Call: HORIZON-HLTH-2021-TOOL-06
Topic: HORIZON-HLTH-2021-TOOL-06-03
Funding Scheme: HORIZON Research and Innovation Actions (RIA)

Grant Agreement no: 101057062

AI powered Data Curation & Publishing Virtual Assistant

Deliverable No. 4.3
Update to Annotation guidelines, tools & training

Contractual Submission Date: 30/04/2023

Actual Submission Date: 31/05/2023
Delayed to ensure alignment with Reference Ontology – describing the concepts and relations to be used for annotation – and included in D2.1 delivered on 31/05/2023.

Responsible partner: P7-Medical University of Graz (MUG)

Approval by the European Commission Pending

Funded by
the European Union

Grant agreement no.	101057062
Project full title	AIDAVA - AI powered Data Curation & Publishing Virtual Assistant

Deliverable number	**D4.3**
Deliverable title	**Update to Annotation guidelines, tools & training**
Type[1]	DEM
Dissemination level[2]	PU
Work package number	WP4
Work package leader	P7-MUG
Author(s)	Markus Kreuzthaler, Sareh Aghaei (MUG), Kris Collins, Stefan Schulz (AVER), Petros Kalendralis (Maastro), Kristian Kankainen (NEMC); Isabelle de Zegher (b!lo)
Keywords	manual annotation tools, manual annotation guidelines

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 or the European Health and Digital Executive Agency (HaDEA).

Neither the European Union nor the granting authority can be held responsible for them.

Part of the deliverable resample general ontology-aware annotation principles which is set forth together with the Manchester NLP group gnTEAM.

Document History

Version	Date	Description
V01	31.05.2023	Deliverable document

[1] **Type**: Use one of the following codes (in consistence with the Description of the Action):

R:	Document, report (excluding the periodic and final reports)
DEM:	Demonstrator, pilot, prototype, plan designs
DEC:	Websites, patents filing, press & media actions, videos, etc.

[2] **Dissemination level**: Use one of the following codes (in consistence with the Description of the Action)

PU:	Public, fully open, e.g. web
SEN:	Sensitive, limited under conditions of the Grant Agreement

Table of Contents

List of Abbreviations and definitions

The abbreviations used in the deliverable are based on the AIDAVA Glossary.

Abbreviation	Full Name
AI	Artificial Intelligence
API	Application Programming Interface
ATC	Anatomical Therapeutic Chemical
CDISC	Clinical Data Standards Consortium
CEN	Committee European de Normalisation
CRO	Contract Research Organisation
CVD	Cardio-vascular disease
DGA	Data Governance Act
DI	Data Intermediary
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DMP	Data Management Plan
DPIA	Data Protection Information Assessment
EHDS	European Health Data Space
EHR	Electronic Health Record
EMA	European Medicine Agency
ETL	Extraction Transformation Load
FAIR	Findable Accessible Interoperable Reusable
FHIR	Fast Healthcare Interoperability Resources
G1	“Generation 1 tools” of AIDAVA (early phase of the project)
G2	“Generation 2 tools” of AIDAVA (later phases of the project)
GA	General Assembly
GDPR	Global Data Protection Regulation
GP	General Practitioner
HCP	Health Care Provider
HDI	Health Data Intermediary
HIMSS	Health Information Management System Society
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7

ICD	International Classification of Disease
ICF	Informed Consent Form
IPR	Intellectual Property Rights
IPS	International Patient Summary
IRB	Institutional Review Board (ethics committee)
KER	Key Exploitable Results
KG	Knowledge Graph
LOINC	Logical Observation Identifiers Names and Code
MCN	Medical Concept Normalization
MDR	Medical Device Regulation
MUG	Medical University of Graz
ML	Machine Learning
NLP	Natural Language Process
NLP	Natural Language Processing
OMOP	Observational Medical Outcomes Partnership
PDEC	Plan for the Dissemination and Exploitation including Communication
PHD	Personal Health Data
PHI	Protected Health Information
PHKG	Personal Health Knowledge Graph
PII	Personal Identifiable Information
RDF	Resource Description Framework
RPA	Remote Process Assistant
RWD	Real World Data
SAB	Sustainability Advisory Board
SC	Steering Committee
SDLC	Software Development Life Cycle
SDTM	Study Data Tabulation Model
SHACL	Shapes Constraint Language
ShEx	Shape Expressions
SNOMED CT	Systematised Nomenclature of Medicine Clinical Terms
TA	Therapeutic Area
TRL	Technology readiness levels

TTP	Trusted Third Party
UMLS	Unified Medical Language System

1. Executive Summary

Manual annotations of clinical narratives are crucial for the adoption and evaluation of Natural Language Processing (NLP) tools, which support an overall AI-assisted data curation approach within AIDAVA. For a symbolic representation of clinical entities of interest and the way how they are related, normalisations that use international standards like SNOMED CT, FHIR or LOINC are crucial. For this deliverable, we updated the first version of the manual annotation guideline (see *AIDAVA Deliverable D4.1*), where requirements for annotation tooling were formulated with respect to the AIDAVA use cases, together with some initial annotation instructions. Grounded on this requirement analysis, INCEpTION was chosen as an annotation tool after a rigorous investigation of available annotation software. A first manual annotation schema was developed and tested, with a focus on the use of SNOMED CT and FHIR for the normalisation of the types of clinical entities (as annotating them with terminology codes) referred to by clinical narratives. Within this preparation phase, INCEpTION was deployed on all three clinical sites (MUG, NEMC, MUMC), with a first version of a consolidated INCEpTION layer definition. A bi-weekly "train the trainers" session was started at the end of 2022, supporting a continuous transition into the piloting phase of the developed guideline, analysing example narratives and how they should be annotated according to the first version of the guideline.

Within the piloting phase lasting from January 2023 to May 2023, in communication with the responsible clinicians, relevant attributes were identified, and a selection of them was used for updating, testing and refinement of the annotation guideline. Annotators were recruited at all three different sites and their feedback was taken into account for the customization and technical set up of INCEpTION. Alignment with Deliverable D2.1 "Reference Ontology as a Global Data Sharing Standard" defining the AIDAVA Reference Ontology was identified as crucial, therefore this deliverable was postponed for one month from April to May 2023.

Building on the first version of the guideline delivered early January 2023, this updated descriptive guideline provides a comprehensive framework and detailed instructions to ensure accurate annotation of clinical narratives. It covers crucial aspects like data standardisation and best practices in annotation (including annotation tool, general principles, specific instructions, concrete examples, and quality control items), ensuring consistent, interoperable, and high-quality annotations. This is invaluable for effective knowledge graph construction, data analysis, and knowledge extraction as central requirements in AIDAVA.

The updated annotation instructions form the core of this deliverable, enabling to start the productive phase of the manual annotation. Manual annotation of texts is iterative and dynamic. It is, therefore, crucial to recognise potential updates and improvements that may arise during the productive phase. Factors that can contribute to the modifications and enhancements of the set of annotation instructions include active feedback from the annotation team, new insights into text phenomena that lead to annotator disagreements, updates in data requirements from use cases, and evolution of project objectives as a result of dissemination and communication activities during the project. To ensure consistency and minimise inconsistencies in the annotation work, a structured feedback mechanism is established, involving documenting any challenges or updates in a shared document, and conducting meetings with the annotation team to address any emerging insights or challenges.

2. Introduction

AIDAVA pursues the goal to represent health data of an individual in a consistent semantic model – more precisely a Personal Health Knowledge Graph constrained by a Reference Ontology – rooted in international standards for electronic health records (EHRs) committed to the FAIR (findable, accessible, interoperable, reusable) principles of data stewardship.

Large amounts of clinical information are only available as narrative content in electronic health records (EHRs). Clinical narratives are characterised by unstandardized language with contextualised jargon expressions, short forms, spelling variants, spelling errors, and typos [3].

Notwithstanding, international standards like SNOMED CT, LOINC, and FHIR promise interoperable and computable representations of clinical terms and the domain entities they denote. Important goals for standards-based interoperable systems have been addressed for a while, but they have only partially been met, even in advanced clinical computing environments:

- Univocal standardised representations of a given portion of clinical reality
- Identification of syntactically different but semantically identical or similar representations
- Transformation of narrative content into such representations

Bridging between human language and semantic standards requires state-of-the-art technology in natural language processing (NLP), machine learning, and particularly deep learning. These systems need data to be trained, in particular by semantically annotated clinical text corpora, performed by human annotators, with the goal to represent relevant patient information in standardised form within a knowledge graph. Such resources constitute the “fuel” for models that precisely and reliably convert the content of clinical narratives into interoperable expressions rooted in the AIDAVA Reference Ontology – building on SNOMED CT, LOINC and HL7 FHIR (see *AIDAVA Deliverable D2.1 – AIDAVA Reference Ontology as a Global Health Data Sharing Standard*). Both HL7 and SNOMED International explicitly recommend the use of these two standards and work on interoperability issues in regular meetings of the SNOMED on HL7 FHIR working group with biweekly meetings.

High-quality human annotations should approximate the following goals.

- With the same input text, different human annotators produce the same target representation.
- With different paraphrases of the same clinical content, different human annotators produce target representations for which semantic equivalence could be stated
- With the translation of the same clinical content to different human languages, different human annotators produce target representations for which semantic equivalence could be stated.

The rationale of the second iteration of this deliverable document is to align the set of principled annotation guidelines together with the first version of the proposed AIDAVA guideline, focused on the Breast Cancer (BC) and Cardiovascular Disease (CVD)[3] use cases, described in *AIDAVA Deliverable 1.1 – Description of Use cases*).

[3] More specifically Myocardial Infarction

3. Prerequisites to Annotation

3.1 Data in Scope

To evaluate this updated version of the annotation guideline, a prioritisation was performed by the clinicians among the data elements identified as necessary for addressing the two use cases. This prioritisation was taken to exemplify how these data elements are represented in clinical texts, which then leads us to define how they should be annotated according to this guideline.

- **For the BC use case**: 001 cTNM; 002 ypTNM; 003 Type of surgery; 004 Neoadjuvant chemotherapy; 005 Adjuvant radiotherapy; 006 Adjuvant radiotherapy dose.
- **For the CVD use case**: 001 Date of birth; 002 Sex; 003 Diabetes mellitus; 004 Stroke; 005 Myocardial infarction; 006 Carotid artery disease; 007 Symptom onset, date/time; 008 Peripheral arterial disease; 009 Creatinine; 010 Total Cholesterol; 011 LDL cholesterol; 012 HDL cholesterol; 013 C-reactive protein; 014 Systolic blood pressure; 015 Smoking; 016 Aspirin used; 017 P2Y12 inhibitors used; 018 Oral anticoagulants used.

Out of the listed elements, selected items were used to guide the data-driven manual annotation examples via INCEpTION. Important for the selection was the existence of needed attributes within semi-structured, non-standardized clinical text documents.

The complete list of Data in scope has been developed with the clinical experts working on the two use cases: a list of 80+ data elements has been defined for the BC use case and 50+ data elements for the CVD use case.

In addition, the overall curation process, including the utilisation of NLP tools, is anticipated to facilitate the extraction of pertinent information associated with the International Patient Summary (IPS), which will be made available to each patient. The IPS, encompassing a diverse range of data elements, represents a broad scope of information on a patient's conditions, vital signs, medications, observations, procedures, diagnostic reports, specimens and other relevant factors. The annotation guideline and the NLP tools will be employed diligently to ensure comprehensive coverage of the IPS, enabling a comprehensive case, detailed summary that supports the broader objectives of the project and enhances the quality of care provided to patients.

All the data in scope are related to concepts defined in the AIDAVA Reference Ontology introduced below.

3.2 AIDAVA Reference Ontology

The development of the AIDAVA Reference Ontology (*AIDAVA Deliverable D2.1 – AIDAVA Reference Ontology as a Global Health Data Sharing Standard*) is an iterative process including ideation, requirement analysis, design, development and active maintenance. The use cases introduced above, the narrative text annotation process, along with Task 2.1 for automation of the data extraction and Task 4.2. for data quality constraints, provide specific requirements on the ontology coverage and ontology management process.

3.2.1 Coverage

In terms of the ontology coverage, following the list of requirements and a review of key initiatives described in Deliverable 2.1, some standards clearly emerged in terms of (potentially lawful[4]) use and

[4] The standards selected for AIDAVA Reference Ontology are included in the EEHRxF recommendations from the Europe Commission [1] and may become mandatory by law as part of the EHDS regulation [2].

acceptance across the healthcare community in Europe and are included as priority standards to be included in the AIDAVA Reference Ontology.

- SNOMED CT as an overarching "hub"
- LOINC for laboratory procedures
- HL7 FHIR information model components, starting with the General Purpose Data Types and including profiles from HL7 FHIR IPS related to the use cases.

HL7 FHIR is becoming the de facto data exchange standard for Health information (as part of EEHRxF [1]). While being an exchange standard, FHIR also includes ontological components that have been considered by the annotation team.

According to FHIR, all contextual information at the instance level should be consistent with the value sets proposed by the FHIR specifications, whereas all ontological information (referred to by code in a FHIR resource) should be provided by an ontology such as SNOMED CT.

For instance, laterality, aetiology, or chronicity of a condition is ontological, as well as dose form and strength of a drug. The same is true for anatomical location of a surgical procedure. All this information should be expressed by SNOMED CT.

In contrast, diagnostic certainty, subject relationship (patient or family), temporal contexts of conditions and procedures are contextual, as well as for units of measurement. FHIR proposes existing HL7 value sets (e.g. with the value "differential" for diagnostic certainty). These values, however, largely overlap between FHIR and SNOMED. This is why both communities are currently working on mappings between FHIR (HL7) value set elements and SNOMED CT codes.

In order to avoid that annotators have to deal with different ontologies, we propose the restriction to SNOMED CT concepts even in these cases where FHIR suggests HL7 values sets. In these cases, SNOMED CT - HL7 mappings are maintained in the background.

3.2.2. Maintenance

While developing the ontology, it has become obvious that we need to represent specific concepts and predicates[5] that are necessary to represent the extracted data during onboarding and annotation of the narratives. This requires definition of an ontology governance process and release executed during the project lifetime. The process covers the maintenance of the so-called AIDAVA Dataset (the collection of all data elements used) and the AIDAVA Reference Ontology itself. Details on the established process are provided in Section 3.4 Governance of D2.1 – Reference Ontology.

[5] Also known as "relations", "linkage concepts in SNOMED CT", "object properties" and "data properties" in OWL

3.3 Infrastructure

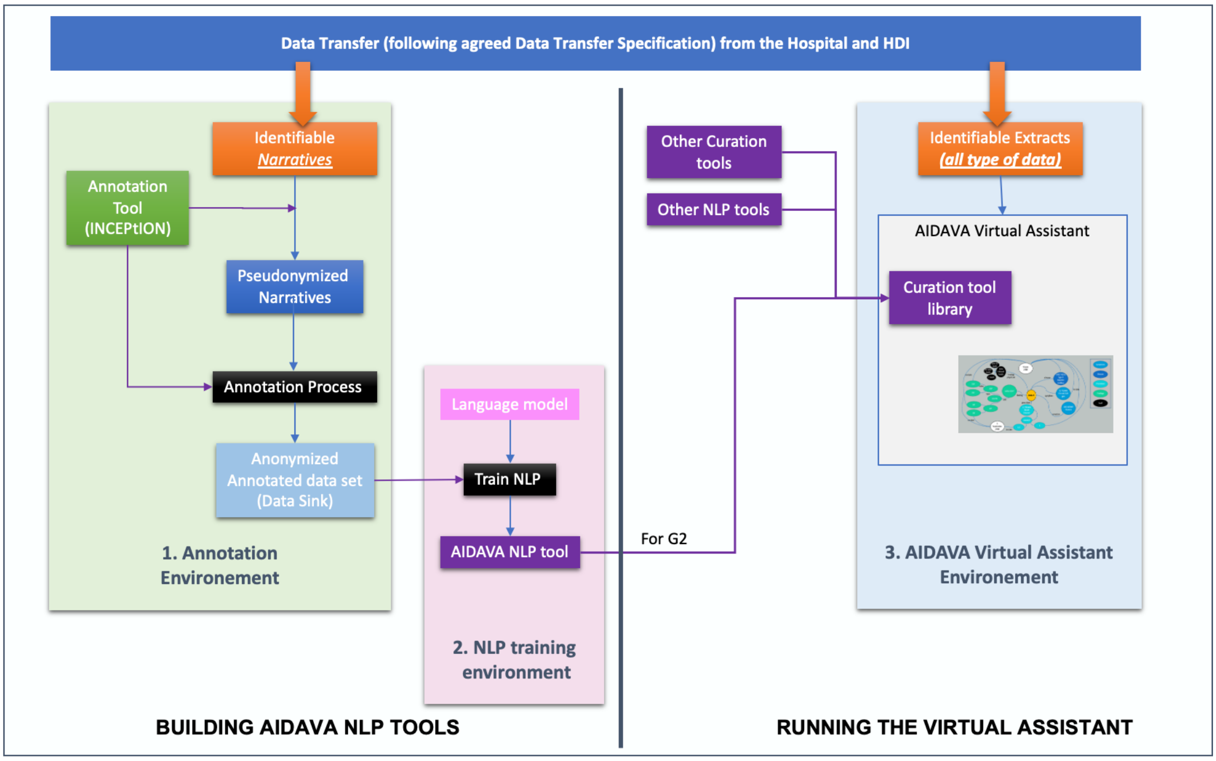

Data Transfer (following agreed Data Transfer Specification) from the Hospital and HDI
Identifiable Narratives
Annotation Tool (INCEPtION)
Pseudonymized Narratives
Annotation Process
Anonymized Annotated data set (Data Sink)
1. Annotation Environement
Language model
Train NLP
AIDAVA NLP tool
2. NLP training environment
BUILDING AIDAVA NLP TOOLS
Other Curation tools
Other NLP tools
For G2
Identifiable Extracts (all type of data)
AIDAVA Virtual Assistant
Curation tool library
3. AIDAVA Virtual Assistant Environement
RUNNING THE VIRTUAL ASSISTANT

Figure 1: Core technical environment blocks with an integrative view of the manual annotation process.

Annotation Environment. The manual annotation environment with the on-premises deployment of INCEpTION focuses on a use case-oriented normalisation of patient information in conformance with international standards. The result of the manual annotation process is exported to a data sink supporting different data formats (see Section 4.4) for the adaptation and the training of NLP components.

NLP Training Environment. The intended setting is that adaptations of NLP components that require the export of manual annotations using INCEpTION are processed within the same platform. At this stage, tools will need to come to the data, and annotated datasets will not leave the environment. For an operational setting this implies the granting of access rights from outside to those who will use the INCEpTION export for NLP training and adaptation. The access rights have to follow regulations at the three clinical sites.

AIDAVA Virtual Assistant Environment. Within G2, the AI-adapted tools should utilise manually annotated datasets to support and enhance data curation tools available through the AIDAVA virtual assistant. Inside this framework, the data sink of the data curation workflow is a knowledge graph representation of patient-based information. It will take into account various modalities, different levels of structure, and different levels of ontology-based standardisation of relevant data.

3.4 Extraction process and data privacy compliance

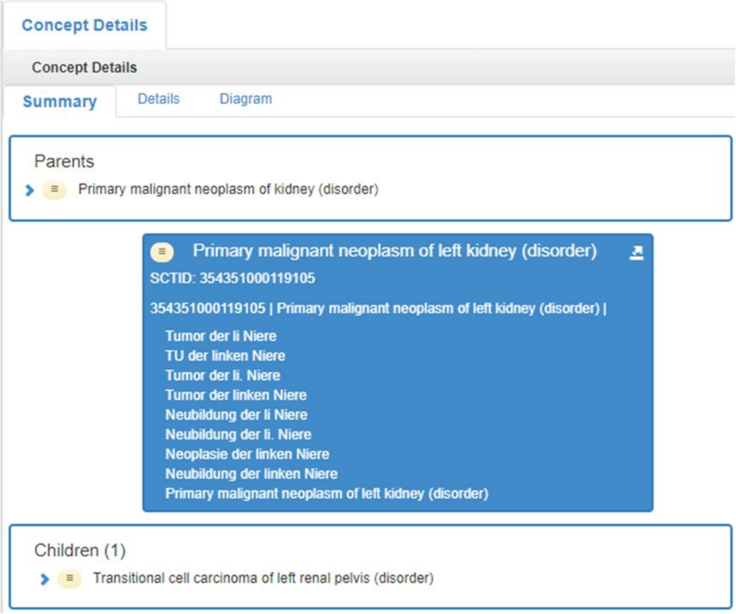
MUG (Medical University Graz) : Clinical narratives that cover the use cases are considered candidates for extracting the data elements of interest. These narratives are then exported and de-identified (see *AIDAVA Deliverable D4.1 – Section 4.2 Ethical committee approval and data handling*) by the mandated local medical data management team into a secure data lake. The export contains both the corresponding tables from the clinical information system and the rendered view as a document. From there, the de-identified narratives are imported into INCEpTION. INCEpTION is hosted via a virtualization service maintained by the Medical University of Graz, with granted project-oriented user access rights via VPN and 2-factor authentication. This setting is in compliance with the local IRB (ethical committee) approval.

NEMC (North Estonia Medical Center): The extraction process has been set up to ensure bidirectionality and thus annotation can later be amended back to the extraction source. Clinical narratives for both CVD and BC cohorts have been identified in a pseudonymized database maintained by NEMC for experimental development purposes. The database replicates structure and content from operational data, except that all identifiers have been obfuscated. The tools INCEpTION tool Averbis Health Discovery have been installed on-site in virtual computers accessible only to selected NEMC accounts and only via VPN. The clinical narratives together with the relevant data fields from a document point of view have been transferred to the INCEpTION tool. In total, 1223 documents (across 3 document types) for the BC use case and 8216 documents (across 4 document types) for the CVD use case have been made available for annotation.

UM/MAASTRO: The BC narratives will be delivered and stored in a secure drive hosted by MAASTRO clinic (working under UM/MUMC). This drive will be named according to the research protocols of MAASTRO for IRB approved studies. For instance, each research project starts with “P” followed by an 8-digit code assigned by the IT department of MAASTRO. The narratives will be extracted from the EHR system of MAASTRO where clinicians and data managers register patient information accompanied by the diagnostic reports. The pseudonymization of the reports is taking place before the data delivery of the narratives in the drive by the MAASTRO IT department. The clinical narratives are imported to INCEpTION via a secure VPN connection of MAASTRO with the corresponding annotator, who will have access to the above-mentioned drive.

UM/MUMC (University of Maastricht/ Maastricht University Medical Center) : The clinical narratives for the CVD use case will be delivered and stored in a secure server hosted by the MUMC data management service provider named “DataHub”. It provides data management services for (non-)clinical studies in both the Faculty of Health, Medicine and Life Sciences of Maastricht University and Maastricht UMC+. Their role is that of a data broker who enables the reuse of data by researchers in the university and the hospital. The data will be anonymized by the IT department of the MUMC before the delivery to the server hosted by DataHub. The clinical narratives will be imported to INCEpTION via a secure VPN connection of UM with the corresponding annotator, who will have access to the above-mentioned server.

3.5 Translation of narratives and terminologies



Parents

› ≡ Primary malignant neoplasm of kidney (disorder)

≡ Primary malignant neoplasm of left kidney (disorder)

SCTID: 354351000119105

354351000119105 | Primary malignant neoplasm of left kidney (disorder) |

Tumor der li Niere
TU der linken Niere
Tumor der li. Niere
Tumor der linken Niere
Neubildung der li Niere
Neubildung der li. Niere
Neoplasie der linken Niere
Neubildung der linken Niere
Primary malignant neoplasm of left kidney (disorder)

Children (1)

› ≡ Transitional cell carcinoma of left renal pelvis (disorder)

Figure 2: Example of additional synonyms for a concept defined in SNOMED CT.

Although translation tasks of clinical narratives and semantic resources had been raised in the original proposal as part of WP4 (T4.3 "Manual Annotation of text documents in 3 languages") in WP 4 meetings the following decisions were taken:

- Clinical narratives are not translated all, because all patient-related data processing will be done on premise, including the annotations as described in Section 4.1.
- Terminology translation in a broader sense is only done regarding the addition of synonyms to the terminologies used for pre-annotation, particularly as external sources to be imported into the Averbis Health Discovery tool, particularly for the German language, see Fig. 2. The extent to which term translations will be used for the Dutch and Estonian sources is still subject of further discussion.
- For a common understanding of the annotation principles, particularly in the context of annotator training, documentation and publication certain short example text passages and data elements will be translated into English and shared among the partners.

4. INCEpTION

For the use of INCEpTION see the INCEpTION documentation [4] and Section 5.1 in *AIDAVA Deliverable D4.1 – Annotation guideline, tools & training, "Tool configuration and use"*. In addition, the following three sections highlight again a view of the INCEpTION annotation scheme, technical considerations regarding the pre-annotations and additional useful INCEpTION settings.

4.1. Annotation schema

Due to its importance, the annotation schema is highlighted here again and its current version 001 is shared with the partners exploiting the MUG NextCloud functionality. The annotation schema in INCEpTION is split into layers [5]. In the first iteration, we define two layers: a Custom MCN (Medical Concept Normalization) layer and a Relation layer.

Custom MCN layer. The Custom layer has the following features, with an emphasis on the possibility to annotate the normalised form according to SNOMED CT. Moreover, ICD and LOINC codes need to be annotated, and the mapping is performed in a later stage of the annotation process.

Feature	Description
Concept : [String]	Standardised description, e.g. SNOMED CT fully specified name ("241998008 |Cardiovascular decompression injury (disorder)|"). The line contains the ID, the preferred name and the semantic tag in parentheses.
Short : [String]	Short readable form for visualisation within INCEpTION.
Comment : [String]	Additional comments for this concept annotation.

Table 1: Feature definitions of the custom layer.

Relation layer. Predicates, which relate concept annotations of interest in combination with their normalised form. The following features are therefore defined within the Relation layer.

Feature	Description
Relation : [String]	Standardised description, e.g. SNOMED CT fully specified name ("363698007 |Finding site (attribute)|"). The line contains the ID, the preferred name and the semantic tag in parentheses.
Short : [String]	Short readable form for visualisation within INCEpTION.
Comment : [String]	Additional comments for this predicate annotation.

Table 2: Feature definitions of the Relation layer.

4.2 Pre-annotations

4.2.1 Averbis Health Discovery

To speed up the manual annotation process under the condition of an unbiased stable inter-rater agreement [6], pre-annotations [7] shall be integrated into the manual annotation tool, which has to be manually confirmed by the user if correct or not. The pre-annotations have not to be necessarily involved in an active learning approach [8] but should support the manual annotator in finding and confirming the elements of the annotation vocabulary. To support pre-annotations, the existing component repository within the Averbis Health Discovery text mining tool should be utilised.

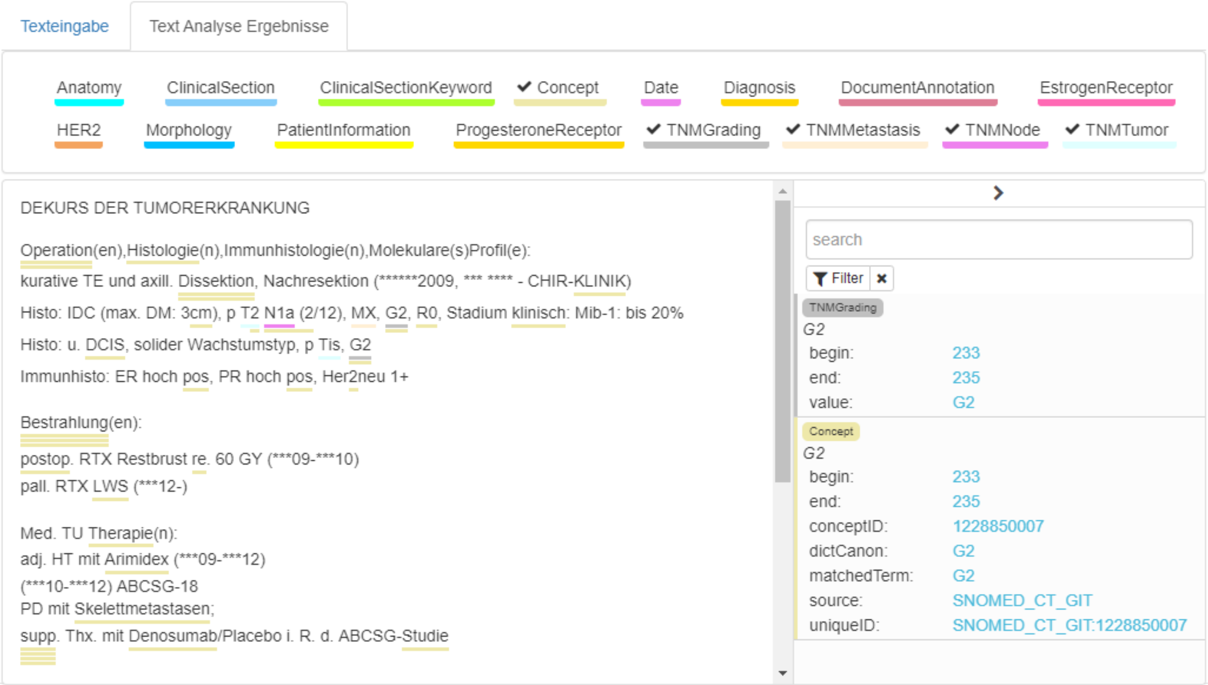
Available components within the **Averbis Health Discovery pathology pipeline**
Generic terminology annotator (SNOMED CT, requires custom vocabulary), ClinicalSections, Laterality, PatientInformation, TNM, LabValues, Diagnoses, Topography, Morphology, GleasonScore, Enumerations, Negations, TumorStage, Receptors, Specimen, DiagnosisStatus, Disambiguation, PathologyDocumentationClassification, HealthPostprocessing

Available components within the **Averbis Health Discovery discharge pipeline**
Generic terminology annotator (SNOMED CT, requires custom vocabulary), ClinicalSections, Laterality, PatientInformation, Organizations, PhysicalTherapies, LabValues, Diagnoses, Procedures, Medications, Enumerations, Negations, DiagnosisStatus, Procedure, Disambiguation, MedicationStatus, HealthPostprocessing

For the integration of pre-annotations, a conversion from the type system used in Averbis Health Discovery to the INCEpTION type system created for AIDAVA has to be done (layer and feature definitions). Averbis Health Discovery provides specific python libraries [9] which support this conversion. An adaptation of these libraries supporting the AIDAVA annotation philosophy implemented in the INCEpTION annotation schema is under development. Bridging annotations from the Averbis Health Discovery component repository to the defined INCEpTION annotation schemas is currently in development. It is based on a combination of the Averbis annotators for diagnoses, lab parameters, drugs and personal health information, combined with the "Generic terminology annotator" using the SNOMED CT German Interface Terminology[6]. It is expected to have pre-annotation functionality for German texts available via this process in Q3 2023. The utility of the tool for the other languages still has to be tested.

A pre-condition for the support of pre-annotation is a clear definition of the consolidated annotation model (layer and feature definitions), which has to be used at all clinical partner sites for consistency. The INCEpTION annotation schema definition is currently populated at version 001, see Section 4.1.

[6] German Interface Terminology for SNOMED CT



Texteingabe | Text Analyse Ergebnisse

Anatomy | ClinicalSection | ClinicalSectionKeyword | ✓ Concept | Date | Diagnosis | DocumentAnnotation | EstrogenReceptor | HER2 | Morphology | PatientInformation | ProgesteroneReceptor | ✓ TNMGrading | ✓ TNMMetastasis | ✓ TNMNode | ✓ TNMTumor

DEKURS DER TUMORERKRANKUNG

Operation(en),Histologie(n),Immunhistologie(n),Molekulare(s)Profil(e):
kurative TE und axill. Dissektion, Nachresektion (******2009, *** **** - CHIR-KLINIK)
Histo: IDC (max. DM: 3cm), p T2 N1a (2/12), MX, G2, R0, Stadium klinisch: Mib-1: bis 20%
Histo: u. DCIS, solider Wachstumstyp, p Tis, G2
Immunhisto: ER hoch pos, PR hoch pos, Her2neu 1+

Bestrahlung(en):
postop. RTX Restbrust re. 60 GY (***09-***10)
pall. RTX LWS (***12-)

Med. TU Therapie(n):
adj. HT mit Arimidex (***09-***12)
(***10-***12) ABCSG-18
PD mit Skelettmetastasen;
supp. Thx. mit Denosumab/Placebo i. R. d. ABCSG-Studie

search

Filter ✖

TNMGrading
G2
begin: 233
end: 235
value: G2

Concept
G2
begin: 233
end: 235
conceptID: 1228850007
dictCanon: G2
matchedTerm: G2
source: SNOMED_CT_GIT
uniqueID: SNOMED_CT_GIT:1228850007

Figure 3: Pre-annotation examples to be integrated into INCEpTION.

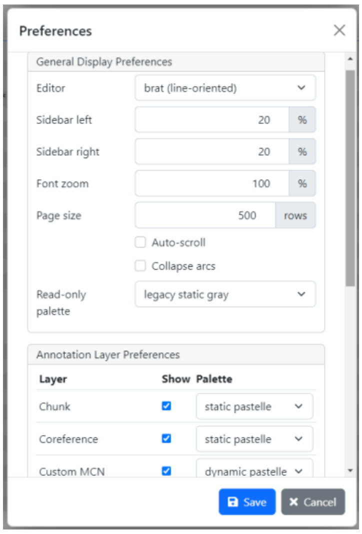
4.2.2 INCEpTION recommender functionality

The recommender functionality supported, which we also refer to as pre-annotation capabilities to assist the human annotator in general in this document, is described by INCEpTION for the suggested String Matcher [10] as follows: “The string matching recommender is able to provide a very high accuracy for tasks such as named entity identification where a word or phrase always receives the same label. If an annotation is made, then the string-matching recommender projects the label to all other identical spans, therefore making it easier to annotate repeated phenomena.” In combination with gazetteers containing specific terms assigned with a label at the end, the recommender can be pre-configured with terms of interest and their corresponding label.

This is useful for specific terms frequently occurring in documents without a need for a disambiguation in context. A full documentation of INCEpTION regarding recommender capabilities can be found in the User Guide [10]. Configuration of this functionality is left open, to the corresponding clinical side, but can be integrated at any point in the manual annotation process.

4.3 Additional configurations

In addition to the base configuration, see Deliverable D4.1 Section 5.1 “Tool configuration and use”, the following view customization should be used, Preferences : General Display Preferences : Editor : brat (line-oriented); Preferences : Annotation Layer Preferences : Custom MCN : dynamic pastelle.



Preferences

General Display Preferences

Editor	brat (line-oriented)	
Sidebar left	20	%
Sidebar right	20	%
Font zoom	100	%
Page size	500	rows
	□ Auto-scroll	
	□ Collapse arcs	
Read-only palette	legacy static gray	

Annotation Layer Preferences

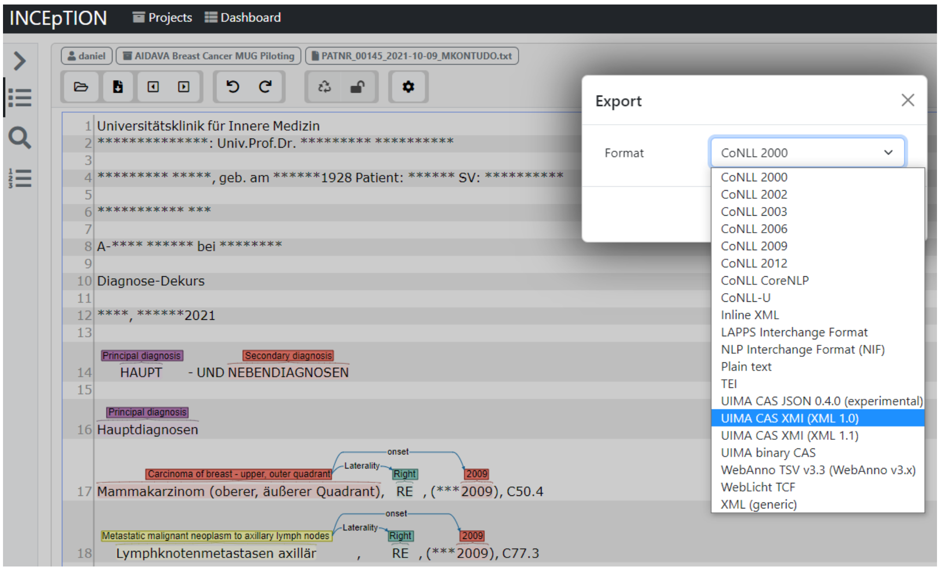
Layer	Show	Palette
Chunk	☑	static pastelle
Coreference	☑	static pastelle
Custom MCN	☑	dynamic pastelle

Save | Cancel

Figure 4: Preferences for the document view in INCEpTION.

4.4 Export format

INCEpTION provides different output formats, which can be used for NLP adaptation and training. An overview is given in the official documentation. Annotation : Chosen document : Export.



Universitätsklinik für Innere Medizin
**************: Univ.Prof.Dr. ********* **********

********* *****, geb. am ******1928 Patient: ****** SV: **********

*********** ***

A-**** ****** bei ********

Diagnose-Dekurs

****, ******2021

HAUPT - UND NEBENDIAGNOSEN

Hauptdiagnosen

Mammakarzinom (oberer, äußerer Quadrant), RE , (***2009), C50.4

Lymphknotenmetastasen axillär , RE , (***2009), C77.3

Format: CoNLL 2000

Figure 5: Possible export formats supported by INCEpTION.

5. Annotation Instructions

The annotation process involves training annotators to identify spans of text in clinical narratives and annotate them by the most appropriate clinical concepts using a subset of preferred items. The prioritised items for CVD and BC are listed in Section 3.1. Detailed annotation instructions are given in Section 6.3. The training of annotators is focused on these preferred items.

After this training, the annotators start with the annotation of the CVD and BC narratives using the concepts and predicates made available for the respective use case.

5.1. Introduction

5.1.1 Naming and graphical conventions

An **Annotation Guideline** for narrative text is a set of principles, instructions and rules that define the process of manually annotating specific elements in a narrative. **Narratives** in our scope are textual parts of electronic health records such as documents or textual entries in database fields. Annotation means the practice of interpreting the text, word by word, and to identify spans that correspond to **Concepts** and **Predicates**. Annotations are performed by human **Annotators**, i.e. specifically trained domain experts. Annotation, in a broader sense, can also mean the same task done by algorithms.

In AIDAVA, annotation is restricted to semantic annotations. **Concepts** correspond to entities of meaning as provided by clinical terminologies. They characterise (mostly in terms of instantiation) the referents, i.e. the things in the domain (a patient and everything around in a clinical treatment episode) to which the words, multiword expressions, abbreviations and even subword strings in the text point to. The sequence of characters to which an annotation is ascribed is named **Span**. [7]

AIDAVA also supports **Predicate** annotations. Predicates are relational elements that link two concept annotations.[8]

The AIDAVA annotation guideline commits to a set of **Annotation Principles** (5.2.). These general principles are related to **Symbols** of an **Annotation Language** (5.3), which are used in **Annotation Tools** (section 4) From the annotation principles, the use cases, and the properties of the tools, a set of **Annotation Instructions** (5.4) is derived. In chapter 6, we will use screenshots from the INCEpTION tool. In this section we visualise our prototypical annotation examples as follows:

- Text is centred, bold, 9pt
- Concept annotations have yellow background colour
- Predicate annotations have green background colour
- Meta annotations have light blue background colour
- The domain entity (or “subject”) of a predicate is shown by a centred arrow pointing north-west or north-east (↖↗)
- The range entity (or “object”) of a predicate is shown by a centred arrow pointing south-west or south-east (↙↘)

<table>
<tr><td>↙</td><td colspan="3">anno:dueTo</td><td colspan="2">↖</td></tr>
<tr><td></td><td colspan="2"></td><td colspan="3">↙ 263502005 |Clinical course (attribute)| ↖</td></tr>
<tr><td>112674009 |Fibrosis (morphologic abnormality)|</td><td colspan="2"></td><td>70232002 |Frequent (qualifier value)|</td><td colspan="2">87628006 |Bacterial infectious disease (disorder)|</td></tr>
<tr><td><b>Fibrosis</b></td><td><b>due</b></td><td><b>to</b></td><td><b>frequent</b></td><td><b>bacterial</b></td><td><b>infections</b></td></tr>
</table>

[7] We refrain from the words “entity”, or “named entity”, which are used in several, partly contradicting senses.

[8] Like a predicate in a simple sentence links subject and object

5.1.2 Standardised scoping

The proposal of an annotation language does not contradict the fact that there are multiple terminologies and ontologies in health care. However, in order to avoid that annotators have to deal with different ontologies, AIDAVA proposes a restriction to concepts defined in the AIDAVA Reference Ontology as described in AIDAVA Deliverable 2.1 and predicates described below (5.3.2).

The Reference Ontology is based on SNOMED CT as core ontology, as recommended through several initiatives (TEHDAS and EEHRxF – see Deliverable D2.1 for in depth discussion). Addition of ontological aspects from the information model underlying FHIR are done by mappings (see *Table 5*) , as well as, inclusion of domain-specific ontologies will be done following the governance process described in Deliverable 2.1. We should be aware of possible gaps regarding the representation of observables, for which we might consider the addition of LOINC codes. The standard approach is, however, the use of SNOMED CT “measurement” and “count” procedure concepts as the recommended strategy to use SNOMED CT alone.

5.2. General principles

Regarding the options (Section 2.3), the following preferences are suggested for annotating clinical text.

All annotations are rooted in document spans, i.e. one or many words. Only semantic annotations, coreference annotations, and potentially, annotations at a document or document section level are performed. No syntactic annotations are performed. The breath of the span is given by the concept in the ontology. Due to precoordination in SNOMED CT, a complex situation can be precisely expressed by a single annotation, which may catch the exact meaning of a whole sentence.

	62564004 |Concussion with loss of consciousness (disorder)|
The patient has	a brain concussion, but she is conscious

Annotations on a subword level are allowed where the word is clearly composed by parts that could alternatively be separated by spaces or hyphens (here between “pT1” and “N1”):

1228957006 |American Joint Committee on Cancer pT1 (qualifier value)|	1229951001 |American Joint Committee on Cancer pN1 (qualifier value)|
pT1N1	

Semantic annotations use the whole depth of the annotation vocabulary. This means that always the annotators choose the concept that comes closest to the passage to be annotated. The delineation of text passages follows the ontology. For instance,

172732009 |Implantation of intracochlear prosthesis (procedure)|
Insertion of an intracochlear implant

instead of

71388002 |Procedure (procedure)|
Insertion of a cochlear implant

(which is easily inferable from the above, in case that only annotations at a high level are required)

Notwithstanding, annotation vocabularies may be constrained by use cases, which limit annotation depth to the granularity of the use case’s sub-ontology. For instance, the annotation vocabulary subset for cancer may not require subclasses of comorbidity types that are known to be irrelevant for cancer. Given that the subgraph under 56265001 |Heart disease (disorder)| has been pruned in such an application oriented ontology, the following annotation is correct

<table>
<tr><td>56265001 |Heart disease (disorder)|</td></tr>
<tr><td>Mitral stenosis</td></tr>
</table>

Nevertheless, we recommend using the whole scope of the ontology and annotating with the highest possible precision:

<table>
<tr><td>79619009 |Mitral valve stenosis (disorder)|</td></tr>
<tr><td>Mitral stenosis</td></tr>
</table>

The rationale is that the former representation can be inferred from the latter one. The decision for low-granularity annotations or for the exclusion of annotations for certain kinds of entities is driven by the optimal utilisation of annotation resources in light of the use cases for which the annotation is done.

Facts, i.e. the connection of annotated spans with a predicate, are annotated as long they can be unambiguously derived from the text without any additional interpretation. This example suggests causality, but it is nevertheless annotated with a temporal predicate, because only a temporal predicate is stated in the text.

<table>
<tr><td colspan="3">↗ anno:after ↘</td></tr>
<tr><td>25064002 |Headache (finding)|</td><td></td><td>110030002 |Concussion injury of brain (disorder)|</td></tr>
<tr><td><b>Headache</b></td><td><b>following</b></td><td><b>brain concussion</b></td></tr>
</table>

This is a counterexample, in which the association between disease and symptom is sufficiently expressed by the preposition “in”

<table>
<tr><td colspan="3">↗ anno:dueTo ↘</td></tr>
<tr><td>162049009 |Left flank pain (finding)|</td><td></td><td>45816000 |Pyelonephritis (disorder)|</td></tr>
<tr><td><b>Left flank pain</b></td><td><b>in</b></td><td><b>pyelonephritis</b></td></tr>
</table>

5.3. Symbols of the annotation language

Annotations use elements from a predefined annotation language. In AIDAVA, we distinguish between symbols for concepts and symbols for predicates. For concepts, we use most of SNOMED CT with some restrictions as explained below. For predicates, we use a closed set of relations derived from SNOMED CT and FHIR (namespace “anno”). The SNOMED annotations are characterised by the use of the typical SNOMED syntax that coordinates identifier and label using the pipe character.

5.3.1. Concepts

By “concepts” we understand the units of non-relational meaning from SNOMED CT (except the descendants of 900000000000441003 |SNOMED CT Model Component (metadata)|). Whether the use of a SNOMED CT concept in an annotation denotes a particular meaning or a universal meaning is not distinguished at the annotation level (“John Doe has asthma” vs. “John Doe is examined for asthma"), but may be subject to downstream interpretations. Top-level concepts in SNOMED CT (the heads of the SNOMED CT hierarchies) are also referred to by "Semantic Types".

SNOMED CT often presents ambiguities, i.e. two or more concepts with same or similar names. This requires guidance in terms of preference rules. We therefore distinguish between focus concepts and non-focus (supportive) concepts, the former being preferred in case of doubt. Focus concepts are ideally fully expressive. They often stand alone and do not modify other concepts. Many focus concepts are partly or fully defined using axioms with predicates and restrictions as prescribed by the SNOMED concept model. Focus concepts come typically from the hierarchies Clinical Conditions (SNOMED findings / disorders / events), Procedures, Observables, Staging and scales, Pharma products. Table 3 provides an overview of the SNOMED CT hierarchies and their use for annotation, Table 4 shows their priorities.

Table 3: SNOMED Hierarchies, their use in annotation, their hierarchy tags (expression in brackets appended to the fully specified name) and their use for annotation.

SNOMED CT hierarchy	Tags	Focus concepts	Use for annotation
Body structure	body structure, morphological abnormality, cell, cell structure		Morphology concepts only in those cases where no corresponding finding concept is available
Clinical finding	finding, disorder	✓	Concepts that correspond to negative fact statements and can easily be expressed with their positive correlate and a verificationStatus assertion should be avoided
Environment or geographical location	environment / location, environment, geographic location		
Event	event	✓	
Observable entity	observable entity	✓	Must have a quantitative or qualitative value
Organism (organism)	organism		
Qualifier value	qualifier value, …		
Pharmaceutical / biologic product	product	✓	Should always be given preference over the annotation with the substance alone
Physical force	physical force		
Physical object	physical object		
Procedure	Procedure, regime/therapy	✓	"Measurement" or "count" concepts treated like observables: Must have a quantitative or qualitative value
Record artifact	record artifact		Only to be used for document or section level annotations
Situation with explicit context	situation		To be avoided. Their meaning should be represented by predicates like verificationStatus and clinicalStatus
SNOMED CT Model Component	metadata		Linkage concepts are the basis of many alias predicates (namespace anno:)
Social context	social context, ethnic group, Lifestyle, Occupation, person, racial group, religion/philosophy, social status		Used, e.g. denoting the family relationship in a family history statement
Special concept	special concept		Not used
Specimen	specimen	✓	
Staging and scales	staging scale, assessment scale, tumor staging,	✓	Like observables. Must have a quantitative or qualitative value
Substance	substance		Use only if no appropriate product concept

Table 4: Priorities between SNOMED Hierarchies in case of ambiguity

First priority	Second priority	Third priority	Example
Core concept	Non-core concept		(default)

Body structure	qualifier value		Always give preference to terms with "Structure" in the fully specified term
disease	morphological abnormality		118600007 |Malignant lymphoma (disorder)| > 1163043007 |Malignant lymphoma (morphologic abnormality)|
finding	morphological abnormality		
observable	finding		75367002 |Blood pressure (observable entity)| > 392570002 |Blood pressure finding (finding)|
observable	disease		
substance	organism		
product	substance		735971005 |Fish (substance)| > 90580008 |Fish (organism)|
*	*	qualifier value	86273004 |Biopsy (procedure)| > 129314006 |Biopsy - action (qualifier value)|
specimen	substance		119297000 |Blood specimen (specimen)| > 256906008 |Blood material (substance)|
procedure	substance		59573005 |Potassium measurement (procedure)| > 88480006 |Potassium (substance)| 182764009 |Anticoagulant therapy (procedure)| > 372862008 |Anticoagulant (substance)|
procedure	Physical object		

Core concepts are typically related to supportive concepts via outgoing predicates (following the SNOMED concept model [11]), such as <Clinical condition, causative agent, Organism> or <Administration of drug or medicament, Direct substance, Substance>. Ideally, supportive concepts should only be used in case they are clinically important and not expressible as focus concepts, and when the interpretation of other parts of the text depends on them. Qualifier values and units of measurement are used only when related to other concepts.

5.3.2. Predicates and their definitions

Table 5 lists the "close to user" alias predicates to be used, together with their origin in SNOMED CT and FHIR. The main reason for specific predicates is to shield users away from the complexity of the internal wiring of SNOMED CT and FHIR, including the possibility of redundant representations. The suggested predicates are expected to cover 95% of the relational assertions needed. Where a new predicate is required that is not in this a new one can be suggested by the annotators, for which a rooting in FHIR, SNOMED or both can subsequently be sought by the guideline maintainers.

Regarding SNOMED CT predicates (linkage concepts in SNOMED CT, object properties in OWL) , ambiguous mappings (e.g. "site") – which are deliberately introduced to keep the set of predicates small – can be disambiguated in terms of finding or procedure sites, regarding their domain type. But they also support redundant representations, e.g. by mapping "site" also to FHIR representations in which it maps to bodySite.

"INV" indicates an inverse predicate, '||' the concatenation operator. "<<" specifies the allowed values for domain and range, according to the SNOMED Expression Constraint Syntax (ECL) [12].

Range restrictions with specific value sets (including the mapping from SNOMED CT codes to HL7-FHIR value sets, are given in Table 5 (following SNOMED International mapping recommendations [13]).

Table 5: Predicate values, their domain and range restrictions and their rooting in SNOMED CT and FHIR

<table>
<tr><th><b>Alias Predicates</b>
(namespace: anno:)</th><th><b>Domain</b>
(as SNOMED ECL expressions)</th><th><b>Relational expression</b>
(rooted SNOMED or FHIR linkage concepts / slots / relations)</th><th><b>Range</b>
(as SNOMED ECL expressions)</th></tr>
<tr><td>abatementTime</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>INV(Condition.code) || Condition.abatement.dateTime</td><td>dateTime</td></tr>
<tr><td>access</td><td><<71388002 |Procedure (procedure)| OR <<404684003 |Clinical finding (finding)|</td><td>260507000 |Access</td><td>362981000 |Qualifier value (qualifier value)|</td></tr>
<tr><td>actionStatus</td><td>763158003 |Medicinal product (product)| OR <<71388002 |Procedure (procedure)|</td><td>INV(MedicationAdministration.medication) || MedicationAdministration.status</td><td>Cf. Table 6</td></tr>
<tr><td>after</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>255234002 |After</td><td><<71388002 |Procedure (procedure)|</td></tr>
<tr><td>beginAge</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>INV(Condition.code) || Condition.onsetAge.quantity.value</td><td>decimal</td></tr>
<tr><td>beginAgeUnit</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>INV(Condition.code) || Condition.onsetAge.quantity.code</td><td>767524001 |Unit of measure (qualifier value)|</td></tr>
<tr><td rowspan="3">beginTime</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>INV(Condition.code) || Condition.onset.dateTime</td><td rowspan="3">dateTime</td></tr>
<tr><td rowspan="2"><<71388002 |Procedure (procedure)| |</td><td>INV(ServiceRequest.code) || ServiceRequest.occurrence.dateTime</td></tr>
<tr><td>INV(Procedure.code) || Procedure.occurrence.dateTime</td></tr>
<tr><td>clinicalStatus</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>INV(Condition.code) || Condition.clinicalStatus</td><td>Cf. Table 6</td></tr>
<tr><td>device</td><td><<71388002 |Procedure (procedure)|</td><td>363699004 |Direct device</td><td><<260787004 |Physical object (physical object)|</td></tr>
<tr><td>dosage</td><td></td><td></td><td></td></tr>
<tr><td>doseForm</td><td>373873005 |Pharmaceutical / biologic product (product)|</td><td>763032000 |Has unit of presentation (attribute)|</td><td>< 736542009 |Pharmaceutical dose form (dose form)|</td></tr>
<tr><td rowspan="2">dueTo</td><td rowspan="2"><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>246075003 |Causative agent</td><td><<410607006 |Organism (organism)| OR <<78621006 |Physical force (physical force)| OR <<105590001 |Substance (substance)|</td></tr>
<tr><td>42752001 |Due to</td><td><< 272379006 |Event (event)| OR << 404684003 |Clinical finding (finding)| OR << 71388002 |Procedure (procedure)|</td></tr>
<tr><td>endAge</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>INV(Condition.code) || Condition.abatementAge.quantity.value</td><td>decimal</td></tr>
<tr><td>endAgeUnit</td><td><<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|</td><td>INV(Condition.code) || Condition.abatementAge.quantity.code</td><td>767524001 |Unit of measure (qualifier value)|</td></tr>
<tr><td>inFamily</td><td><<404684003 |Clinical finding (finding)|</td><td>INV(FamilyMemberHistory.condition) || FamilyMemberHistory.relationship</td><td><<303071001 |Person in the family (person)|</td></tr>
</table>

Alias Predicates (namespace: anno:)	**Domain** (as SNOMED ECL expressions)	**Relational expression** (rooted SNOMED or FHIR linkage concepts / slots / relations)	**Range** (as SNOMED ECL expressions)
		INV(246090004 |Associated finding (attribute)|) || 408732007 |Subject relationship context (attribute)|	
informant	<<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|	INV(Condition.code) || INV(Provenance.target) || Provenance.agent.type	125676002 |Person (person)|
	<<71388002 |Procedure (procedure)|	INV(Procedure.code) || INV(Provenance.target) || Provenance.agent.type	
	<<363787002 |Observable entity (observable entity)| OR 254291000 |Staging and scales (staging scale)|	INV(Observation.code) || INV(Provenance.target) || Provenance.agent.type	
ingredient	373873005 |Pharmaceutical / biologic product (product)|	127489000 |Has active ingredient (attribute)|	< 105590001 |Substance (substance)|
familyDeath	<<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|	INV(FamilyMemberHistory.relationship) || FamilyMemberHistory.condition.contributedToDeath	boolean
laterality	<<123037004 |Body structure (body structure)|	272741003 |Laterality	<<182353008 |Side (qualifier value)|
	<<404684003 |Clinical finding (finding)|	363698007 |Finding site	85421007 |Structure of right half of body (body structure)| OR 31156008 |Structure of left half of body (body structure)|
	<<71388002 |Procedure (procedure)|	405813007 |Procedure site - Direct	85421007 |Structure of right half of body (body structure)| OR 31156008 |Structure of left half of body (body structure)|
	123037004 |Body structure (body structure)|	272741003 |Laterality (attribute)|	<< 182353008 |Side (qualifier value)|
medicationStatus			
morphology	<<71388002 |Procedure (procedure)|	363700003 |Direct morphology	<<118956008 |Body structure, altered from its original anatomical structure (morphologic abnormality)|
morphology	<<404684003 |Clinical finding (finding)|	116676008 |Associated morphology (attribute)|	<<118956008 |Body structure, altered from its original anatomical structure (morphologic abnormality)|
Or	138875005 |SNOMED CT Concept (SNOMED RT+CTV3)|	Logical OR	138875005 |SNOMED CT Concept (SNOMED RT+CTV3)|
requestIntent	<<71388002 |Procedure (procedure)|	INV(ServiceRequest.code) || ServiceRequest.intent	385661002 |Considered and not done (qualifier value)| OR 397943006 |Planned (qualifier value)|
sameAs		coreference	
severity	<<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|	INV(Condition.code) || Condition.severity	<<272141005 |Severities (qualifier value)|
		246112005 |Severity (attribute)|	
site	<<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|	363698007 |Finding site	<<123037004 |Body structure (body structure)|
		INV(Condition.code) || Condition.bodySite	

Alias Predicates (namespace: anno:)	**Domain** (as SNOMED ECL expressions)	**Relational expression** (rooted SNOMED or FHIR linkage concepts / slots / relations)	**Range** (as SNOMED ECL expressions)
	<<71388002 |Procedure (procedure)|	405813007 |Procedure site - Direct	
		INV(Procedure.code) || Procedure.bodySite	
siteIndirect	<<71388002 |Procedure (procedure)|	405814001 |Procedure site - Indirect	<<123037004 |Body structure (body structure)|
specimen	<<71388002 |Procedure (procedure)|	116686009 |Has specimen	<<123038009 |Specimen (specimen)|
strength	373873005 |Pharmaceutical / biologic product (product)|	732945000 |Has presentation strength numerator unit (attribute)|	decimal
value	<<363787002 |Observable entity (observable entity)| OR 254291000 |Staging and scales (staging scale)| OR <<386053000 |Evaluation procedure (procedure)| OR << 404684003 |Clinical finding (finding)|	INV(Observation.code) || Observation.value.quantity.value	decimal OR
		INV(Observation.code) || Observation.value.CodeableConcept	<<362981000 |Qualifier value (qualifier value)|
		363713009 |Has interpretation (attribute)|	
valueLow		INV(Observation.code) || Observation.value.Range.low	decimal
valueHigh		INV(Observation.code) || Observation.value.Range.high	decimal
valueComparator	<<363787002 |Observable entity (observable entity)| OR 254291000 |Staging and scales (staging scale)| OR <<386053000 |Evaluation procedure (procedure)|	INV(Observation.code) || Observation.value.quantity.comparator	Cf. Table 6
unit	<<363787002 |Observable entity (observable entity)| OR <<386053000 |Evaluation procedure (procedure)|	INV(Observation.code) || Observation.value.quantity.code	<<767524001 |Unit of measure (qualifier value)|
	373873005 |Pharmaceutical / biologic product (product)|	732945000 |Has presentation strength numerator unit (attribute)|	
verificationStatus	<<404684003 |Clinical finding (finding)| OR <<272379006 |Event (event)|	INV(Condition.code) || Condition.verificationStatus	Cf. Table 6
		INV(246090004 |Associated finding (attribute)|) || 408729009 |Finding context (attribute)|	

Table 6: Predicate values and their rooting in SNOMED CT and FHIR

FHIR value	Default	Predicate alias prefix anno:	Corresponding SNOMED CT concepts
<		anno:valueComparator	276139006 |Less-than symbol < (qualifier value)|
< =			276137008 |Less-than-or-equal symbol <= (qualifier value)|
>			276140008 |Greater-than symbol > (qualifier value)|
> =			276138003 |Greater-than-or-equal symbol >= (qualifier value)|
Unconfirmed		anno:verificationStatus	410590009 |Known possible (qualifier value)| 415684004 |Suspected (qualifier value)|
Provisional			410592001|Probably present (qualifier value)|
Confirmed	x		410605003 |Confirmed present (qualifier value)|
Refuted			410594000 |Definitely NOT present (qualifier value)| 410516002 |Known absent (qualifier value)|
Entered-in-error			723510000 |Entered in error (qualifier value)|
Unknown			261665006 |Unknown (qualifier value)|
Active	x	anno:clinicalStatus	394774009 |Active problem (qualifier value)|

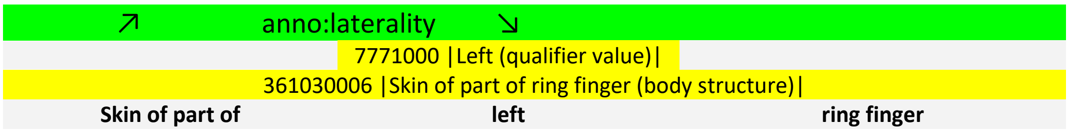
Inactive			394775005 |Inactive problem (qualifier value)|
Resolved			410513005 |In the past (qualifier value)|
Recurrence			255227004 |Recurrent (qualifier value)|
Remission			277022003 |Remission phase (qualifier value)|
Relapse			263855007 |Relapse phase (qualifier value)|
in-progress		anno:actionStatus	385651009 |In progress (qualifier value)|
not-done			385660001 |Not done (qualifier value)|
on-hold			385655000 |Suspended (qualifier value)|
completed	x		385656004 |Ended (qualifier value)| 410513005 |In the past (qualifier value)|
entered-in-error			723510000 |Entered in error (qualifier value)|
stopped			410545000 |Stopped before completion (qualifier value)|
unknown			410537005 |Action status unknown (qualifier value)|
proposal		anno:requestIntent	385661002 |Considered and not done (qualifier value)|
plan			397943006 |Planned (qualifier value)|

The main interesting point of the predicatess tabulated in Table 5 is that they can be transformed into a set of connected concepts, FHIR slots, or SNOMED CT predicates ("linkage concepts", corresponding to OWL object and datatype properties), or both in further steps, which provides a significant level of interoperability.

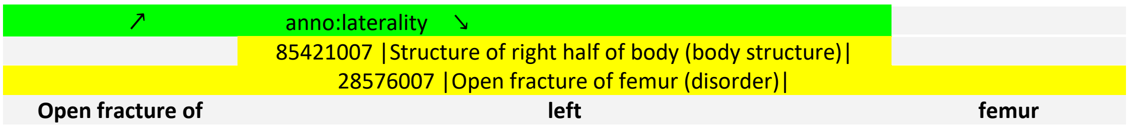
5.4. Specific annotation instructions

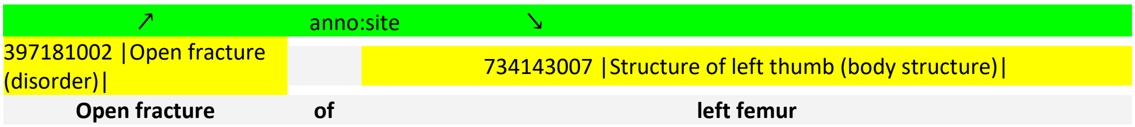
5.4.1. Delineation of annotation spans

If there is more than one alternative, e.g. annotating token t1 and annotating token t2+t3 with one code, or t1+t2 and t3, preference should be given to the alternative in which the focus concept (see below) has the most detailed annotation.


↗ anno:laterality ↘
7771000 |Left (qualifier value)|
361030006 |Skin of part of ring finger (body structure)|
Skin of part of
left
ring finger

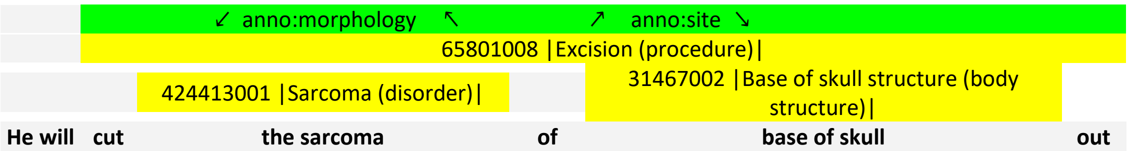
According to the constraints of the SNOMED CT concept model, this is not always the method of choice. For instance, in the following example, the disorder concept would be linked to the qualifier via the laterality relation, which however is only allowed for body structures. Therefore


↗ anno:laterality ↘
85421007 |Structure of right half of body (body structure)|
28576007 |Open fracture of femur (disorder)|
Open fracture of
left
femur


↗ anno:site ↘
397181002 |Open fracture (disorder)|
734143007 |Structure of left thumb (body structure)|
Open fracture
of
left femur

5.4.2. Non-contiguous passages

Within a sentence, one annotation can refer to a non-contiguous passage, i.e. spanning over tokens not included in the annotation. Annotations can therefore overlap (the concept Excision spans over the phrasal verb "cut…out").



↙ anno:morphology ↖ ↗ anno:site ↘

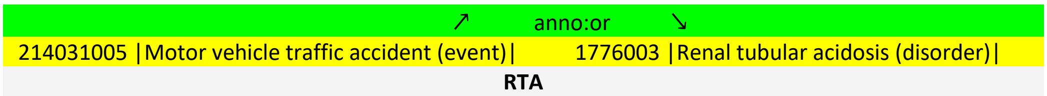
65801008 |Excision (procedure)|

424413001 |Sarcoma (disorder)| 31467002 |Base of skull structure (body structure)|

He will cut the sarcoma of base of skull out

5.4.3. Ambiguities

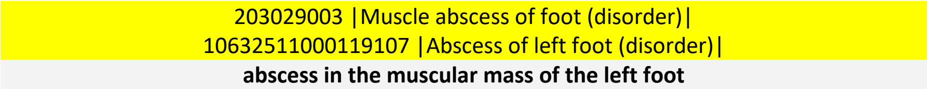
Ambiguities that can be resolved out of the context of a single sentence are resolved for the annotation (e.g. if the syntactic context makes clear that "RTA" means "renal-tubular acidosis" and not "road traffic accident". If not, then we allow for assigning more than one code, which then have to be related using the OR relator (pairwise). This is to distinguish those cases where one entity requires two or more codes that complete each other.



↗ anno:or ↘

214031005 |Motor vehicle traffic accident (event)| 1776003 |Renal tubular acidosis (disorder)|

RTA



203029003 |Muscle abscess of foot (disorder)|

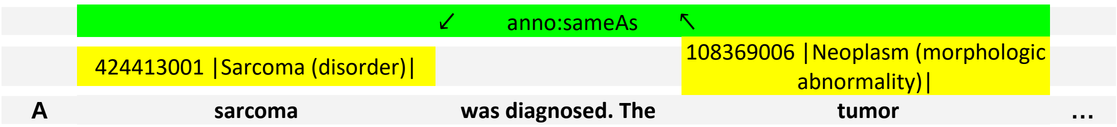
10632511000119107 |Abscess of left foot (disorder)|

abscess in the muscular mass of the left foot

In the example, with OR both relata have the same span. Since OR is symmetric and transitive, the direction does not matter. In the case of three or more relata, the only concern is that all relata are directly or indirectly related.

5.4.4. Coreference

Annotation is done to the granularity level given by the sentence, regardless of additional knowledge the reader has. In the case of nominal anaphora, a coreference link ("sameAs") is used.

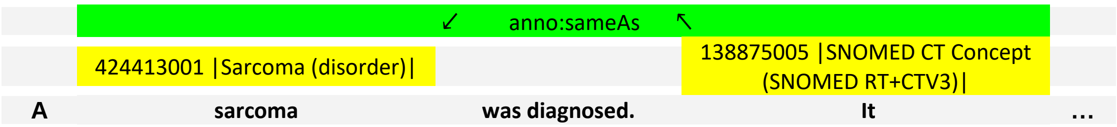


↙ anno:sameAs ↖

424413001 |Sarcoma (disorder)| 108369006 |Neoplasm (morphologic abnormality)|

A sarcoma was diagnosed. The tumor …

The same with nominal anaphora. The pronoun is annotated with the top concept of SNOMED.



↙ anno:sameAs ↖

424413001 |Sarcoma (disorder)| 138875005 |SNOMED CT Concept (SNOMED RT+CTV3)|

A sarcoma was diagnosed. It …

5.4.5. Ellipsis

Ellipses such as expressing “patient wears glasses” just by “glasses” are common in clinical texts. In obvious cases, annotators should use the concepts that are really meant (i.e. focus concepts) for annotation.

↗	anno:beginTime	↘
225582009 |Wears glasses (finding)|		108369006 |Neoplasm (morphologic abnormality)|
Glasses	**since the age of**	**13 years**

Very common are ellipses in expressions such as “Kidney: NAD” (“no abnormality detected”) or “Neurological examination: normal”

↗ anno:value	↘
364180004 |Kidney feature (observable entity)|	
Kidney	**NAD**

Wherever the whole statement is expressed by a single concept, use this:

↗ anno:value	↘
225544001 |Skin appearance normal (finding)|	
Skin	**NAD**

12236201000119103 |Conjunctivitis of right eye (disorder)|	301926003 |Conjunctiva normal (finding)|
Unilateral conjunctivitis right	**Left side NAD**

5.4.6. Unspecific annotations

Unspecific annotations due to limited granularity of the annotation vocabulary receive a meta-annotation “Incomplete”.

				Incomplete
	225589000 |Chokes when swallowing (finding)|			227607007 |Candy (substance)|
The patient	**choked**	**on a**	**swallowed**	**Mars bar**

5.4.7. Drug products

The mention of a drug is annotated with a code from the hierarchy Pharmaceutical / biologic product.

374627000 |Product containing precisely diclofenac sodium 50 milligram/1 each conventional release oral tablet (clinical drug)|
Voltaren 100 tablets

In cases where there is no concept from this hierarchy, the ingredients and other characteristics need to be linked:

↗		anno:doseForm	↘
763158003 |Medicinal product (product)| 1119336002 |Bamlanivimab (substance)|	↘ ↙	anno:ingredient	764485007 |Conventional release cutaneous patch (dose form)|
Bamlanivimab			**cutaneous patch**

5.4.8. Combination of annotations

One entity may have two or more coinciding annotations. Three cases need to be distinguished, for which the following conventions are used:

1. The entity is annotated by a set of concepts. All are true and necessary. The order does not matter.

<table>
<tr><td>1528001 |Product containing folinic acid (medicinal product)|
3127006 |Product containing fluorouracil (medicinal product)|
327032007 |Product containing oxaliplatin (medicinal product)|</td></tr>
<tr><td><b>FOLFOX regimen</b></td></tr>
</table>

2. Two or more concepts are given for an ambiguous expression, e.g. an acronym. It is not clear from the context which is the correct interpretation.

<table>
<tr><td colspan="2">anno:or</td></tr>
<tr><td>48724000 |Mitral valve regurgitation (disorder)|</td><td>22298006 |Myocardial infarction (disorder)|</td></tr>
<tr><td colspan="2"><b>MI</b></td></tr>
</table>

3. Domain and range in one word (typical in languages with single word compounds). We do not perform annotations on a subword level. Here the word “tumorfree” is the origin and end of the arrow.

<table>
<tr><td>108369006 |Neoplasm (morphologic abnormality)|</td><td rowspan="2">↘ anno:verificationStatus ↙</td></tr>
<tr><td>2667000 |Absent (qualifier value)|</td></tr>
<tr><td><b>tumorfree</b></td><td></td></tr>
</table>

4. Implied concepts: in case of ellipsis, annotations are done as if no tokens were dropped, like here: Right (eye) nornal (vision).

<table>
<tr><td></td><td colspan="2">↙ anno:site ↖</td></tr>
<tr><td>339301000119109 |Myopia of left eye (disorder)|</td><td>18944008 |Right eye structure (body structure)|</td><td>45089002 |Normal vision (finding)|</td></tr>
<tr><td><b>Left eye nearsighted</b></td><td><b>Right</b></td><td><b>normal</b></td></tr>
</table>

5.4.9. Negation, uncertainty, clinical status and severity

Concepts with a negative meaning such as 162062008 |No vomiting (situation)| are to be avoided whenever they are expressible by combining the positive meaning (as given by SNOMED CT) with the value “refuted” in Condition.verificationStatus as given by FHIR. The use of SNOMED CT concepts with negative meaning is limited to those cases where there is no alternative, e.g. 249695006 |Absence of rib (finding)| because there is no “presence of rib”, or where they correspond to overly popular terms such as “non-smoker”.

<table>
<tr><td colspan="2">↙ anno:verificationStatus ↖</td></tr>
<tr><td>410594000 |Definitely NOT present (qualifier value)|</td><td>422400008 |Vomiting (disorder)|</td></tr>
<tr><td><b>No</b></td><td><b>vomiting</b></td></tr>
</table>

<table>
<tr><td colspan="2">↙ anno:verificationStatus ↖</td></tr>
<tr><td>415684004 |Suspected (qualifier value)|</td><td>77386006 |Pregnancy (finding)|</td></tr>
<tr><td><b>Suspected</b></td><td><b>pregnancy</b></td></tr>
</table>

<table>
<tr><td colspan="3">↙ anno:clinicalStatus ↖</td></tr>
<tr><td>263855007 |Relapse phase (qualifier value)|</td><td></td><td>24700007 |Multiple sclerosis (disorder)|</td></tr>
<tr><td><b>relapse</b></td><td><b>of</b></td><td><b>MS</b></td></tr>
</table>

In situations of doubt, it is principally possible to annotate with both alternatives:

<table>
<tr><td></td><td colspan="2">↙ anno:verificationStatus ↖</td></tr>
<tr><td></td><td>410594000 |Definitely NOT present (qualifier value)|</td><td>65568007 |Cigarette smoker (finding)|</td></tr>
<tr><td colspan="3">8392000 |Non-smoker (finding)|</td></tr>
<tr><td><b>The patient</b></td><td><b>does not</b></td><td><b>smoke</b></td></tr>
</table>

5.4.10. Procedure status and intent

Procedure status is only annotated if the procedure was done prior to the episode of care described in the document.

<table>
<tr><td></td><td>236886002 |Hysterectomy (procedure)|</td><td rowspan="2">↘ anno:actionStatus
↙</td></tr>
<tr><td></td><td>410513005 |In the past (qualifier value)|</td></tr>
<tr><td><b>The patient had had</b></td><td><b>a hysterectomy</b></td><td></td></tr>
</table>

<table>
<tr><td>236886002 |Hysterectomy (procedure)|</td><td rowspan="2">↘ anno:actionStatus
↙</td></tr>
<tr><td>410513005 |In the past (qualifier value)|</td></tr>
<tr><td><b>hysterectomised</b></td><td></td></tr>
</table>

Planned procedures are expressed with anno:requestIntent.

<table>
<tr><td colspan="2">↙ anno:requestIntent ↖</td></tr>
<tr><td>397943006 |Planned (qualifier value)|</td><td>236886002 |Hysterectomy (procedure)|</td></tr>
<tr><td><b>We planned</b></td><td><b>a hysterectomy</b></td></tr>
</table>

5.4.11. Qualitative values

Concepts of the hierarchies "Observable entity" and "Staging and scales" are only used with a value. Wherever this hierarchy does not provide a concept to express the measurement of something, use subconcepts of 785673007 |Measurement of level of substance in blood (procedure)| instead.

A current drawback of Observables is that they are often not related to their defining concepts, e.g. 446089006 |Volume of lower limb (observable entity)| is not related to the lower limb concept. It is therefore undefined how to refine observables via post-coordination, such as Volume of left lower limb. We here suggest that for laterality, the predicate anno:laterality is used in the same way as for body parts.

Values can be quantitative or qualitative. Qualitative values, like the attribution of "elevated" to an observable such as "systolic blood pressure" just requires the linkage of two concepts.

For relating observable entities (as well as SNOMED procedures substituting observables) with values, we use the predicate anno:value. It can be linked to codes when qualitative values are used:

↗ anno:value ↘		
364075005 |Heart rate (observable entity)|		75540009 |High (qualifier value)|
Heart frequency	**is**	**elevated**

Also, other types of concepts may be refined by the association of values, e.g.

↙ anno:value ↖	
87628006 |Bacterial infectious disease (disorder)|	87628006 |Bacterial infectious disease (disorder)|
bacterial infections	**frequent**

5.4.12. Quantitative values

Quantitative values are more complex. Quantitative values are more diverse, e.g. “Heart rate < 40 /min” vs. “Heart rate 40 /min” vs. “Heart rate 40”. For relating observable entities with quantitative values, we often have to consider units and in rare cases comparators. If there is no unit in the text, it is left out. As general, missing or unclear information is not guessed in the annotation process. Only concepts of the hierarchies “Observable entity” and “Procedure” use (non-temporal) quantitative values.

↗ anno:valueUnit ↘			
↗ anno:value ↘			
↗ anno:valueComparator ↘			
364075005 |Heart rate (observable entity)|	276139006 |Less-than symbol < (qualifier value)|	40.0	286549009 |per minute (qualifier value)|
Heart frequency	**<**	**40**	**/ min**

↗ anno:value ↘		
401201003 |Cigarette pack-years (observable entity)|		
1.0		
Smokes	**1**	**py**

Numeric values are annotated as decimals. Note that the decimal separator is always the period.

5.4.13. Temporal values

Everything can be related to a temporal value, such as a date, but also a time period or an age.

	↗ anno:beginTime ↘		
	22298006 |Myocardial infarction (disorder)|		2021-12-03
The patient had a	**heart attack**	**on**	**Dec 3, 2021**

Reference to time can also be related with clinical status values.

	↗ anno:beginTime	↘
↙ anno:clinicalStatus ↖		
410513005 |In the past (qualifier value)|	82313006 |Suicide attempt (event)|	2008-02-xx
History of	**Suicide attempt**	**(02/2008)**

If there is only one time reference (or a coarse-grained interval such as a month) we use anno:beginTime (although there is no end time).

↗	anno:beginAgeUnit		↘
↗ beginAge		↘	
22298006 |Myocardial infarction (disorder)|		61.0	258707000 |year (qualifier value)|
Heart attack	**at**	**61**	**years**

For very frequent conditions, there are observable concepts that take an age or a time as a value, such as 228488005 |Age at starting smoking (observable entity)|. For the sake of annotation homogeneity we do not use them, instead

↗	anno:beginAgeUnit		↘
↗ beginAge		↘	
65568007 |Cigarette smoker (finding)|		13.0	258707000 |year (qualifier value)|
She started smoking	**when she was**	**13**	**Years old**

5.4.14. Family History

↙	anno:inFamily	↖
27733009 |Sister (person)|		46635009 |Diabetes mellitus type 1 (disorder)|
Sister	**has**	**Type 1 diabetes**

	↙	anno:familyDeath ↖
↙	anno:inFamily	↖
66839005 |Father (person)|	TRUE	363358000 |Malignant tumor of lung (disorder)|
Father	**died from**	**lung cancer**

5.4.15. Informant

It is often important to specify the person that informs about something that happened.

	↗	anno:informant	↘
	271594007 |Syncope (finding)|		106304008 |Teacher (occupation)|
The child	**fainted**	**according to the**	**teacher**

5.4.16. Codes

Apart from literals denoting personal health information (clinicians, patients, institutions, person IDs), which have been automatically removed prior to annotation (in case of failure see measures to be taken 7.2.1), and dates and numbers (cf. 5.4.12 and 5.4.13), the occurrence of codes (particularly ICD-10, LOINC) in narratives is frequent. They should be treated like abbreviations and annotated according to our annotation rules. In the cases in which they are obviously more coarse-grained in meaning compared to already annotated passages (which is frequent) they are considered redundant and can be ignored[9].

The example shows that the coarse-grained nature of ICD-10 entails an ambiguity, due to the coverage of two different types of metastases ("Secondary malignant neoplasm of liver and intrahepatic bile duct") by the code C78.7. This ambiguity would need to be resolved in an additional step. There is no information gain because the annotation of "Liver metastases" is already completely sufficient. Therefore "C78.7" should not be annotated.

94381002 |Metastatic malignant neoplasm to liver (disorder)|	126853008 |Neoplasm of biliary tract (disorder)|	↘ anno:or ↙
	94381002 |Metastatic malignant neoplasm to liver (disorder)|	
Liver metastases	**C78.7**	

A counterexample is the following. Here the information in the code (Malignant neoplasm: Upper-inner quadrant of breast) is indeed more specific than Breast cancer alone. Here, the annotation of the ICD code contributes with additional information.

254837009 |Malignant neoplasm of breast (disorder)|	373082000 |Malignant neoplasm of breast upper inner quadrant (disorder)|
Breast cancer	**C50.2**

[9] For German narratives that are pre-annotated using Averbis Health Discovery, automatic pre-annotation of codes can be expected

6. Exemplification

This section shows how the designed guideline can be instantiated and applied to BC use case and CVD use case by examples. However, new examples encountered during the annotation process can provide valuable insights and help identify areas where the guidelines can be enhanced. Therefore, the annotation guideline (including the principles, rules, instructions, and introduced predicates) is requested to be refined and expanded within the project by incorporating more data.

6.1 Assumptions

Several assumptions have been formulated in the annotation guideline that are considered essential to be followed throughout the entire annotation process. These assumptions are outlined below.

- The default subject throughout the entire process is the patient (e.g., the subject of care) and there is no need to annotate it.
- Conditional, hypothetical and imperative expressions, as well as questions, should not be annotated.
- The default value for the presence of a finding is known (yes) and no need to annotate it (the default values of the introduced predicates are shown in the 'default' column of the Table 6).

6.2 Process template

In the following, the instructions mostly related to the BC use case and CVD use case are summarised.

- The steps to annotate data related to situation concepts include:
  a. Identify the subject of the record, which may be the patient or a relevant family member (e.g., mother) and then annotate it only if it refers to the family member (refers to Section 6.4.2).
  b. Identify the clinical finding.
  c. Connect the subject (detected in step a) to the clinical finding using the ‘inFamily’ predicate.

     If any Protected Health Information (PHI) [14] persists despite the data de-identification procedure outlined in the first version of this guideline (Deliverable D4.1 – Section 4.2), annotators are requested to promptly report such instances to the designated local data steward responsible for the de-identification process[10].

  d. Identify a qualifier value indicating the presence of the clinical finding (i.e., present, absent, or unknown) if applicable (according to Section 6.4.2, if the presence of a finding is already ‘known’, there is no need to annotate it).
  e. Connect the clinical finding to the qualifier value determined in Step d using the ‘verificationStatus’ predicate.
  f. Identify a qualifier value indicating the temporal context of the clinical finding (i.e., current, past, current and past, or unknown) if applicable.
  g. Connect the clinical finding to the qualifier value determined in step f using the ‘clinicalStatus’ predicate.
- To annotate certain findings such as smoking behaviours in detail, the following steps[11] need to be followed:

[10] This ensures that appropriate measures can be taken to address and rectify any remaining PHI to safeguard data privacy and compliance.

[11] However, while it is possible to create general instructions for annotating any clinical finding, they are not applied in the AIDAVA use case as the focus is to utilise SNOMED CT as extensively as possible.

a. Annotate the date or duration if there is any mention in the text according to the following Table.

Table 7: smoking behavior

Pattern in Text	**SNOMED CT Concept**
age at start	Age at starting smoking (observable entity)
age at stop	Age at stopping smoking (observable entity)
start n time-unit ago	Time since started smoking (observable entity) << 258700003 |Non-International System of Units unit of time (qualifier value)|
stop n time-unit ago	Time since stopped smoking (observable entity) << 258700003 |Non-International System of Units unit of time (qualifier value)|
start in the year1	Date of onset (observable entity) **AND** Tobacco use and exposure
stop in the year2	Date ceased smoking (observable entity)
time-unit duration	Total time smoked (observable entity) << 258700003 |Non-International System of Units unit of time (qualifier value)|

b. Annotate the smoking quantity unit according to the following table:

Table 8: Quantity units for smoking behavior

Quantity Unit	**SNOMED CT Concept**
pack	Pack (physical object)
cigarette	Cigarette (physical object)

c. Annotate the denominator of time regarding the amount of smoking using << 282363004 |Denominators of time (qualifier value)|
d. Annotate the values numbers for duration, years, and smoking quantity
e. Use the introduced relations in Table 5 to establish links among the annotated spans (i.e, valueUnit, value, valueLow, valueHigh, and so on)

6.3. Annotation examples using INCEpTION

In this section, we will illustrate the annotation process through several examples that demonstrate the application of the guideline. Each subsection will focus on a specific example from CVD or BC and provide a detailed explanation of the approach and results. By following these examples, annotators will gain a better understanding of how to apply the instructions described in Section 6.1. to their own documents. The template used to organise the examples induces the input text, a screenshot of the annotations made using INCEpTION, and a description of how the general manual annotation principles are applied. Examples are an important resource for annotator training. During the annotation process, annotators and trainers will collect more examples, both on English texts (for better mutual discussion) and increasingly in their own language.

6.3.1. Smoking behaviour

Table 9: Example of smoking behaviour

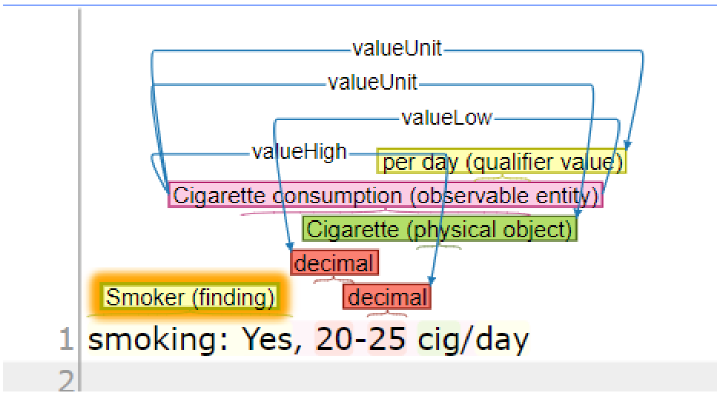
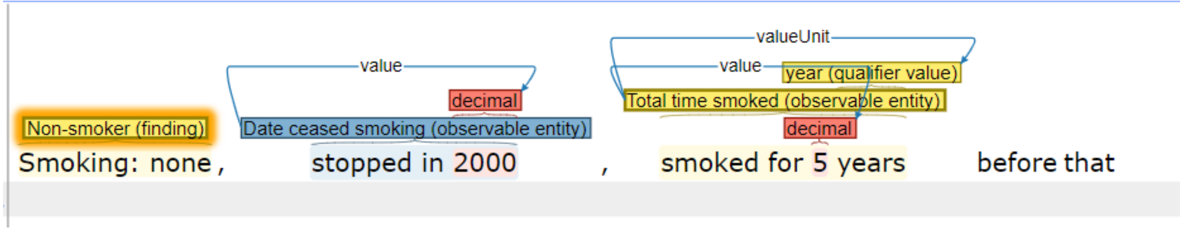
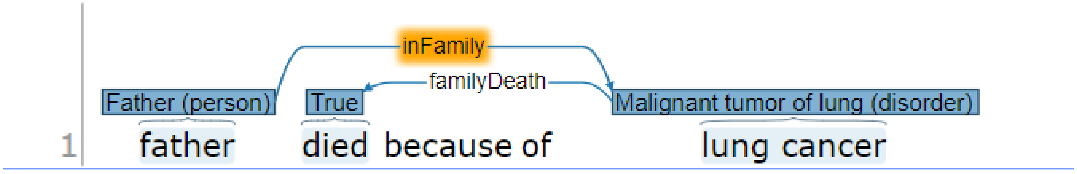
Input	smoking: Yes, 20-25 cig/day
INCEpTION screenshot	 valueUnit valueUnit valueLow valueHigh per day (qualifier value) Cigarette consumption (observable entity) Cigarette (physical object) decimal Smoker (finding) decimal 1 smoking: Yes, 20-25 cig/day 2
Adjudication Description	The input text presents the duration of smoking per day. Therefore, following the explanation in Section 6.2, the range of values (20-25), the time unit (day) and quantity unit (cigarette) of smoking are annotated. According to Table 5, the predicate 'valueUnit' is used to establish connections from the observable entity (i.e., Cigarette consumption) to the detected time units and quantity units. Similarly, the predicates 'valueLow' and 'valueHigh' are applied to make connections from the observable entity to the decimals.

Table 10: Example of smoking behaviour

Input	Smoking: none, stopped in 2000, smoked for 5 years before that
INCEpTION screenshot	 value decimal valueUnit value year (qualifier value) Total time smoked (observable entity) decimal Non-smoker (finding) Date ceased smoking (observable entity) Smoking: none , stopped in 2000 , smoked for 5 years before that
Adjudication Description	Two predicted patterns (stop at a specific date, and time duration) in smoking data appear in the input text. Therefore, instructions X and Y are followed to annotate ‘stopped in 2000’ and ‘smoked for 5 years’ as ‘Date ceased smoking’ and ‘Total time smoked’, respectively. Also, the numbers 2000 and 5 were coded as decimals. Moreover, the time-unit needs to be identified (as instructed in Section 6.2), so ‘years’ correspond to ‘year (qualifier value)’. As shown in Table 5, the predicate ‘value’ has observable entity and decimal as its domain and range, respectively. Thus, we use the predicate ‘value’ between the identified observable entities and their corresponding decimals. The same explanation applies to ‘valueUnit’ between the observable value and the qualifier value (i.e., year).

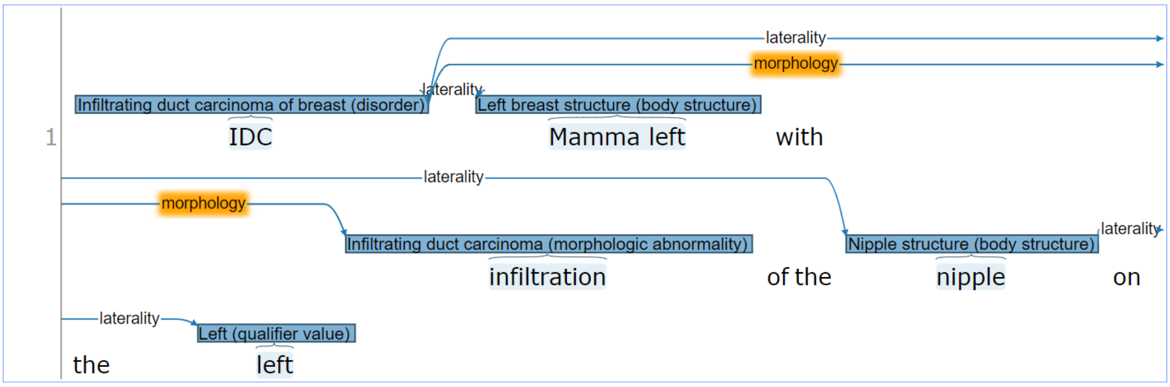
6.3.2. Family history

Table 11: Example of family history

Input	father died because of lung cancer
INCEpTION Screenshot  inFamily familyDeath Father (person) True Malignant tumor of lung (disorder) 1 father died because of lung cancer	
Adjudication Description	Since the reference is to a family member (not the main patient), it should be annotated with <<303071001 |Person in the family (person)|, which in this case is the father. Then, the predicate 'inFamily' is used to establish a link between the disorder (i.e., lung cancer) and the person according to Table 5. As the range of the predicate 'familyDeath' is a boolean value (True or False), the term 'died' was marked as True, and then the predicate was established from the lung cancer to True (as per Instruction Family History in Section 5.4.14).

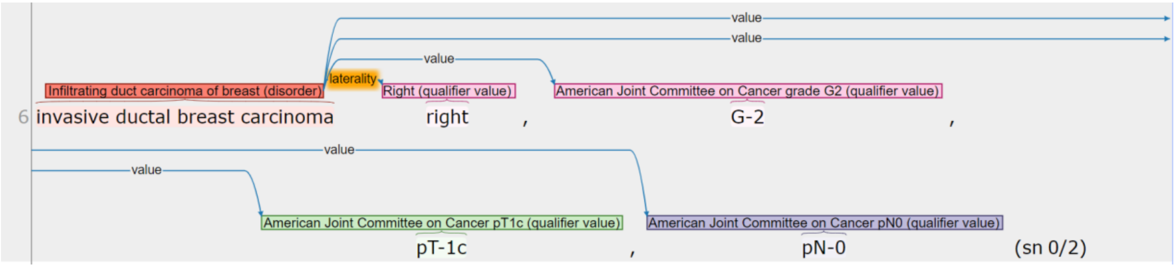
6.3.3. Histology

Table 12: Example of histology

Input	IDC Mamma left with infiltration of the nipple on the left
INCEpTION Screenshot  laterality morphology laterality Infiltrating duct carcinoma of breast (disorder) Left breast structure (body structure) 1 IDC Mamma left with laterality morphology Infiltrating duct carcinoma (morphologic abnormality) Nipple structure (body structure) laterality infiltration of the nipple on laterality Left (qualifier value) the left	
Adjudication Description	According to Section 5.3.1 and Section 6.2, the core concepts and their corresponding qualifier values are coded. To establish links among the identified concepts, Table 5 is used to find appropriate predicates between the concepts based on the domain and range of the specified predicates. For example, the predicate 'morpholoy' is used to connect a clinical finding (i.e., IDC) with a morphologic abnormality (i.e., infiltration).

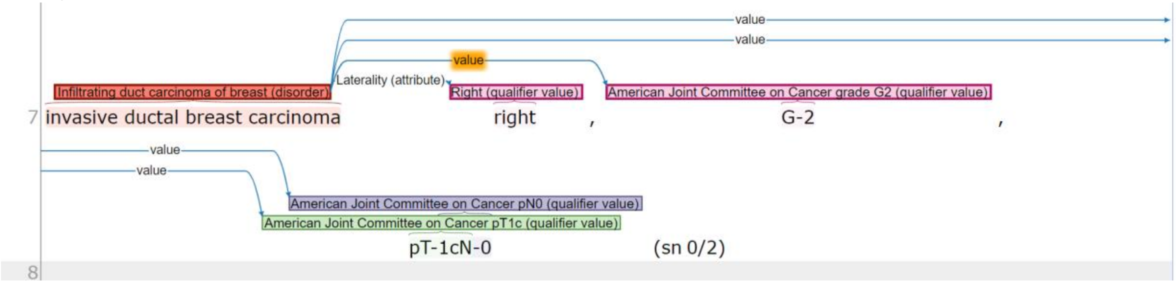
6.3.4. TNM stage

Table 13: Example of TNM stage

<table>
<tr><td><b>Input</b></td><td>invasive ductal breast carcinoma right, G-2 pT-1c, pN-0(sn 0/2)</td></tr>
<tr><td colspan="2"><b>INCEpTION Screenshot</b>

value
value
value
laterality
Infiltrating duct carcinoma of breast (disorder)
Right (qualifier value)
American Joint Committee on Cancer grade G2 (qualifier value)
6 invasive ductal breast carcinoma
right
,
G-2
,
value
value
American Joint Committee on Cancer pT1c (qualifier value)
American Joint Committee on Cancer pN0 (qualifier value)
pT-1c
,
pN-0
(sn 0/2)

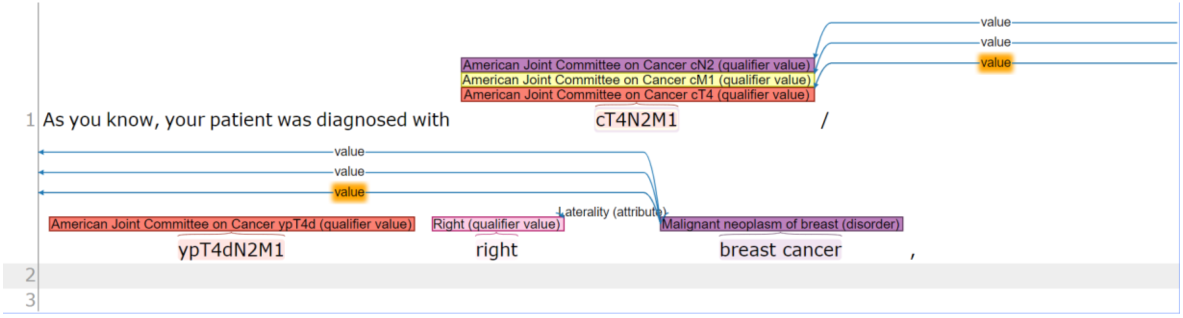
</td></tr>
<tr><td><b>Adjudication Description</b></td><td>The span ‘G-2 pT-1c, pN-0(sn 0/2)’ is clearly composed by parts that are separated by comma and each part needs to be coded separately (Section 6.2). Thus, G-2, pT-1c, and pN-0 are normalised with the corresponding concepts. According to Table 5, to associate a disorder with a side qualifier value, the predicate 'laterality' should be used. Similarly, the predicate 'value' is used to establish a link from the disorder to the qualifier values indicating TNM staging.</td></tr>
</table>

Table 14: Example of TNM stage

<table>
<tr><td><b>Input</b></td><td>invasive ductal breast carcinoma right, G-2 pT-1cN-0(sn 0/2)</td></tr>
<tr><td colspan="2"><b>INCEpTION Screenshot</b>

value
value
value
Laterality (attribute)
Infiltrating duct carcinoma of breast (disorder)
Right (qualifier value)
American Joint Committee on Cancer grade G2 (qualifier value)
7 invasive ductal breast carcinoma
right
,
G-2
,
value
value
American Joint Committee on Cancer pN0 (qualifier value)
American Joint Committee on Cancer pT1c (qualifier value)
pT-1cN-0
(sn 0/2)
8

</td></tr>
<tr><td><b>Adjudication Description</b></td><td>As stated in Section 5.2, annotations on a subword level are permitted. Therefore, 'N-0' is coded with '1229947003 |American Joint Committee on Cancer pN0 (qualifier value)|'. Additionally, the predicate 'value' is employed to establish relationships between a disorder and qualifier values based on Table 5.</td></tr>
</table>

Table 15: Example of TNM stage

Input	As you know, your patient was diagnosed with cT4N2M1/ypT4dN2M1 right breast cancer,
INCEpTION Screenshot  As you know, your patient was diagnosed with cT4N2M1 / ypT4dN2M1 right breast cancer , value value value American Joint Committee on Cancer cN2 (qualifier value) American Joint Committee on Cancer cM1 (qualifier value) American Joint Committee on Cancer cT4 (qualifier value) value value value Laterality (attribute) American Joint Committee on Cancer ypT4d (qualifier value) Right (qualifier value) Malignant neoplasm of breast (disorder)	
Adjudication Description	Similar to the two previous examples explained, 'CT4N2M1' can be annotated at the subword level (Section 5.2). Furthermore, the predicate 'value' with the domain of clinical finding and the range of qualifier values (Table 5) is used to establish a link between the breast cancer disorder and the identified qualifier values.

6.3.5. Example therapy

Table 16: Example of therapy

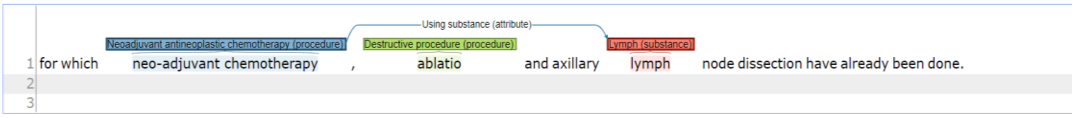
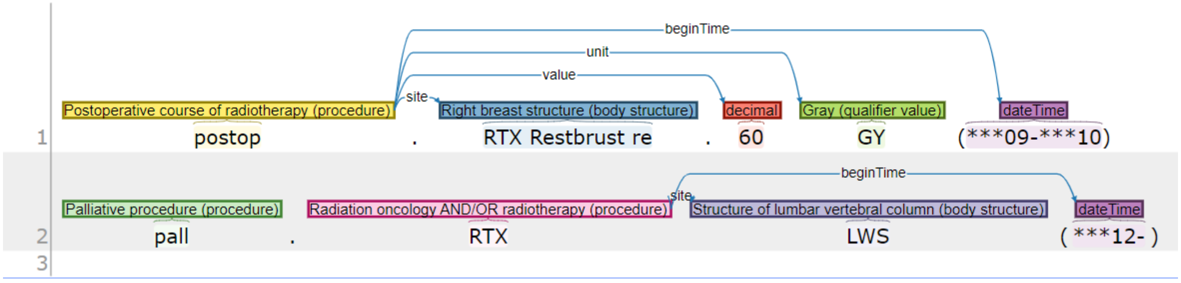
Input	for which neo-adjuvant chemotherapy, ablatio and axillary lymph node dissection have already been done.
INCEpTION Screenshot  for which neo-adjuvant chemotherapy , ablatio and axillary lymph node dissection have already been done. Using substance (attribute) Neoadjuvant antineoplastic chemotherapy (procedure) Destructive procedure (procedure) Lymph (substance)	
Adjudication Description	According to Table 3, procedures (considered core concepts) and substances (if there is no appropriate product) must be identified and coded. While Table 5 does not provide a specific predicate with the domain and range for procedures and substances, respectively, annotators are strongly recommended to use the SNOMED CT Browser to explore which attributes can be applied. In this example, the attribute 'Using substance' can be employed to establish a connection between the procedure and substances. Moreover, as indicated in Table 6, the default value for the predicate 'actionStatus' is 'done/completed'. Therefore, the span related to the status does not need to be coded or annotated.

Table 17: Example of therapy

Input	postop. RTX Restbrust re. 60 GY (***09-***10) pall. RTX LWS (***12-)
INCEpTION Screenshot  beginTime unit value site Postoperative course of radiotherapy (procedure) Right breast structure (body structure) decimal Gray (qualifier value) dateTime 1 postop . RTX Restbrust re . 60 GY (***09-***10) beginTime site Palliative procedure (procedure) Radiation oncology AND/OR radiotherapy (procedure) Structure of lumbar vertebral column (body structure) dateTime 2 pall . RTX LWS (***12-) 3	
Adjudication Description	All abbreviations are annotated using the appropriate concepts (it is recommended to use Google to determine the meaning of an abbreviation). Additionally, Table 5 is referenced to assert predicates between concepts based on the domain and range of the predicates. For instance, 'beginTime' is used to connect a procedure to a 'dateTime'.

6.3.6. Common examples

Table 18: Common example

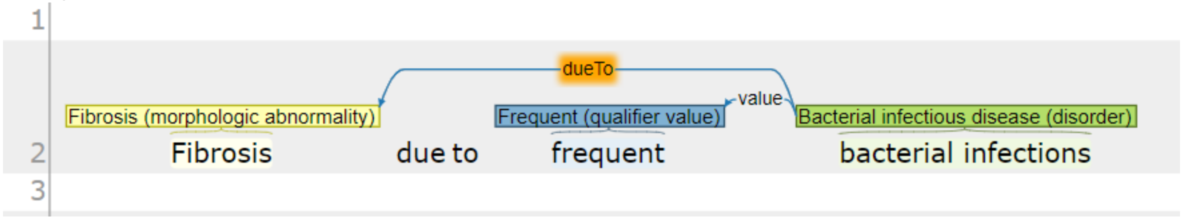
Input	Fibrosis due to frequent bacterial infections
INCEpTION Screenshot  1 dueTo value Fibrosis (morphologic abnormality) Frequent (qualifier value) Bacterial infectious disease (disorder) 2 Fibrosis due to frequent bacterial infections 3	
Adjudication Description	According to Table 3, the mentioned concepts of core concepts as well as qualifier values are identified and coded. To assert a predicate from a disorder to a morphologic abnormality, the predicate 'dueTo' is allowed to be applied. Moreover, the detected disorder is linked to the qualifier value using the predicate 'value' according to Table 5.

Table 19: Common example

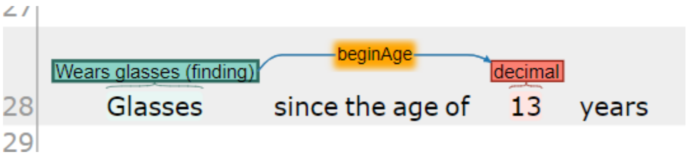
Input	Glasses since the age of 13 years
INCEpTION Screenshot	 Wears glasses (finding) beginAge decimal Glasses since the age of 13 years
Adjudication Description	As stated in Section 5.4.5, to handle ellipses, annotators must utilize the most suitable concepts. Therefore, in this case, 'Glasses' is coded as 'Wears glasses'. The decimal value representing the age is annotated as a decimal, and then the predicate 'beginAge' is employed to connect the finding to the decimal, as specified in Table 5.

Table 20: Common example

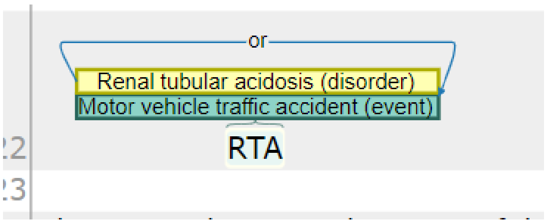
Input	RTA
INCEpTION Screenshot	 or Renal tubular acidosis (disorder) Motor vehicle traffic accident (event) RTA
Adjudication Description	As discussed in Section 5.4.3, in cases of ambiguities, despite the principles and expertise of the annotators, the predicate 'OR' is applied to connect candidate concepts of the ambiguous span.

Table 21: Common example

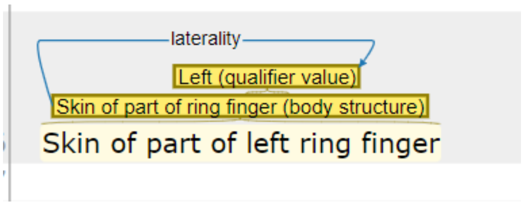
Input	Skin of part of left ring finger
INCEpTION Screenshot	 laterality Left (qualifier value) Skin of part of ring finger (body structure) Skin of part of left ring finger
Adjudication Description	Annotators must consider the most detailed annotations throughout the entire process, as explained in Section 5.4.1. Therefore, in this example, the most detailed concept is used to annotate 'Skin of part of the ring finger', and then the predicate 'Left' is employed to assert a predicate between the detected body structure and its corresponding side, which is the left side.

Table 22: Common example

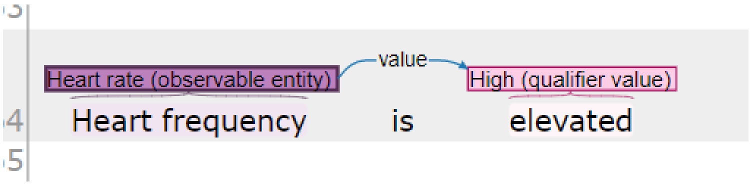
Input	Heart frequency is elevated
INCEpTION Screenshot  Heart rate (observable entity) value High (qualifier value) Heart frequency is elevated	
Adjudication Description	According to Table 5, the predicate 'value' is used with the domain and range of observable entities and qualifier values to assert a predicate between the detected concepts.

Table 23: Common example

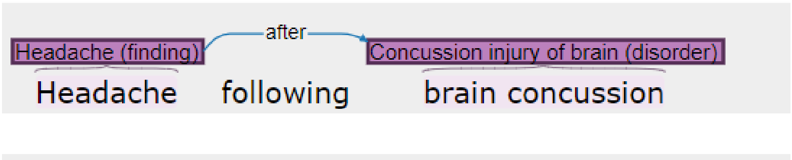
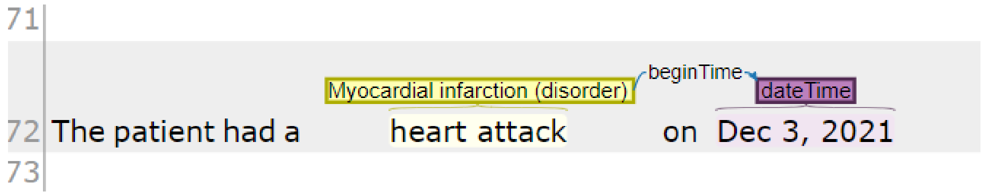
Input	Headache following brain concussion
INCEpTION Screenshot  Headache (finding) after Concussion injury of brain (disorder) Headache following brain concussion	
Adjudication Description	As discussed in Section 5.2, annotators are not permitted to infer causality. Therefore, in this example, the only applicable predicate is a 'after'.

Table 24: Common example

<table>
<tr><td><b>Input</b></td><td>The patient had a heart attack on Dec 3, 2021</td></tr>
<tr><td colspan="2"><b>INCEpTION Screenshot</b>

The patient had a heart attack on Dec 3, 2021
</td></tr>
<tr><td><b>Adjudication Description</b></td><td>According to the assumption in Section 6.1, there is no need to annotate the patient. Furthermore, the 'beginTime' is used to establish a link between the disorder and dateTime, as stated in Table 5.</td></tr>
</table>

7. Quality control

7.1. Training of annotators

MAASTRO/UM: For the CVD case, two medical students (3rd or 4th year) who already work as part-time research assistants at the cardiology department of the Maastricht University Medical Centre will be hired for the annotation procedure. The students who will be hired will be already familiar with medical CVD terms and the Dutch language. For the breast cancer case, two medical students (2nd and 3rd year) will be hired in collaboration with the clinical trials office (CTO) department of MAASTRO clinic. The students will already have experience with dealing with radiotherapy reports that include radiotherapy related terms (both diagnostic and referral reports).

MUG: Three medical students responsible for the manual annotations for the CVD and BC use cases have been hired. They have profound language skills in English to understand the manual annotation guideline, as well as a medical domain-specific language understanding. Based on the manual annotation guideline provided in this document, they will be trained by the local trainers on site for the production phase.

NEMC: The corresponding CVD and BC teams have been set up to consist of one or more annotators, one manager, and two lead curators. The annotators are resident physicians in their field and thus have a deep understanding of their domain. The manager is a computational linguist with education in health information management and fills four roles explained below. The lead curators are one medical doctor and one medical secretary. The medical secretaries are the ones with the most experience in the data entry task that the annotation-trained language model is to automate. The medical doctors have been introduced to the aim of annotation and have been made familiar with the tools and workflow. The four roles of the manager are the following: • Improve process quality by listening to feedback from the participators. • Solve running problems concerning the annotation tool and clinical document data sets. • Train new annotators and disseminate new knowledge. • Communication management, trying to answer questions within the group before involving lead curators.

7.1.1. Training schema

To train the hired medical students as annotators, the following steps will be taken by the local responsible trainer at MUG, MAASTRO/UM, and NEMC:

Familiarise the annotators with the annotation task. Provide a detailed explanation of the annotation task and its objectives. Clarify the specific annotations required for each medical case (CVD and breast cancer) and the desired outcomes. Ensure that the annotators have a clear understanding of the medical conditions, relevant terminology, and the importance of accurate and consistent annotations.

Provide comprehensive annotation guideline. Develop detailed annotation guideline (following the steps of sections 5 and 6) that outline the specific criteria, instructions, and examples for performing the annotations.

Conduct training sessions. Organise training sessions to educate the annotators about the annotation process. These sessions can include presentations, demonstrations, and hands-on exercises with real-world data. There are already “Train the trainers” sessions on how to interpret and apply the annotation guideline effectively by the project coordinators of the different centers. Focus on areas such as identifying relevant medical terms, understanding context, resolving ambiguities, and maintaining consistency in annotations.

Practice annotation exercises. Provide sample data or cases for the annotators to practise their annotation skills. These exercises should closely resemble the real annotation task and cover various scenarios and challenges they may encounter.

7.2. Quality process and quality checks

In addition to the quality monitoring possibilities described in D4.1 – Section 5.4., an adjudication of disagreement though strictly applying the manual annotation guideline will be established after the piloting phase in transition to the productive phase (June 2023 - April 2024). This process involves the intervention of the annotation expert of each centre (MUG, MAASTRO/UM and NEMC), who can review the conflicting annotations and make a final decision based on the task definition of the annotation guideline. Additionally, each site keeps the documentations on questions arising, which will be discussed in detail and resolved within the ongoing bi-weekly train-the-trainers sessions. Adjudication helps resolve discrepancies and ensures the accuracy and quality of the annotations. Quality control checks will be performed by establishing a feedback loop with the annotators to provide regular evaluations and constructive feedback on their performance. This iterative feedback process helps maintain and enhance the quality of the annotations over time and the addition of available data. Possible necessary updates to the annotation guideline after adjudication, will be accessible via a shared document throughout all the partners.

7.2.1 De-identification

Based on the de-identification procedure described in D4.1 – Section 4.2 this process is further strengthened that via the manual annotation process possible still existing Protected Health Information (PHI) is identified. The PHI information has to be masked in the corresponding document and reported back to the local data steward responsible for the de-identification process.

8. Persisting annotated datasets

There are only a few clinical annotated data sets in German, even less so in Dutch and in Estonian. Considering the value of the AIDAVA annotated data sets for further research projects, and the time spent annotating them, it is important to keep them for further research. However, we need to ensure compliance to data privacy and data protection.

We expect that most narratives will be extracted from attributes with specific focus (e.g. notes on histology, notes on surgery, etc.) and therefore will already be anonymous; in addition, as part of the extraction process, data will be de-identified (see D4.1 – Section 4.2 on Ethical Committee approval). However, we cannot have complete certainty that some narratives will not contain some PHI. To ensure data privacy, we included an additional check (see Section 7.2 of this document) to ensure that any occurrence of PHI is removed, and ensure that the whole annotation process delivers anonymized annotation data sets.

From a data protection perspective, these anonymized datasets will not be made public but will remain under full control of the institutions who produced them, respectively MUG, UM and NEMC; their reuse would be subject to local Ethical Committee approval. Availability of these data sets will however be communicated through the AIDAVA website and through the OpenScience Cloud. Researchers who are interested in reusing them could submit a request to each of the local institutions.

9. Next steps

Annotator team. Manual Annotation is a dynamic process that allows the annotation team to learn through practical experience. Although the annotation guideline itself will remain unchanged, addition of concepts with additions to SNOMED CT codes and LOINC codes in the AIDAVA Reference Ontology may be necessary during the productive phase. Furthermore, regular meetings with the annotation teams provide an opportunity for valuable feedback, which may lead to necessary updates and modifications. All these updates, along with best practices learned from the teams, will be documented in a shared document that will be included in the report to be delivered together with the annotated datasets in April 2024. Furthermore, the planned manual annotators will receive training from local trainers to ensure their proficiency in the annotation process.

Train the trainers. The train-the-trainers session will continue in a bi-weekly session, where updates, feedback from the local annotators and questions will be discussed (see Section 7.2). Within this session, possible needed updates to the annotation guidelines will be decided and shared via the document mentioned above and accessible from all three responsible institutions.

Coordination Ontology team. Responsible people from the reference ontology team are invited to the regular Task 4.3 sessions (“updates”, “train-the-trainers”). The main rationale of this is to avoid decisions of using international standardised knowledge resources as part of the manual annotation process, which will not be used for knowledge graph (see Section 3.2). Any request for change to the AIDAVA Reference Ontology – based on the needs of the annotator team – will follow the governance process defined by the ontology team.

Coordination NLP team. Communication with the NLP team will be fulfilled via participating in the regular Task 5.1 meetings. The deployment requirements of NLP tools which will be adjusted by use of the exported manually annotated data set will be discussed. A prototypical NLP adaptation scenario considering all technical requirements, to be aligned with the local site architects to ensure the infrastructure is in place, will be investigated in Q3 2023.

The final set of annotated clinical narratives is expected to be finished by 04 2024 and will be confirmed with the deliverable “D4.7 Annotated datasets (3 languages/2 TA) with report”.

10. References

[1] “Recommendation on a European Electronic Health Record exchange format,” *Shaping Europe’s digital future*. https://digital-strategy.ec.europa.eu/en/library/recommendation-european-electronic-health-record-exchange-format (accessed May 25, 2023).

[2] “European Health Data Space,” *Public Health*. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (accessed May 25, 2023).

[3] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, “Clinical natural language processing in languages other than English: opportunities and challenges.” Journal of biomedical semantics. 2018 Dec;9(1):1-3.

er than English: opportunities and challenges,” *J. Biomed. Semantics*, vol. 9, no. 1, p. 12, Mar. 2018.

[4] “Documentation,” *INCEpTION*. https://inception-project.github.io/documentation/ (accessed May 23, 2023).

[5] “INCEpTION user guide.” https://inception-project.github.io/releases/26.1/docs/user-guide.html (accessed May 23, 2023).

[6] T. Lingren *et al.*, “Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements,” *J. Am. Med. Inform. Assoc.*, vol. 21, no. 3, pp. 406–413, May-Jun 2014.

[7] M. Mikulová, M. Straka, J. Štěpánek, B. Štěpánková, and J. Hajic, “Quality and Efficiency of Manual Annotation: Pre-annotation Bias,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Jun. 2022, pp. 2909–2918.

[8] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, “A study of active learning methods for named entity recognition in clinical text,” *J. Biomed. Inform.*, vol. 58, pp. 11–18, Dec. 2015.

[9] *averbis-python-api: Conveniently access the REST API of Averbis products using Python*. Github. Accessed: May 23, 2023. [Online]. Available: https://github.com/averbis/averbis-python-api

[10] “INCEpTION user guide.” https://inception-project.github.io/releases/23.4/docs/user-guide.html (accessed May 23, 2023).

[11] D. Markwell, “6. SNOMED CT Concept Model – SNOMED CT Starter Guide – SNOMED Confluence.” https://confluence.ihtsdotools.org/display/DOCSTART/6.+SNOMED+CT+Concept+Model (accessed May 24, 2023).

[12] D. Markwell, “Expression constraint language – specification and guide.” https://confluence.ihtsdotools.org/display/DOCECL (accessed May 24, 2023).

[13] P. G. Williams, “Free SNOMED CT set for FHIR - SNOMED on FHIR - SNOMED Confluence.” https://confluence.ihtsdotools.org/display/FHIR/Free+SNOMED+CT+set+for+FHIR (accessed May 24, 2023).

[14] HIPAA Journal, “What is Considered PHI under HIPAA? 2023 Update,” *HIPAA Journal*, Feb. 15, 2023. https://www.hipaajournal.com/considered-phi-hipaa/ (accessed May 24, 2023).