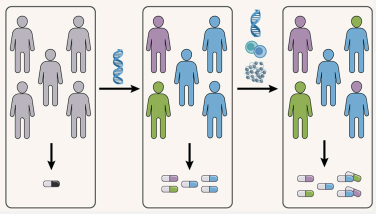
Minimal dataset for NSCLC registry

Which variables need to be collected?

Background

Real-world data can improve the evaluation and appropriate use of medicines

The treatment landscape for non-small cell lung cancer (NSCLC) is rapidly evolving, which is attributed to the rising number of subtypes within NSCLC and the rapid development of new medicines. Consequently, evaluating and efficiently applying these new medicines is becoming increasingly challenging due to a lack of comprehensive knowledge of (cost)effectiveness and appropriate use of NSCLC medicines.

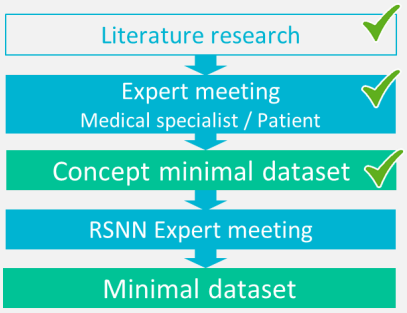


Verkerk *et al.* Cell. 2024

Therefore, there is a need to gather additional information about the application and effectiveness of these medicines post-market approval. Real-world data (RWD) from patient registries have the potential to systematically capture disease-specific data from clinical practice according to high-quality standards. To structurally answer questions about effectiveness, cost-effectiveness and appropriate use of NSCLC medicines using registry RWD, it will be necessary to determine which data must be collected in the registry: the NSCLC minimal dataset.

Defining the minimal dataset for NSCLC is part of the More-EUROPA project. Within this project, one of the case studies investigates how the assessment of drugs in NSCLC patients can be improved by using RWD from the registries of the Dutch Institute for Clinical Auditing (DICA).

Method defining minimal dataset



All participants of the RSNN expert meeting received a draft version of the minimal dataset in advance. This initial dataset was developed through a literature review of existing datasets, followed by an expert meeting held in December 2023. Eleven experts from diverse professional backgrounds – including pulmonologists (with experience in registries or cieBOM), hospital pharmacists (with registry experience), and a patient representative - attended the meeting, providing a strong foundation for the minimal dataset discussed at the RSNN expert meeting. The input gathered from the RSNN expert group was used to finalize the minimal dataset.

RSNN Expert meeting

The expert meeting took place on **25 September 2024** under the Chatham House Rule.

Aim: to gather input on the concept NSCLC minimal dataset

Discussion questions:

1. What type of questions need to be addressed with this dataset?
2. Review of the concept NSCLC minimal dataset:
   a. Which variables are missing?
   b. Which variables can be left out?

15 participants, including representatives from:

- Regulator
- Health technology assessment (HTA) expert
- Patient representative
- Industry
- Academia

Key messages

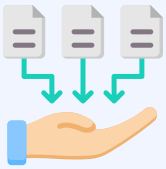
All stakeholders acknowledge the added value of registry RWD, especially in:

- Contextualizing single-arm trial results: *e.g.,* insight into the natural course of the disease
- Optimizing treatment protocols: dose-finding, identifying subgroups who benefit most

However, RWD has not been sufficient in establishing primary real-world evidence for the correlation between therapy and time-dependent outcomes, especially progression-free survival (PFS), due to heterogeneity in recording.

Revisions of the concept NSCLC minimal dataset include:

- Addition of eight new variables (VO2 max, referral, edition TNM, treatment goal, resection complete, complications surgery, dose reduction adverse event, PROM: financial impact).
- Replacement of two existing variables (country of birth, type of molecular diagnostics).
- Revision of mutation definition to include sub-mutations and follow-up data.



Data collection of the minimal dataset must be feasible:

- Keep the NSCLC minimal dataset as small as possible, and temporarily collect additional data (prospectively or retrospectively) if needed for specific questions.
- Consider using proxies when data collection is not feasible or too time-consuming. For example, time to next treatment (TTNT) could be an interesting potential proxy for PFS.

Future step: Define a feasible NSCLC minimal dataset for registry RWD by:

- Investigating which variables can be extracted from electronic health records and other data sources, considering the impact of time-consuming manual data collection.

Summary

1.What type of questions need to be addressed with this dataset?

The **industry, regulators, and HTA** highlighted that registry RWD could be used to contextualize results from single-arm trials. Using RWD as external control arm for primary evidence remains challenging due to the difficulties in matching RWD to clinical trial data. Besides that, the **regulatory and patient representatives** discussed the possibility of using RWD to track the natural course of diseases, particularly in monitoring the genetic mutations over time. Finally, **HTA** underscored the value of using RWD as an external comparator to validate the long-term survival extrapolations to improve the accuracy of cost-effectiveness assessments. The **industry** focused on the applications of RWD for optimizing treatment protocols. For example, dose-finding is an area where RWD could provide crucial information about the optimal dose. Besides that, RWD could help identify subgroups of patients who will benefit most from a particular therapy (effect modification). The experts discussed how RWD could provide valuable information about the consequences of side effects, such as when adverse events lead to dose adjustments or treatment discontinuation. One of the gaps identified by **HTA** was the challenge in predicting subsequent lines of treatment, as the subsequent lines used in practice are often unknown. This uncertainty impacts the accuracy of cost-effectiveness predictions.

Overall, the group agreed that RWD offers significant potential. However, challenges remain. RWD has not yet been utilized as a primary source of evidence of time-to-event (PFS, OS) analyses in regulatory- and HTA decision-making, largely due to the difficulties in matching between RWD and clinical trial data, making conclusions from these comparisons hard to interpret.

2. Review concept minimal dataset

Based on the input of the RSNN expert meeting, the concept minimal dataset was revised. Added variables and revised variables and definitions are indicated in red in table 1. Ethnicity was missing from the minimal dataset as it influences effectiveness and safety due to differences in physiology. However, ethnicity is hard to collect due to privacy and registration issues. Therefore, Country of residence was replaced for country of birth for a proxy of ethnicity. Molecular diagnostics is refined to type of molecular diagnostics, as this has impact on the mutation status. Besides that, mutation was further specified for sub-mutation and follow-up data. Sub-mutations are needed for contextualization of single-arm trials. Finally, it was noted that capturing side effects is important for a better understanding of their impact, course, and risk factors. However, **regulators and HTA** mentioned that RWD on side effects would not directly impact their assessment. Moreover, HTA emphasized that quality of life, as measured by EQ-5D, is a critical 'need-to-have' for their evaluation.

Variables added to the concept minimal dataset:

- VO2 max, as this can impact whether to perform surgery
- Referral for treatment, as it is important to know whether the complete treatment pathway is captured. It is currently not feasible to link data on a patient level across hospitals due to privacy regulations, such as restrictions on using citizen service number (BSN), which is essential for capturing the complete treatment pathway of all patients.
- Edition TNM staging system, as it affects the TNM stage.
- Treatment goal, as this influences the choice of treatment.
- Resection complete and complications after surgery since the influence of neoadjuvant treatment could be assessed.
- Dose reduction because of adverse events, to capture the impact of side effects on therapy and to identify the reason for dose reduction for dose-finding studies.
- Financial impact: Question 28 of EORTC QLQ-C30, since this was the only question missing in the EORTC QLQ-C30. HTA assessment needs the whole QLQ-C30 questionnaire since that is validated. Moreover, it was stressed that EQ-5D is a need to have for HTA assessment.

The group agreed that the minimal dataset should be feasible for data collection. Reducing the registration burden for healthcare professionals was a key priority. **HTA** argued that keeping this dataset as small as possible would ensure that the data collection process remains manageable, with additional data collected temporarily if needed to address specific research questions. Besides that, there was a wide interest in using proxies for data collection and several examples were discussed. TTNT was suggested as a potential proxy for PFS, as PFS is a difficult variable to systematically collect in a registry. At the same time, **patient and academia representatives** emphasized the importance of involving patients in the data collection process. By allowing patients to input their own data, such as in the PRO-Lung app where data is collected to track side effects, they can not only contribute to the data but also receive meaningful feedback about their health status.

Future steps:

It is important to define a feasible NSCLC minimal dataset for registry RWD, taking the registration burden into account. Some of the variables within this minimal dataset might not be feasible to routinely collect in a registry due to heterogeneity of the data and/or the registration burden. Therefore, within the MORE-EUROPA project we will investigate which variables from this minimal dataset can be (automatically) extracted from electronic health records and other data sources.

Contact

For further information or questions, please reach out: g.grit@dica.nl

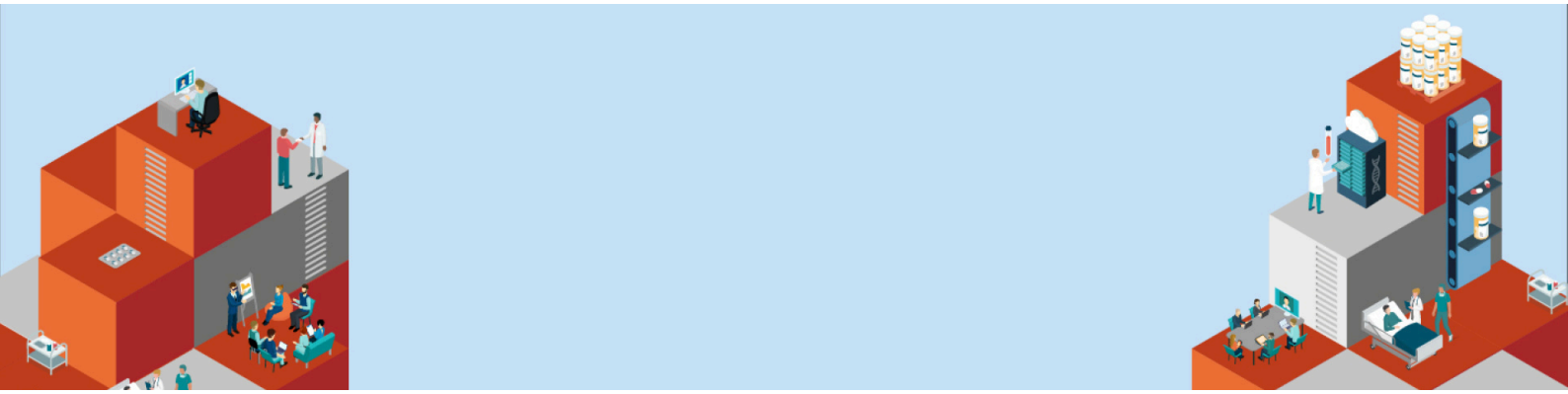


Table 1. Minimal dataset with revisions from the RSNN expert meeting indicated in red.

Variabel	Definition
Patient characteristics	
Patient identification	
Healthcare institution	Identification of hospital
Patient number	
Last name	
Postal code patient	0000AA
General patient characteristics	
Birthdate	~~Age~~ yyyy
Sex	Female/male at birth
~~Country of residence~~ Country of birth	Country of origin
Weight	Number (kg)
Height	Number (cm)
Episode	
Disease-specific patient characteristics	
ECOG	Performance score (0-4)
Smoking	Active smoker, stopped smoking (for a minimum of 1 year), never smoker (a person who does not smoke now and has smoked fewer than 100 cigarettes or similar amount of other tobacco products in his/her lifetime), unknown
Comorbidity	All comorbidities defined in 'uitkomstgerichte zorg longkanker gegevensset' + comorbidities to calculate Charlson Score
Organ function	Kidney, liver, hematological function (blood levels)
FEV1 (%)	Pulmonary function testing
DLCO (%)	Pulmonary function testing
VO2 max	Pulmonary function testing
Healthcare facility details	
Date first outpatient visit	dd-mm-yyyy
Referral for treatment	Yes/no
Diagnosis / tumor characteristics	
Date first CT / PET scan	dd-mm-yyyy
Method diagnosis	Pathological, imaging only
Date of diagnosis (Pathology)	dd-mm-yyyy (pathology report result)
Status diagnosis	NSCLC or suspected NSCLC, not pathological proven
Typing NSCLC (Pathology)	Adenocarcinoma, squamous cell carcinoma, adenosquamous carcinoma, large cell carcinoma, etc.
~~Molecular diagnostics~~ Type of molecular diagnostics	~~performed yes/no~~ NGS / WGS / other / unknown
Date of molecular diagnostic results	dd-mm-yyyy
PD-L1	<1%, 1-50% >50%
Mutation	EGFR, KRAS, ALK, EML4ALK, ROS1, BRAF V600, BRAF niet V600, RET, MET, HER2, NTRK1/2/3, NRG, KEAP1, STK11, TP53, RTK + recent updates of all treatable mutations + further specified sub mutations (*e.g.*, EGFR exon19/20/21) Including date. Frequency: follow-up
Clinical T stage	cT0, cT1, cT1a, cT1b, cT2, cT2a, cT2b, cT3, cT4, cTX, unknown
Clinical N stage	cN0, cN1, cN2, cN2a, cN2b, cN3, cNX, unknown
Clinical M stage	cM0, cM1, cM1a, cM1b, cM1c, cM1c1, cM1c2, cMX, unknown
Edition TNM staging system	7th, 8th, 9th edition
Brain imaging studies	MRI or CT brain yes/no
Brain metastases	Brain and myelum, cerebral meninges, brain, cerebral ventricle, cranial nerve, brain stem Including date
Pathalogical T stage	Determined after surgery: pT0, pT1, pT1a, pT1b, pT2, pT2a, pT2b, pT3, pT4, pTX, unknown

Pathological N stage	Determined after surgery: pN0, pN1, pN2, pN2a, pN2b, pN3, pNX, unknown
Pathalogical M stage	Determined after surgery: pM0, pM1, pM1a, pM1b, pM1c, pM1c1, pM1c2, pMX, unknown
Treatment	
Healthcare utilization data	Healthcare utilization within the hospital. Nice to have: health utilization data outside of hospital: *e.g.*, general practitioner
Treatment goal	Curative, palliative (cancer treatment with a palliative intent), unknown
Surgery	
Type of surgery	Pneumonectomy, bilobectomy, lobectomy, wedge excision, segmentectomy
Surgery date	dd-mm-yyyy
Resection complete	Yes, no
Complications after surgery	Reoperation after lung surgery + Number of clinical days of admission after lung surgery + ICU readmission after lung surgery
Radiotherapy	
Number of fractions given	Number of fractions given during radiotherapy per x
Cumulative dose	Number of fractions given in radiotherapy in total
Type of radiotherapy	AP-PA, 3D cRT, IMRT, dynamic arc therapy VMAT, proton therapy, stereotactic radiation, whole brain radiotherapy WBRT, gamma knife Update definition
Start date of radiotherapy	dd-mm-yyyy
End date of radiotherapy	dd-mm-yyyy
Radiotherapy in combination with systemic therapy	
Main treatment, neoadjuvant, adjuvant	
Concurrent or sequential systemic therapy with radiotherapy	
Start date of systemic therapy with radiotherapy	dd-mm-yyyy
End date of systemic therapy with radiotherapy	dd-mm-yyyy
Systemic therapy	
Main treatment, neoadjuvant, adjuvant	
Start date systemic therapy	dd-mm-yyyy
End date systemic therapy	dd-mm-yyyy
Type of medication	Active ingredient
Dosage schedule (including number of doses)	Dose, interval, number of doses, cumulative dose
Method of administration	Intravenous/subcutaneous/oral etc
Stopped medication early?	Stopped earlier than planned yes/no
Reason to stop medication	Systemic therapy toxicity, lung cancer progression, other
Dose reduction because of adverse event	Yes, no
Co-medication	co-medication given during a course of systemic therapy, including the following categories: 1) combination treatment 2) medication to manage adverse events 3) co-morbidities/other
Therapy compliance	Takes medication as prescribed
Side effects due to systemic therapy: CTCAE grade III-V	Cytopenias, infection, skin reaction, pneumonitis, cough, dyspnea, or other lung toxicity, oesophagitis, mucositis oral, nausea, vomiting, diarrhea, constipation, or other GI toxicity, neuropathy, tinnitus, hearing impaired, or other neurologic toxicity, kidney injury. Including date
Follow-up	
Outcomes	
Lost to follow-up	No declaration/hospital activities in the past 6 months
Death	Yes, no
Death Date	dd-mm-yyyy
Death Cause	Lung cancer related, treatment related, other
Tumor progression/recurrence	Yes, no
Tumor progression/recurrence Method	Clinical, imaging (chest x-ray, CT scan), PA proven, other
Tumor progression/recurrence Date	dd-mm-yyyy

Response	Complete response, partial response, stable disease
Response date	dd-mm-yyyy
Imaging	Method and date (frequency of progression measurement)
PROMs	
Quality of life	EQ-5D
Overall health/quality of life	EORTC QLQ-C30, questions 29, 30
Social functioning	EORTC QLQ-C30, questions 26, 27
Physical functioning	EORTC QLQ-C30, questions 1 to 5
Emotional functioning	EORTC QLQ-C30, questions 21 to 24
Cognitive functioning	EORTC QLQ-C30, questions 20, 25
Role functioning	EORTC QLQ-C30, questions 6,7
Fatigue	EORTC QLQ-C30, questions 10, 12, 18
Pain	EORTC QLQ-C30, questions 9, 19 + EORTC QLQ-LC29, questions 40-42
Shortness of breath	EORTC QLQ-C30, question 8 + EORTC QLQ-LC29, questions 33-35
Cough	EORTC QLQ-C29, question 31
Nausea/vomiting	EORTC QLQ-C30, questions 14, 15
Lack of appetite	EORTC QLQ-C30, question 13
Trouble sleeping	EORTC QLQ-C30, question 11
Constipation	EORTC QLQ-C30, question 16
Diarrhea	EORTC QLQ-C30, question 17
Allergic reaction	EORTC QLQ-C29, question 43
Financial impact	EORTC QLQ-C30, question 28