Does the data source collect information with a focus on a specific disease? This might be a patient registry or other similar initiatives.

Yes

Disease details

Lung cancer metastatic

Non-small cell lung cancer

Non-small cell lung cancer metastatic

Small cell lung cancer

Disease details (other)

Lung cancer, all stages and all histologies: non-small-cell lung cancer (NSCLC), small-cell lung cancer (SCLC), pleural mesothelioma, thymoma and thymic carcinoma.

Rare diseases

Are rare diseases captured? In the European Union a rare disease is one that affects no more than 5 people in 10,000.

Yes

Pregnancy and/or neonates

Does the data source collect information on pregnant women and/or neonatal subpopulation (under 28 days of age)?

No

Hospital admission and/or discharge

Yes

ICU admission

Is information on intensive care unit admission available?

Yes

Cause of death

Not Captured

Prescriptions of medicines

Captured

Prescriptions vocabulary

not coded

Dispensing of medicines

Not Captured

Advanced therapy medicinal products (ATMP)

Is information on advanced therapy medicinal products included? A medicinal product for human use that is either a gene therapy medicinal product, a somatic cell therapy product or a tissue engineered products as defined in Regulation (EC) No 1394/2007 [Reg (EC) No 1394/2007 Art 1(1)].

No

Contraception

Is information on the use of any type of contraception (oral, injectable, devices etc.) available?

No

Indication for use

Does the data source capture information on the therapeutic indication for the use of medicinal products?

Captured

Indication vocabulary

Not coded (Free text)

Medical devices

Is information on medicinal devices (e.g., pens, syringes, inhalers) available?

No

Administration of vaccines

No

Procedures

Does the data source capture information on procedures (e.g., diagnostic tests, therapeutic, surgical interventions)?

Captured

Procedures vocabulary

Not coded (Free text)

Healthcare provider

Is information on the person providing healthcare (e.g., physician, pharmacist, specialist) available? The healthcare provider refers to individual health professionals or a health facility organisation licensed to provide health care diagnosis and treatment services including medication, surgery and medical devices.

Yes

Clinical measurements

Is information on clinical measurements (e.g., BMI, blood pressure, height) available?

Yes

Genetic data

Are data related to genotyping, genome sequencing available?

Captured

Genetic data vocabulary

Other

Genetic data vocabulary, other

Not coded. Somatic molecular alterations are extracted as free text from pathology and molecular biology reports, with no mapping to a reference terminology to date

---

Biomarker data

Does the data source capture biomarker information? The term “biomarker” refers to a broad subcategory of medical signs (objective indications of medical state observed from outside the patient), which can be measured accurately and reproducibly. For example, haematological assays, infectious disease markers or metabolomic biomarkers.

Captured

Biomarker data vocabulary

Other

Biomarker vocabulary, other

Not coded. Biomarker results are extracted as free text from clinical, pathology and laboratory reports

Patient-reported outcomes

Is information on patient-reported outcomes (e.g., quality of life) available?

No

Patient-generated data

Is patient-generated information (e.g., from wearable devices) available?

No

Units of healthcare utilisation

Are units of healthcare utilisation (e.g., number of visits to GP per year, number of hospital days) available or can they be derived? Units of healthcare utilisation refer to the quantification of the use of services for the purpose of preventing or curing health problems.

No

Unique identifier for persons

Are patients uniquely identified in the data source?

No

Diagnostic codes

Not Captured

Medicinal product information

Captured

Medicinal product information collected

Active ingredient(s)

Brand name

Dosage regime

Dose

Medicinal product vocabulary

Not coded (Free text)

Other

If 'other,' what vocabulary is used?

Line of treatment, treatment start and stop dates, reason for discontinuation and best response to treatment

Quality of life measurements

Not Captured

Lifestyle factors

Captured

Lifestyle factors

Other

Tobacco use

Sociodemographic information

Captured

Sociodemographic information collected

Age

Health area

Sex

Quantitative descriptors

Population Qualitative Data

Population age groups

Adult and elderly population (≥18 years)

Adults (18 to < 65 years)

Adults (18 to < 46 years)

Adults (46 to < 65 years)

Elderly (≥ 65 years)

Adults (65 to < 75 years)

Adults (75 to < 85 years)

Adults (85 years and over)

Estimated percentage of the population covered by the data source in the catchment area

10 % of new lung cancers in 2025

Description of the population covered by the data source in the catchment area whose data are not collected (e.g., people who are registered only for private care)

Patients who exercised their right to object (opt-out) are excluded from the data source. Patients are recruited in approximately 30 participating centres in France, including comprehensive cancer centres (CLCC), university hospitals (CHU), general hospitals (CH) and private facilities.

The data source does not provide exhaustive national coverage: the catchment population corresponds to the active lung-cancer population of the participating centres

Population

Population size

29.219

Active population size

15.000

Population by age group

Age group	Population size	Active population size
Adults (18 to < 46 years)	701	360
Adults (46 to < 65 years)	7.977	4.095
Adults (65 to < 75 years)	10.606	5.445
Adults (75 to < 85 years)	8.181	4.200
Adults (85 years and over)	1.754	900

Median observation time

Median time (years) between first and last available records for unique individuals captured in the data source

1,00

Median time (years) between first and last available records for unique active individuals (alive and currently registered) capt

0,97

Data flows and management

Access and validation

Governance details

Documents or webpages that describe the overall governance of the data source and processes and procedures for data capture and management, data quality check and validation results (governing data access or utilisation for research purposes).

Charte Comité scientifique LUCC (vfinale).pdf

English (264.25 KB - PDF) View document

https://www.lifen.fr/database/lucc-poumon

Biospecimen access

Are biospecimens available in the data source (e.g., tissue samples)?

No

Access to subject details

Can individual patients/practitioners/practices included in the data source be contacted?

Yes

Description of data collection

Medical records are collected from participating hospitals either automatically through Lifen interoperability connectors (nominal flow) or, for non-connected sites, via Datalab Bridge, a free web application that pseudonymises documents before transfer to a secure SFTP server.

Documents are pseudonymised before ingestion into the clinical data warehouse (EDS) and re-pseudonymised with a project-specific UUID in each project workspace. Variables are then extracted from unstructured text by a fine-tuned open-source Mistral AI model, with targeted human validation by clinical research professionals (hybrid AI-human workflow).

The variable set is defined and validated by the cohort's scientific committee.

Hosting is HDS-certified (health data hosting, France).

Event triggering registration

Event triggering registration of a person in the data source

Disease diagnosis

Event triggering de-registration of a person in the data source

Other

Event triggering de-registration of a person in the data source, other

Exercise of the right to object

Data source linkage

Linkage

Is the data source described created by the linkage of other data sources (prelinked data source) and/or can the data source be linked to other data source on an ad-hoc basis?

Yes

Linkage description, pre-linked

None. The data source is not created by the linkage of other data sources: it is built from medical documents collected directly from participating hospitals

Linkage description, possible linkage

Ad-hoc linkage with French administrative healthcare databases (SNDS) is considered, in particular as a strategy to recover missing data such as vital status.

Data management specifications that apply for the data source

Data source refresh

Monthly

Informed consent for use of data for research

Not Required

Possibility of data validation

Can validity of the data in the data source be verified (e.g., access to original medical charts)?

Yes

Data source preservation

Are records preserved in the data source indefinitely?

No

Approval for publication

Is an approval needed for publishing the results of a study using the data source?

Yes

Data source last refresh

01/10/2026

Common Data Model (CDM) mapping

CDM mapping

Has the data source been converted (ETL-ed) to a common data model?

Yes