Funded by the European Union

DELIVERABLE 1.5

Report and summary of the Guidelines and SOPs for structured data

WP1 – Guidelines, standards, and SOPs for RWD collection based on fully structured data

Delivery date: *31st May 2024*

Due date: *M18*

Deliverable type: *Report (R)*

Dissemination level: *Public (PUB)*

Authors

Name: Samu Eränen (HUS), Juho Lähteenmaa (HUS), Cecilie Ane Koefoed-Nielsen (RHP)

Lead contributor	1. HUS
Other contributors	2. RHP

Document history

VERSION	DATE	AUTHOR	DESCRIPTION
v.0.1	20th Mar 2024	Samu Eränen	Establishing the document
v.0.5	15th May 2024	Samu Eränen	Version shared with partners
v.1.0	28th May 2024	Samu Eränen	Modified based on reviews
v.final	31st May 2024	Samu Eränen	Finalized for submission to the EC

Internal review history

REVIEWED BY	DATE	DESCRIPTION
Ilaria Massa	22nd May 2024	Review of the document
Andrea Roncadori	22nd May 2024	Review of the document
Nea Hellman	23rd May 2024	Review of the document

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Table of contents

Executive summary

The main objective of WP1 is to develop guidelines and standard operating procedures for the collection of fully structured data, including clinical outcome measures, as part of routine clinical work in cancer hospitals. Helsinki University Hospital (HUS) and Rigshospitalet (RHP) have been collaborating on developing and testing a structured real-time data collection pathway for breast cancer and non-small cell lung cancer (NSCLC). HUS has led the design, construction, and testing of the breast cancer pathway while RHP has focused on NSCLC. Later in the project, HUS and RHP will cross-validate their respective cases.

This public deliverable report describes the guidelines and standard operating procedures (SOP) for the collection, processing, and basic analytics of the structured data. These guidelines and SOPs are a basis for education material to further dissemination and exploitation activities and to educate important stakeholders of the project. The documents can be used by other cancer centres that aim at building fully structured data collection practices into their clinical routines.

It is noteworthy that Work Package (WP) 1 will still continue untill the end of November 2024 including validation tasks and evaluating use cases of additional cancer types. In that sense this report is still provisional. The final version of guidelines and SOPs for the collection, processing, and basic analytics of the structured data will be finalized after all tasks are completed.

List of abbreviations and definitions

Abbreviation	Definition
AI	Artificial intelligence
ATC	Anatomical Therapeutic Chemicals
ECOG	Eastern Cooperative Oncology Group
EMR	Electronical medical record
ER	Estrogen receptor
HER	Human epidermal growth factor receptor
HTA	Health technology assessment
HUS	Helsinki University Hospital (Helsinki, Finland)
IHC	Immunohistochemical
ISH	In situ hybridization
MDT	Multi-disciplinary Team
NSCLC	Non-small-cell lung cancer
OMOP	The Observational Medical Outcomes Partnership
pCR	Pathological complete response
PR	Progesterone receptor
QoL	Quality of life
RCT	Randomized controlled trial
RegEx	Regular expression
RHP	Rigshospitalet (Copenhagen, Denmark)
RWD	Real-world data
RWE	Real-world evidence
TKI	Tyrosine Kinase Inhibitors
TNM	Tumor, nodules and metastasis
WP	Work package

1. Background

Real-world evidence (RWE) has become an important component in evaluating healthcare outcomes. Although randomized controlled trials (RCTs) are the primary method for assessing effects of new therapies, RWE can offer important benefits. The results of RCTs have external validity only in similar patient populations that have been studied, and the strict eligibility criteria of RCTs may significantly differ from real-world population and their outcomes. Real-world data (RWD) studies typically require less time and expense, allowing for larger sample size and longer-term follow-up. Furthermore, RWD studies can be more accessible in a regulatory and ethical manner. (Hall 2017, Slattery et al. 2020, Silverman 2009, Yang et al. 2010)

Due to the continual increase in the global cancer prevalence and rising prices of novel cancer therapies, combined with the increase of the overall aging population and the rise in people diagnosed with cancer, the global healthcare system for the treatment of cancer is at risk of becoming unaffordable. One of the present challenges in effectively utilizing RWD is the absence of a standardized data model for clinical cancer treatment information, which constrains data sharing across various registries (Kent et al. 2021). Additionally, regulatory and technical obstacles can create further complexities in integrating data from multiple sources. (Boyle et al. 2021)

For easing these burdens, Horizon Europe has funded the ONCOVALUE project. In this project coordinated by HUS, a consortium of leading European cancer hospitals in collaboration with private companies will build data collection and processing capabilities to create a high-quality clinical data source for assessing RWE. Besides structured data, unstructured data originating from medical notes and medical images will be transformed into structured data with the use of artificial intelligence (AI) technologies to enable analytics and RWE creation. For that, the primary goal of the project is to provide an end-to-end infrastructure for RWD reporting in health regulatory and health technology assessment (HTA) decision-making and to address the legal constraints in the cancer hospitals to ensure secure and legal access to RWD. Furthermore, ONCOVALUE will ensure the implementation of the developed guidelines and methodologies by providing trainings for the collection and management of high-quality RWD in European cancer centers and for the use of this data by HTA and regulatory bodies. As such, ONCOVALUE is positioned to contribute to increased cost-effectiveness and subsequent sustainability of cancer care.

2. Introduction

The main objective of WP1 is to develop guidelines and standard operating procedures for the collection of fully structured data, including clinical outcome measures, as part of routine clinical work in cancer hospitals. The outcome measurements also contain quality of life (QoL) questionnaires that fully integrate to the electronic medical records system and are collected with patient portals/applications.

HUS and RHP have been collaborating on developing and testing a structured real-time data collection pathway for breast cancer and non-small cell lung cancer (NSCLC) in structural electronic medical record (EMR) environment. HUS has led the design, construction, and testing of the breast cancer pathway while RHP has focused on NSCLC. Later in the project, HUS and RHP will cross-validate their respective cases.

As the ONCOVALUE project progresses, HUS and RHP will expand their structured data entry system to encompass additional cancer types. The selection of additional cancer types has been discussed in the ONCOVALUE Scientific and Clinical coordination group meetings during 2023 and 2024. HUS and RHP will be testing and validating data collection for colon cancer. Also melanoma vill be validated if it is feasible during project timeline.

This deliverable report describes the guidelines and standard operating procedures (SOP) for the collection, processing, and basic analytics of the structured data. These guidelines and SOPs are a basis for education material to further dissemination and exploitation activities and to educate important stakeholders of the project. The documents can be used by other cancer centres that aim at building fully structured data collection practices into their clinical routines.

The stage and tumor characteristics of cancer have impact on the treatment setting to be chosen and care pathways differ remarkably depending on the setting. The guidelines and SOPs have been reported from the perspective of specific use cases of breast cancer and NSCLC to demonstrate the feasibility of collecting the current real-world-data (RWD) of these treatment settings from a structured data repository environment. These use cases have been:

- Pembrolizumab in the neoadjuvant treatment of triple-negative breast cancer
- Real-world progression free survival and overall survival in patients with metastatic NSCLC treated with tyrosine kinase inhibitors in first line

HUS has also chosen one specific metastatic breast cancer use case for data evaluation which will be evaluated later in 2024.

- Trastuzumab Deruxtecan for the treatment of HER2-positive metastatic breast cancer

It is noteworthy that Work Package 1 and especially *TASK 1.3 – Validating the structured data entry for selected cancer types* and *Task 1.5 - Guidelines and SOPs for the collection, processing, and basic analytics of the structured data* will continue untill the end of November 2024. RHP will validate the structured data entry approach for breast cancer developed at HUS and HUS will validate the structured data entry approach for NSCLC cancer developed at RHP. In that sense the guidelines and SOPs presented in this report are still provisional. The scope for data collection has been quite strictly in data entities relating to use cases (RCTs) at this phase. The scope may be expanded to a broader set of variables as HUS and RHP validate their respective cases. In addition, both cancer centers will also analyse and validate data entries for 1-2 other cancer types. The final version of guidelines and SOPs for the collection, processing, and basic analytics of the structured data will be finalized once these tasks are completed.

3. Structural EMR and data repository environment

3.1 Benefits of structural documentation

Structural documentation practices require the user to document clinical data by choosing from predetermined set of parameters. This kind of documentation plays a crucial role in ensuring the efficiency, quality and safety of healthcare delivery. Advantages of structural documentation in healthcare include for example:

- As healthcare systems become more interconnected, structural documentation becomes essential for interoperability. Standardized documentation formats and structures facilitate the exchange of information between different healthcare systems and providers.
- Provides a foundation for data collection and analysis. Researchers can use this information to study real-world healthcare trends, assess the effectiveness of intervention sand established treatments, and contribute to evidence-based practices.
- Facilitates effective communication within healthcare teams. It provides a standardized way to convey information about roles, responsibilities, workflows, and processes. This, in turn, helps in minimizing misunderstandings and errors.
- Supports continuous quality improvement initiatives. By documenting processes and organizational structures, healthcare providers can identify areas for improvement, implement changes, and track the impact of those changes over time.
- Helps ensure compliance with regulations by providing a clear record of the organization's structure, policies, procedures, and practices.
- Understanding the structural aspects of healthcare organizations enables better resource allocation. This includes optimizing staff roles, managing equipment and facilities, and ensuring that resources are allocated effectively to meet patient needs.
- Comprehensive documentation assists in risk management by identifying potential risks associated with specific processes or roles. This allows healthcare organizations to implement strategies to mitigate these risks and improve patient safety.
- Well-documented structures and processes contribute to the delivery of patient-centered care. Healthcare providers can use this information to streamline workflows, reduce wait times, and improve overall patient satisfaction.

Structural documentation enables coherent data within a hospital. However, used data parameter sets are typically country-specific or even organization-specific. Aggregating and analyzing data from different hospitals requires data harmonization process and common data models such as The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) (OHDSI Data Standardization 2024) which is an open community data standard, designed to standardize the structure and content of observational data. This process is specified in detail in Task 5.1 of WP5 of ONCOVALUE.

3.2 Structural documentation in HUS and RHP

Both HUS and RHP share the same EMR provided by EPIC Systems Corporation. Clinical documentation in EPIC is strongly based on structural data fields or so-called smart elements in free text which enables data to be stored in structural form. The advantage of having the same EMR between HUS and RHP is that the validation of use cases will be easier as some fundamentals of documentation practices are similar. In addition, both hospitals can learn from each other and find best practices for their documentation. HUS, for example, has built documentation templates for clinical visits of breast cancer patients during this project and RHP can exploit this work in their workflows.

Hospital EMRs rarely cover all the clinical or administrational documentation needed for RWD collection. In HUS, for example, laboratory and pathology information is documented in separate clinical systems, which include essential variables for RWD collection. Concerning RWD collection for specific use case, it is essential to identify all the associated clinical systems in which relevant data is documented.

In addition to EMRs, hospitals benefit from having separate data processing, analysis and reporting environments which are linked to broader set of health-related data than just EMR originated clinical data. Secondary data repositories have many benefits when extracting RWD for clinical or regulatory purposes. It enables, for example:

- Gathering data from several data sources over time, including data entities not directly stored in the EMR database
- Processing large quantity of data at the same time
- Anonymization and pseudonymization capabilities and thus easier access to data from data privacy GDPR law perspective
- Versatile data processing, curation and analysis tools attached

HUS, for example, has implemented a Datalake data repository which retrieve real-time data from several data sources, including cost and administrative data in parallel to clinical data. The simplified picture of HUS Data architecture is presented in Figure 1. below.

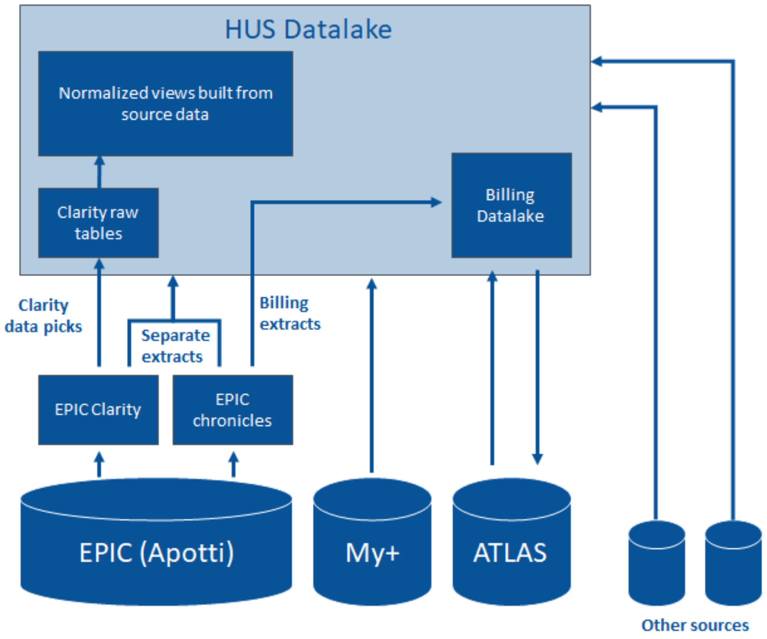

Figure 1. HUS Data architecture (The Datalake has dozens of different data source systems, but only relevant data sources have been included in the figure. My+ is the laboratory information system used in HUS and detailed cost breakdown is derived from the ATLAS billing system.)

This report of guidelines and SOPs for structured data is independent of the EMR the cancer center is using. Data collection of the use cases presented in this report is completed in EPIC environment, but the guidelines and SOPs are presented in such a general level that they can be utilized in any structural EMR environment.

3.3 Data quality requirements for documentation in structural data environment

Good data quality requires thorough assessment of documenting practices through following principles: (Statistics Finland 2023)

- Data accuracy and validity: Data is documented according to standardized vocabularies and care guidelines. For example, used terminology and gradings are documented in coherent manner.
- Data consistency: Data is documented as structurally as possible. When care workflow or documentation template does not support structured data entry, it is still valuable to use as coherent terminology as possible. Structured documentation applies to both professional and patient-related documentation.
- Data timeliness: Documentation time points or intervals are jointly defined and used within organization. Documentation intervals may differ between different treatment lines or patient groups but should be coherent within them. For example, patient questionnaires are collected at equal time intervals within specific treatment line.
- Data relevance: Data is documented at reasonable scope within hospital. Essential data variables are jointly agreed within organization and their documentation should be unavoidable. Data should be regularly documented only if it is utilized – either in clinical or secondary use.
- Data completeness: Relevant data is documented comprehensively. The objectives for completeness of data (variables) have been jointly agreed and monitored within organisation. For example, desirable response rates have been set for patient questionnaires. If not achieved, remedial actions are deployed.

Structural data environment and documentation ensures data quality from several of the principles described above. However, achieving good quality RWD requires also good planning and cooperation with end users performing the actual documentation. Mutual understanding among end users about the needs and advantages of adequate documentation practices enables to achieve satisfactory data completeness, which is crucial for fully exploiting RWD.

Data quality procedures are assessed more thoroughly in Task 5.2 – Data quality procedures of WP5 in the ONCOVALUE project.

4. Guidelines for collection, processing, and basic analytics of structured data

Collection of relevant structural data is entirely dependent on the documentation practices in use in the cancer center. There is natural variation in documentation practices between centers in different coutries as local legislative and regulative requirements vary, clinics use different EMRs systems, and there are differences in clinical pathways. Pursuing to harmonize these kinds of fundemantal organizational practices between all cancer centers would be very time consuming or even impossible. The approach for guidelines and SOPs for collection of structured data is not to form unambiguous and single documentation practice for each relevant data entity in use cases included in this report. Rather, the approach of the report is to specify the desired data entities and show different ways of collecting them. Each cancer center must analyze their documentation practices before choosing the most appropriate way of data collection. If some relevant data entity is not documented structurally or the completeness of its documentation is very low, appplicable documentation practices must be implemented.

The scope of data collection in this report has been mainly the data content in chosen use cases. Later in the Oncovalue project *WP2 – Working towards a future proof HTA framework based on hybrid RWD collection* will define the indicators necessary for HTA regulators. They will be taken into account also in the scoping of data collection in WP1.

4.1 Guidelines for breast cancer documentation in structural data environment

The use case of breast cancer neoadjuvant setting in this deliverable has been Pembrolizumab in the neoadjuvant treatment of early triple-negative breast cancer (Schmid et al. 2020).

4.1.1 Care path of early-stage breast cancer

This chapter is referring to Finnish diagnostics and clinical practice guideline for breast cancer specified by Finnish breast cancer group (Finnish breast cancer group 2024). These clinical guidelines form the basis for guidelines and SOPs for collecting structured data in this report.

In Finland breast cancer diagnostics start usually outside the cancer center, for example in primary healthcare, and the patient arrives with a referral to secondary or tertiary healthcare. This report does not include diagnostic phases prior to the referral. However, certain diagnostics must be performed prior to referral is approved by cancer center.

In recent years, the early-stage breast cancer care path has changed so that an increasing number of patients begin treatment with systemic therapy. In neoadjuvant therapy, systemic treatments are initiated before surgery. Its advantage lies in the rapid initiation of the therapy and the possibility of assessing treatment response, as well as tailoring further treatment if complete treatment response is not achieved. Neoadjuvant therapy offers the opportunity to use new cancer drugs for which there is no equivalent evidence in the adjuvant setting.

In Finnish cancer centers, the beginning of the patient's care path is planned either in a preoperative multidisciplinary meeting or by experienced specialist physicians handling referrals. The decision to start neoadjuvant therapy is made in a multidisciplinary meeting. The neoadjuvant therapy assessment requires determination of hormone receptor, Ki67, and HER2 status from core needle biopsy to enable treatment pathway selection.

Before starting neoadjuvant therapy, the Finnish breast cancer treatment guideline (Finnish breast cancer group 2024) requires to:

- Determine and record the preoperative clinical radiological clinical TNM stage based on tumor size and lymph node involvement.
- Determine the bioprofile of the breast cancer tissue obtained from PNB: ER, PR, HER2, proliferation, histology, grade.
- Determine and record the diagnosis code (C50.xx) and clinical cancer registry notification.
- Perform a body CT scan for patients with cT3-4 or cN2-3 status.
- Assess the need for and testing criteria for hereditary breast cancer susceptibility.
- (If necessary, perform hCG testing for women of childbearing age, and estradiol and FSH testing for women in menopause.
- If the family size is not yet complete, consult a gynecologist for fertility preservation and potential future fertility treatments)

4.1.2 Data collection for neoadjuvant setting of breast cancer

Defining the cohort

As the breast cancer use case in this report is associated with neoadjuvant treatment setting, the essential starting point for collecting use case data is to extract the neoadjuvant patient cohort.

The most explicit way to identify neoadjuvant patients is clinical documentation practice and corresponding structural documentation field in EMR in which healthcare professional fills the decision about (neoadjuvant) treatment setting when appropriate treatment indications are actualized. This single data field would ease the recognition of desirable cohort and correct data collection. Obviously, this data element must be considered as intention-to-treat variable and later cohort definition requires the evaluation of the actualization of treatment. If that kind of data field is not directly available, neoadjuvant patient cohort can be extracted using other documentation entities. These data entities in prioritized order are for example:

- Procedure codes relating to neoadjuvant treatment
- Neoadjuvant treatment terminology in use in medication orders / order bundles. For example, neoadjuvant term used in medication order naming
- Medication order dates of systemic therapy, which occur between preoperative pathology and surgery
- (Breast cancer diagnosis codes)

In many cases, cohort extraction requires using combination of several data entities. If there have been changes in clinical documentation practices during the analysis period, different extraction combinations may be needed for different patients. In the initial phase of cohort extraction, it is also important to identify possible patients or patient groups that do not follow the standard care pathway. For example, there may some neoadjuvant patients which does not end up having surgery after neoadjuvant treatment. In that case the extraction of neoadjuvant cohort must be performed without surgical documentation. Appendix I includes a link to Oncovalue GitHub environment which includes the coding used in HUS for cohort definitions and data collection.

Baseline characteristics

Characteristics of the patients at baseline corresponds to the data in Table 1 in use case under assessment. In the case of Pembrolizumab in the neoadjuvant treatment of early triple-negative breast cancer at least the following variables must be extracted.

Table 1: Description of the different baseline variables and the source of extraction for breast cancer use case

Variable	Source of extraction
Patient age	Extracted from patient basic documentation
Menopausal status (premenopausal / postmenopausal)	Extracted either from documentation by healthcare professional during pretreatment visits or by patient filled preliminary information form.
ECOG performance-status score	Extracted from documentation by healthcare professional during pretreatment visits Gathered as a set over a timespan from the arrival of the core-needle biopsy to the initiation of the neoadjuvant treatment, with respect to the specific patient, organ, and neoadjuvant episode. Data from different sources get concatenated and the maximum value is selected.
Administration of carboplatin	Extracted from medication administration
Primary tumor classification (T1-T4)	Extracted from pretreatment pathology reports.
Nodal involvement (positive / negative)	Extracted from pretreatment radiology reports or from pathology report (fine needle aspiration biopsy)
Overall disease stage	Extracted from pretreatment pathology reports. Can be derived from combining tumor classification and nodal involvement.
HER2 status score (ISH / IHC) (negative, 1+, 2+, 3+ or proportion 0-100%)	Extracted from pretreatment pathology report and diagnosis. Gathered as a set over a timespan from the arrival of the core-needle biopsy to the initiation of the neoadjuvant treatment, with respect to the specific patient, organ, and neoadjuvant episode. ISH results (positive or negative) can be used for defining classical HER2 status. More granular HER2 status requires IHC classification. If patient has multiple IHC values, the

	maximum value is selected. Multifocal tumor cases require still further assessment.

Other baseline characteristics

In addition to the data entities of neoadjuvant use case assessed in this report, there are other data variables of baseline characteristics, which may be relevant to other neoadjuvant treatment use cases. These are optional at this phase of the project and the final scope of baseline characteristics will be specified until the end of WP1.

Table 2: Description of the additional baseline variables and the source of extraction for breast cancer use case

Variable	Source of extraction
Cancer histology	Extracted either directly from pretreatment pathology report or by combining information with ICD10 (diagnoses' second digit). Gathered as a set over a timespan from the arrival of the core-needle biopsy to the initiation of the neoadjuvant treatment, with respect to the specific patient, organ, and neoadjuvant episode. Special caution must be taken with extraction as patient may have multiple histologies and reported histologies may change over time. Multicentric tumor cases require still further assessment.
Pre-operative ER and PR statuses (positive / negative)	Extracted from pretreatment pathology report and diagnosis. Gathered as a set over a timespan from the arrival of the core-needle biopsy to the initiation of the neoadjuvant treatment, with respect to the specific patient, organ, and neoadjuvant episode.
Pre-operative tumor grade	Extracted from pretreatment pathology report and diagnosis and from pathological texts. Gathered as a set over a timespan from the arrival of the core-needle biopsy to the initiation of the neoadjuvant treatment, with respect to the specific patient, organ, and neoadjuvant episode. After extracting the tumor grades as a set, selecting the maximum grade.
Pre-operative MIB-1 (Ki-67)	Extracted from pretreatment pathology report and diagnosis. Gathered as a set over a timespan from the arrival of the core-needle biopsy to the initiation of the neoadjuvant treatment, with respect to the specific patient, organ, and neoadjuvant episode.

Medication data

All medication data related to neoadjuvant episode is extracted from EMR medication administration. Data should include at least all systemic therapy administered to patients:

- Types of therapies and medicinal products, using:
  - ATC codes
  - Trade names of the medicinal product
  - Active substances
- Dates of all therapy cycles
- Number of therapy cycles
- Dosages
- Unplanned medication changes:
  - discontinuation of treatment
  - change of treatment line
  - dosage changes
  - reason for medication changes

Extraction of other medication data unrelated to cancer systemic therapy is not excluded in this report as it is not studied in the RCT of use case. Nevertheless, this data entity is still under assessment in WP1. The challenge with concurrent medication data is that it may not be exist or can be incomplete in hospital EMR. Its extraction requires access or integration to data repositories outside hospital EMR.

Table 3: Description of the different post treatment variables and the source of extraction for breast cancer use case

Variable	Source of extraction
histopathologic TN **clinical M status**	Extracted from post-operative pathology reports. Can be derived from combining tumor classification and nodal involvement.
Primary tumor classification (T1-T4)	Extracted from post-operative pathology reports.
Post-operative invasive tumor diameter	Extracted from post-operative pathology report.
Nodal involvement (positive / negative)	Extracted from post-operative pathology report
Post-operative number of removed lymph nodes	Extracted from post-operative pathology report.
Post-operative number of malignant lymph nodes	Extracted from post-operative pathology report.
Post-operative largest lymph node metastases	Extracted from post-operative pathology report.
Post-operative extracapsular extension node	Extracted from post-operative pathology report.
Cancer histology	Extracted either directly from post-operative pathology report. Special caution must be taken with extraction as patient may have multiple histologies. Multicentric tumor cases require still further assessment.
HER2 status (ISH / IHC) (negative, 1+, 2+, 3+ or proportion 0-100%)	Extracted from post-operative pathology report and diagnosis.
Post-operative ER and PR statuses (positive / negative)	Extracted from post-operative pathology report and diagnosis.
Post-operative tumor grade	Extracted from post-operative pathology report and diagnosis and from pathological texts.
Post-operative MIB-1 (Ki-67)	Extracted from post-operative pathology report and diagnosis.
Post-operative tumor pathological response	Extracted from post-operative pathology report.
Post-operative nodal pathological response	Extracted from post-operative pathology report.
Recidual cancer burden	Extracted from post-operative pathology report.

Adverse events

Documenting adverse events is fundamental part of RCTs as regulators requires comprehensive reporting of medication adverse events for medicine assessment and licensing. Adverse events have great impact on the quality of life of patients and thus they also have influence on HTA guidelines of new cancer drugs. In the clinical treatment pathway, the occurrence of severe adverse events may lead to changes in systemic treatment or even discontinuation of treatment.

Although the naming conventions and grading of adverse events are universally standardized, the list of different adverse events is long and heterogeneous. The variety of occurring adverse events depends strongly on many simultaneous factors like medicine administered, treatment setting, phase of the treatment and phase of the disease. This means that implementing one general documenting practice for different use cases does not produce adequate adverse event data and different methods must be applied to their data collection.

Most explicit group of adverse events include all symptoms which are verified through laboratory results. The data is extracted from laboratory information systems and adverse event grading can be performed automatically according to threshold values agreed. Data collection time points are set by treatment guidelines and thus consistent within treatment setting.

Other adverse events are assessed and documented either by healthcare professionals or patients in accordance with national laws or regulations. From the quality perspective, documentation by healthcare professionals will produce more homogeneous and comprehensive adverse event data, but it includes many practical challenges:

- Building templates or data fields for comprehensive adverse event data documentation may require quite massive system building work. Adverse events are medicine specific, and templates or data fields must be built specifically for each medicine. Furthermore, new templates and data fields need to be updated continuously as new medicine emerges to market.
- General (medicine independent) documentation templates or data fields covering all relevant adverse event entities lead to long drop-down lists from which end user chooses the desirable ones. The usability of this kind documentation is low and very unlikely to be adapted by healthcare professionals. This approach also needs quite remarkable system building work as the grading of different adverse events varies and grading must be built individually for each.
- Adverse events are documented by healthcare professionals only during patient contacts. Documentation between planned clinical visits must not cause significant additional workload for healthcare professionals.

Patient-centered adverse event documentation practice requires electronic documentation templates or questionnaires which are integrated into hospital EMR and data repositories. To achieve good data completeness, major emphasis needs to be placed to usability and easiness of documentation.

HUS is currently in the process of implementing adverse event documentation questionnaires in the patient portal of its EMR. The implementation will be completed later this year. PRO-CTCAE is used as content and grading of events.

Quality of Life measures

Quality of life (QoL) measures data is not directly included in use case RCT of HUS. However, QoL data will be essential for HTA regulators. Collecting Quality of Life data is presented in *Deliverable 1.4 – Standard and report for the collection and analytics for QoL data*.

4.2 Guidelines for NSCLC documentation in structural data environment

4.2.1 Description of the scope of the report

The diagnostic pathway

The diagnostics of NSCLC is carried out in the setting of lung medicine and the diagnosis of NSCLC is settled on a multi-disciplinary team (MDT) meeting after a process that for the majority of patients involves FDG-PET CT and broncoscopy with biopsies and pathology diagnostics. The MDT consists of physicians from different relevant medical specialties – e.g. lung medicine, pathology, radiology, nuclear medicine and oncology. The MDT determines the diagnosis of NSCLC or whether further diagnostics are to be made and evaluates the staging according to the cancer staging system TNM. Depending on the staging and thereby the further treatment strategy – also settled by the MDT – the patient is referred to the relevant department of thoracic surgery or oncology.

All data regarding the diagnosis of NSCLC are in the contest of secondary care on the regional Hospitals. In the eastern part of Denmark (The Capitol Region and The Region of Zealand) all data is processed and stored in the EMR provided by EPIC Systems Corporation. Different relevant diagnoses are reported structurally by clinicians in a diagnoses list using ICD 10 nomenclature. Also, the TNM staging and location of metastases can be reported using ICD 10 nomenclature in the diagnoses list. Pathology data is both reported unstructured in a text format and documented structurally using SNOMED (SNOMED 2024) nomenclature. Radiology data remain to be reported structurally.

Treatment settings

For the early NSCLC stages (TNM stage I and II), surgery is possible whenever surgeons find the patient suitable for surgery. Both presurgical staging and the grade of tumor resection during surgery determine whether patients are offered adjuvant medical oncological treatment. For locally advanced stages (TNM stage III) radiotherapy given concomitant with chemotherapy is offered when oncologists find the tumor relevant for radiotherapy and the patient suitable for the full treatment course. This concomitant treatment can be supplemented with immunotherapy. For the late stages of NSCLC (TNM stage IV), which most of the patients are diagnosed with, either chemotherapy, immunotherapy, a combination of these or targeted therapy can be offered. The oncological treatment being offered depends on different cancer characteristics, e.g. histology, biomarkers and mutations.

Requirements for documentation rely on the communication between the different medical specialties and between doctors internally on a department. Thus, MDT notes and clinicians' notes are all available in the EMR – but in unstructured text format (clinicians notes). All data regarding treatment is structured, all medication is documented based on the ATC nomenclature, procedures, hereunder radiotherapy and surgery, is documented using SKS and UTC nomenclature.

4.2.2 Description of the use case

Background

Tyrosine kinase inhibitors (TKI) is a targeted antineoplastic therapy and RCT's have established the documentation for a significantly prolonged progression free and overall survival for patients with metastatic NSCLC harboring a targetable driver mutation compared with chemotherapy (Lee et al. 2015, Lee et al. 2017, Soria et al. 2018, Ramalingam et al. 2020, Solomon et al. 2014, Solomon et al. 2018, Mok et al. 2020, Hotta et al. 2022, Ahn et al. 2022, Solomon et al. 2023). This efficacy of TKI in real-world populations of patients with metastatic NSCLC remains to be evaluated.

Study population

Inclusion criteria are:

- NSCLC harboring a targetable mutation (EGFR, ALK or ROS1)
- Diagnosis after 1st of January 2016
- Treatment with TKI in 1st line

There are no exclusion criteria.

Study objects and endpoints

Study objects will be the *clinicopathological characteristics* in patients diagnosed with metastatic NSCLC harboring a targetable mutation (Epidermal Growth Factor Receptor (EGFR) exon 19 deletion, EGFR exon 21 insertions or uncommon EGFR mutations, Anaplastic Lymfoma Kinase (ALK) rearrangement, or c-ROS oncogene 1 (ROS1) translocations) and treated with a tyrosine kinase inhibitor (EGFRex19del or EGFRex21mut: 1st generation gefitinib, erlotinib or afatinib, 2nd generation dacomitinib, or 3rd generation 21simertinib. ALK rearrangements: 1st generation crizotinib, 2nd generation alectinib, brigatinib or ceritinib, or 3rd generation lorlatinib. ROS1 translocations: entrectinib, crizotinib) in first line treatment. Primary endpoints are real world progression free survival (rwPFS), time to treatment change (rwTTC) and overall survival (rwOS). Secondary analyses will assess the relationship of the explanatory variables and the primary endpoints.

Table 4: Description of the different baseline variables and the source of extraction for NSCLC use case

Variable	Description	Source of extraction
Gender	Biological gender (sex) at birth	Directly from EMR
Age	Age (days)	Directly from EMR
BMI	Body Mass Index	Directly from EMR
ECOG_PS	European Collaborative Oncology Group performance status	Directly from EMR

Previous_cancer	Previous cancer (EPIC period)	Extracted directly from the EMR “Diagnosis list” using ICD-10
Comorbidities	ICD10 + ATC	Extracted directly from the EMR “Diagnosis list” using ICD-10, supplemented with ATC codes from national register “Shared Medication Record” (FMK)
Smoking_status	Smoking status	Directly from EMR
Alcohol_consumption	Alcohol consumption	Directly from EMR
Brain_metastases	Presence of CNS metastases	Radiology reports
Bone_metastases	Presence of bone metastases	Radiology reports
Liver_metastases	Presence of liver metastases	Radiology reports
Lymph_node_metastases	Presence of lymph node metastases	Radiology reports
Lung_metastases	Presence of lung metastases	Radiology reports
Adrenal_gland_metastases		Radiology reports
Metastases	Number of different metastatic sites	Radiology reports
Histology	NSCLC histology type	Pathology reports
PD.L1 status	Biomarker PD.L1 status	Pathology reports
Driver_mutation	EGFR (exondel19, L585R or uncommon), ALK or ROS1	Pathology reports
T_stage	Generic t stages across all cancers	Radiology reports or from clinicians' notes
N_stage	Generic n stages across all cancers	Radiology reports or from clinicians' notes
M_stage	Generic m stages across all cancers	Radiology reports or from clinicians' notes
FIGO_stage	Generic combined TNM stage across all cancers	Radiology reports or from clinicians' notes

4.2.3 Collecting data

Collecting data to define the study cohort

- The diagnosis of NSCLC is extracted from the "diagnosis list" using ICD-10 terminology. Information on timing is integrated in this data output
- The histology and mutation status are derived from the National Pathology Registry, "Patobank" using SNOMED terminology
- Treatment plans derived from "Beacon" in the EMR contains information on date and ATC codes for every antineoplastic treatment ordered and administered to the patient since 2016

Collecting patient and cancer characteristics

- Age and sex are continuously updated in the EMR with data from the "Danish Population Register"
- Other patient characteristics such as weight, height, BMI, smoking status and alcohol consumption will be extracted directly from the EMR
- Cancer characteristics will be extracted from the EMR "Diagnosis list" using ICD-10 terminology, "Xero Viewer" using text format from radiology reports (unstructured) and the National Pathology Registry, "Patobank" using SNOMED terminology. Both Xero Viewer and Patobank is integrated in the EMR platform.
- Comorbidities will be extracted from the EMR "Diagnosis list" using ICD-10 and supplemented with ATC codes from the national register "Shared Medication Record" (FMK)

Collecting outcome measures

- The vital status is continuously updated in the EMR with data from the "Danish Population Register"
- Treatment responses, i.e. progression, are extracted from "Xero Viewer" using text format from radiology reports (unstructured)
- Data regarding the time on different treatment lines and different treatment changes are derived from "Beacon" in the EMR using ATC codes

Line of Treatment (LoT) is defined as the medical antineoplastic treatment given to the patient corresponding to one cancer diagnosis. The treatment intent of LoT can be curative or palliative. The LoT continues until next LoT, death, or loss to follow-up, i.e. censoring.

Treatment change is defined as any changes in initial treatment strategy: discontinuation of treatment, change of LoT or adding any antineoplastic medication, radiotherapeutic or surgical procedure to a current LoT.

The initial treatment is defined as the antineoplastic treatment given to the patient within the first 3 months after baseline.

Unstructured data

Information about numbers and location of metastases requires manual review of radiological reports and will be collected by medical doctors.

ECOG performance status will manually be collected by medical students from clinician's notes in SP when such appears.

The assessment of radiological interpretation of the treatment responses, i.e. complete or partial response, stable disease or progression, requires manual review of radiological reports and will be collected by medical doctors. Since treatment responses are not automatically extracted, but relies on manual evaluation, some degree of individual interpretation is unavoidable when it comes to this outcome.

All manually reported data are collected in the Research Electronic Data Capture database (RedCAP) and exported to Azure Machine Learning (AML). RedCAP is a web-based software platform designed to support data capture for research studies hosted at Copenhagen Capitol Region.

Data management

All data come directly from the EMR or are gathered through REDCap, and a unique and anonymized ID is made for every patient to secure interoperability.

Data is stored in the cloud-based platform Azure Machine Learning (AML) as csv files for further definition of cohort and variables and secondarily for the statistical analysis in RStudio.

References

The following sources have been referred to in this document:

Ahn MJ, Kim HR, Yang JCH, Han JY, Li JY, Hochmair MJ, Chang GC, Delmonte A, Lee KH, Campelo RG, Gridelli C, Spira AI, Califano R, Griesinger F, Ghosh S, Felip E, Kim DW, Liu Y, Zhang P, Popat S, Camidge DR. Efficacy and Safety of Brigatinib Compared With Crizotinib in Asian vs. Non-Asian Patients With Locally Advanced or Metastatic ALK-Inhibitor-Naive ALK+ Non-Small Cell Lung Cancer: Final Results From the Phase III ALTA-1L Study. Clin Lung Cancer. 2022 Dec;23(8):720-730. doi: 10.1016/j.cllc.2022.07.008. Epub 2022 Jul 21. PMID: 36038416.

Boyle J., Hegarty G., Frampton C., Harvey-Jones E., Dodkins J., Beyer K., George G., Sullivan R., Booth C., Aggarwal A.; Real-world outcomes associated with new cancer medicines approved by the Food and Drug Administration and European Medicines Agency: A retrospective cohort study. Eur J Cancer. 2021 Sep;155:136-144. doi: 10.1016/j.ejca.2021.07.001. Epub 2021 Aug 6. PMID: 34371443; PMCID: PMC8442759.

Finnish Breast Cancer Group 2024; National diagnostics and clinical practice guideline for breast cancer; https://rintasyoparyhma.yhdistysavain.fi/@Bin/199460/Rintasy%C3%B6v%C3%A4n+valtakunnallinen+diagnostiikka-+ja+hoitosuositus+2024.pdf (30 Jan 2024)

Hall P.; Real-world data for efficient health technology assessment. Eur J Cancer. 2017 Jul;79:235-237. doi: 10.1016/j.ejca.2017.04.003. Epub 2017 May 15. PMID: 28522211.

Hotta K, Hida T, Nokihara H, Morise M, Kim YH, Azuma K, Seto T, Takiguchi Y, Nishio M, Yoshioka H, Kumagai T, Watanabe S, Goto K, Satouchi M, Kozuki T, Shukuya T, Nakagawa K, Mitsudomi T, Yamamoto N, Asakawa T, Yoshimoto T, Takata S, Tamura T. Final overall survival analysis from the phase III J-ALEX study of alectinib versus crizotinib in ALK inhibitor-naïve Japanese patients with ALK-positive non-small-cell lung cancer. ESMO Open. 2022 Aug;7(4):100527. doi: 10.1016/j.esmoop.2022.100527. Epub 2022 Jul 14. PMID: 35843080; PMCID: PMC9434408.

Kent S., Burn E., Dawoud D., Jonsson P., Østby J., Hughes N., Rijnbeek P., Bouvy J.; Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics. 2021 Mar;39(3):275-285. doi: 10.1007/s40273-020-00981-9. Epub 2020 Dec 18. PMID: 33336320; PMCID: PMC7746423.

Lee CK, Wu YL, Ding PN, Lord SJ, Inoue A, Zhou C, Mitsudomi T, Rosell R, Pavlakis N, Links M, Gebski V, Gralla RJ, Yang JC. Impact of Specific Epidermal Growth Factor Receptor (EGFR) Mutations and Clinical Characteristics on Outcomes After Treatment With EGFR Tyrosine Kinase Inhibitors Versus Chemotherapy in EGFR-Mutant Lung Cancer: A Meta-Analysis. J Clin Oncol. 2015 Jun 10;33(17):1958-65. doi: 10.1200/JCO.2014.58.1736. Epub 2015 Apr 20. PMID: 25897154.

Lee CK, Davies L, Wu YL, Mitsudomi T, Inoue A, Rosell R, Zhou C, Nakagawa K, Thongprasert S, Fukuoka M, Lord S, Marschner I, Tu YK, Gralla RJ, Gebski V, Mok T, Yang JC. Gefitinib or Erlotinib vs Chemotherapy for EGFR Mutation-Positive Lung Cancer: Individual Patient Data Meta-Analysis of Overall Survival. J Natl Cancer Inst. 2017 Jun 1;109(6). doi: 10.1093/jnci/djw279. PMID: 28376144.

Mok T, Camidge DR, Gadgeel SM, Rosell R, Dziadziuszko R, Kim DW, Pérol M, Ou SI, Ahn JS, Shaw AT, Bordogna W, Smoljanović V, Hilton M, Ruf T, Noé J, Peters S. Updated overall survival and final progression-free survival data for patients with treatment-naive advanced ALK-positive non-small-cell lung cancer in the ALEX study. Ann Oncol. 2020 Aug;31(8):1056-1064. doi: 10.1016/j.annonc.2020.04.478. Epub 2020 May 11. PMID: 32418886.

OHDSI Data standardization 2024; https://www.ohdsi.org/data-standardization (30 May 2024)

Ramalingam SS, Vansteenkiste J, Planchard D, Cho BC, Gray JE, Ohe Y, Zhou C, Reungwetwattana T, Cheng Y, Chewaskulyong B, Shah R, Cobo M, Lee KH, Cheema P, Tiseo M, John T, Lin MC, Imamura F, Kurata T, Todd A, Hodge R, Saggese M, Rukazenkov Y, Soria JC; FLAURA Investigators. Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC. N Engl J Med. 2020 Jan 2;382(1):41-50. doi: 10.1056/NEJMoa1913662. Epub 2019 Nov 21. PMID: 31751012.

Schmid, P., Cortes, J., Pusztai, L., McArthur, H., Kümmel, S., Bergh, J., Denkert, C., Park, Y. H., Hui, R., Harbeck, N., Takahashi, M., Foukakis, T., Fasching, P. A., Cardoso, F., Untch, M., Jia, L., Karantza, V., Zhao, J., Aktan, G., Dent, R., & O'Shaughnessy, J. (2020). Pembrolizumab for early triple-negative breast cancer. *New England Journal of Medicine, 382*(9), 810-821. https://doi.org/10.1056/NEJMoa1910549P.

Silverman S.; From randomized controlled trials to observational studies.; Am J Med. 2009 Feb;122(2):114-20. doi: 10.1016/j.amjmed.2008.09.030. PMID: 19185083.

Slattery, J., Xavier K.; Assessing strength of evidence for regulatory decision making in licensing: What proof do we need for observational studies of effectiveness? Pharmacoepidemiology and Drug Safety 29.10 (2020): 1336-1340.

SNOMED international 2024; https://www.snomed.org (30 May 2024)

Solomon BJ, Mok T, Kim DW, Wu YL, Nakagawa K, Mekhail T, Felip E, Cappuzzo F, Paolini J, Usari T, Iyer S, Reisman A, Wilner KD, Tursi J, Blackhall F; PROFILE 1014 Investigators. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. N Engl J Med. 2014 Dec 4;371(23):2167-77. doi: 10.1056/NEJMoa1408440. Erratum in: N Engl J Med. 2015 Oct 15;373(16):1582. PMID: 25470694.

Solomon BJ, Kim DW, Wu YL, Nakagawa K, Mekhail T, Felip E, Cappuzzo F, Paolini J, Usari T, Tang Y, Wilner KD, Blackhall F, Mok TS. Final Overall Survival Analysis From a Study Comparing First-Line Crizotinib Versus Chemotherapy in ALK-Mutation-Positive Non-Small-Cell Lung Cancer. J Clin Oncol. 2018 Aug 1;36(22):2251-2258. doi: 10.1200/JCO.2017.77.4794. Epub 2018 May 16. PMID: 29768118.

Solomon BJ, Bauer TM, Mok TSK, Liu G, Mazieres J, de Marinis F, Goto Y, Kim DW, Wu YL, Jassem J, López FL, Soo RA, Shaw AT, Polli A, Messina R, Iadeluca L, Toffalorio F, Felip E. Efficacy and safety of first-line lorlatinib versus crizotinib in patients with advanced, ALK-positive non-small-cell lung cancer: updated analysis of data from the phase 3, randomised, open-label CROWN study. Lancet Respir Med. 2023 Apr;11(4):354-366. doi: 10.1016/S2213-2600(22)00437-4. Epub 2022 Dec 16. PMID: 36535300.

Soria JC, Ohe Y, Vansteenkiste J, Reungwetwattana T, Chewaskulyong B, Lee KH, Dechaphunkul A, Imamura F, Nogami N, Kurata T, Okamoto I, Zhou C, Cho BC, Cheng Y, Cho EK, Voon PJ, Planchard D, Su WC, Gray JE, Lee SM, Hodge R, Marotti M, Rukazenkov Y, Ramalingam SS; FLAURA Investigators. Osimertinib in Untreated EGFR-Mutated Advanced Non-Small-Cell Lung Cancer. N Engl J Med. 2018 Jan 11;378(2):113-125. doi: 10.1056/NEJMoa1713137. Epub 2017 Nov 18. PMID: 29151359.

Statistics Finland; Tiedon laatukriteerit ja mittaristo – soveltamisohje Tiedon laatukehikko, 2023; https://stat.fi/media/uploads/org/tilastokeskus/tiedonlaatu/tiedon_laatukriteerien_soveltamisohje.pdf (30 Jan 2024)

Yang W., Zilov A., Soewondo P., Bech O., Sekkal F., Home P.; Observational studies: going beyond the boundaries of randomized controlled trials.; Diabetes Res Clin Pract. 2010 May;88 Suppl 1:S3-9. doi: 10.1016/S0168-8227(10)70002-4. PMID: 20466165.

APPENDIX I

ONCOVALUE GitHub environment: https://github.com/ONCOVALUE