DELIVERABLE

Project Acronym: **ASSESS CT**

Grant Agreement number: **643818**

Project Title: **Assessing SNOMED CT for Large Scale eHealth Deployments in the EU**

D2.1 – Multilingual and multidisciplinary study of terminology coverage and quality - interim report

Authors:

Stefan Schulz	Medical University of Graz (Austria)
Jose Antonio Miñarro-Giménez	Medical University of Graz (Austria)
Daniel Karlsson	University of Linköping
Kirstine Rosenbeck Gøeg	University of Ålborg
Kornél Markó	Averbis GmbH

Project co-funded by the European Commission within **H2020-PHC-2014-2015/H2020_PHC-2014-single-stage**		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services	
RE	Restricted to a group specified by the consortium (including the Commission Services	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

Revision	Date	Author	Organisation	Description
1	2016-02-23	Stefan Schulz	MUG	First Draft
2	2016-02-24	Jose Antonio Miñarro-Giménez	MUG	Provision of data
3	2016-02-26	Daniel Karlsson	LIU	Provision of data, addition of text
4	2016-02-27	Stefan Schulz	MUG	Pre-final 1
5	2016-02-27	Daniel Karlsson	LIU	Pre-final 2
6	2016-02-28	Kirstine Rosenbeck Gøeg	AAU	Review and suggestion for restructure of method sections
7	2016-02-28	Stefan Schulz	MUG	Review and suggestion for restructure of method sections
8	2016-03-08	Dipak Kalra	Eurorec	Final review
9	2016-03-09	empirica	empirica	Editorial revision
10	2016-03-09	Stefan Schulz	MUG	Final version

Date of delivery	Contractual:	29.02.2016	Actual:	10.03.2016
Status	final			

Abstract (for dissemination)	Two studies are described that scrutinize the fitness for purpose of SNOMED CT, compared to other terminology settings. Terminologies were tested for coverage and agreement in clinical text annotation, as well as in manual binding to clinical information models. Whereas the latter use case showed a better performance both regarding concept coverage and agreement for SNOMED CT, the former one showed equivalence of English SNOMED CT free text annotations to an alternative, UMLS-based scenario, superiority of the Swedish SNOMED CT version, but inferiority of French and Dutch, where only subset translations were available.
Keywords	SNOMED CT, UMLS, Terminology coverage and quality, Electronic Health Records, Semantic Interoperability, eHealth in Europe

Statement of originality

Table of Content

1 Executive Summary

Workpackage 2 of ASSESS CT seeks empirical evidence for the fitness for purpose of SNOMED CT, compared to other terminology settings. As a testbed for the measurement of (i) concept coverage, (ii) term coverage, and (iii) inter-annotator agreement as quality indicators, two experiments were conducted: manual terminology annotation of a parallel corpus of clinical text snippets in six languages, and manual binding of terminology codes to clinical information models.

For the annotation use case, a SNOMED CT-only terminology setting was compared to a hybrid terminology, based on an extended subset of the Unified Medical Language System (UMLS) meta-thesaurus. For the binding use case, SNOMED CT was compared to a hybrid of four international terminologies.

The results of the text annotation experiment showed no significant superiority of the extended UMLS terminologies compared with SNOMED CT for languages in which a full translation of SNOMED CT is available (English and Swedish). The coverage of translations of SNOMED CT subsets (in French and Dutch), however, was lower than available alternatives. The benefit of the availability of synonyms could also be clearly shown. Apart from the English alternative scenario, analogously built alternative terminologies in Dutch, French, and Swedish showed much lower concept and term coverage compared with what could be shown in the English SNOMED CT scenario.

The terminology binding experiment, constructed to reflect a key standardized terminology use case, showed a better performance of SNOMED CT both regarding concept coverage and agreement compared with a set of the four widely used international standard terminologies ATC, ICD-10, LOINC, and MeSH.

The fact that SNOMED CT is a single-source product, with periodic releases, downwards compatibility and a uniform licence management, issued by an international non-for-profit organization is already an advantage over hybrid terminology settings, such as those constructed for these experiments which are partly dependent on non-European sources and provide a good coverage only for English. The outcome of our experiments suggests that SNOMED CT is capable of meeting the needs of annotating free text and binding to clinical models, in languages for which a translation exists, at least as well as an alternative hybrid solution, and better in the case of clinical model binding.

However, the restriction to SNOMED CT subsets as an alternative to large-scale terminology localisation (as done in Belgium for French and Dutch) must be carefully checked against the use cases to be addressed.

2 Introduction

2.1 About this document

This document presents results and ongoing work of ASSESS CT Workpackage 2, regarding multilingual clinical terminology coverage and quality. It is limited to the scrutiny of the languages English, Dutch, Swedish, and French, i.e. languages in which SNOMED CT is partly of fully available. Due to the nature of this report (interim), the results given here, as well as their interpretations, are still incomplete and subject to changes. This document is closely coupled to the deliverable D2.2, which builds on D2.1 and scrutinizes the use of terminologies for representing structured and unstructured clinical content. This document will be completed as the final deliverable D2.3, with a complete record of project results and a more in-depth discussion. The final deliverable will be submitted with extended annexes, including manuscripts to be submitted to scientific journals.

2.2 Goals and objectives of ASSESS CT

The goal of ASSESS CT is to improve semantic interoperability of eHealth services in Europe by investigating the fitness of the international clinical terminology SNOMED CT as a potential standard for EU-wide medical documentation.

SNOMED CT is the world's most comprehensive multilingual healthcare terminology which enables machine-processable representation of clinical content in electronic health records. Rooted in an ontological framework, SNOMED CT provides controlled terms, including synonyms and translations in different languages. The use of SNOMED CT in Electronic Health Record (EHR) is expected to improve communication and semantic retrieval, thus improving real time decision support to more accurate retrospective reporting for research and management. SNOMED CT is maintained by the IHTSDO, an international standards development organisation. Currently it has 28 member countries, in which SNOMED CT can be freely used. It is distributed in English and Spanish, with other translations being provided by member countries, such as Swedish and Danish (completed), as well as French and Dutch (in development). However, some important EU languages are not served, such as German, Italian and East European languages, and numerous EU member states are not IHTSDO members e.g. Germany, France, Italy, Finland, Austria, Ireland, Hungary, Croatia, Romania, Bulgaria and Greece.

As health care systems are organized nationally, the EU has not taken any steps so far towards the adoption of a standardized health terminology. However, as the mobility of EU citizens is increasing and national boundaries are loosened for a more internationalized market for health care services, the question of interoperability of health care data gains importance at a European level. The ASSESS CT consortium is addressing this challenge by investigating the current use of SNOMED CT, analysing reasons for adoption / non-adoption, and identifying success factors, strengths and weaknesses related to SNOMED CT and to alternative terminologies.

ASSESS CT makes use of diverse methodological approaches, like literature reviews, surveys, focus groups interviews, and workshops. It scrutinizes the current state of use of SNOMED CT and the fulfilment of semantic interoperability use cases, known technical and organisational drawbacks, and the way the terminology is improved and maintained. It

analyses the impact of SNOMED CT adoption from a socio-economic viewpoint, encompassing management, business, organisational, and governance aspects.

SNOMED CT adoption is scrutinized against two alternative scenarios, *viz.* (i) to abstain from actions at the EU level, and (ii) to devise an EU-wide semantic interoperability framework alternative without SNOMED CT. These scenarios were addressed in WP2 through three different terminology settings: SNOMED CT only (SCT_ONLY), a UMLS-derived alternative terminology set (UMLS_EXT), and a German-only terminology setting (LOCAL) corresponding to a scenario where each country maintains their own terminology without or with minimal EU level coordination.

The connection between the terminology settings and the three alternative scenarios exposed in the workplan (ADOPT = EU-wide interoperability by using SNOMED CT; ALTERNATIVE = EU-wide interoperability by using terminologies; ABSTAIN = no action at EU level) is not straightforward, due to the following reasons:

- SCT_ONLY scrutinizes SNOMED CT in isolation (for methodological reasons), whereas the ADOPT scenario is not exclusive to the introduction of SNOMED CT.
- UMLS_EXT exposes a setting without SNOMED CT, thus addressing the scenario ALTERNATIVE. However, because several sources in UMLS_EXT are localised and in use in EU member states, the scenario ABSTAIN is also addressed.
- Finally, the setting LOCAL provides a picture of what is possible with an optimised mix of terminologies that already exist for one language. Thus, the scenario ABSTAIN is addressed.

2.3 ASSESS CT Workpackage 2

The ASSESS CT Workpackage 2 is conducting comparative studies, all of which attempt to answer the following two questions:

- How well does SNOMED CT address selected use cases, compared to an alternative setting, which uses a mix of existing terminologies without SNOMED CT, adapted to the needs of EU member states?
- How well does SNOMED CT address selected use cases, compared to the current state of affairs, i.e. sticking to the terminologies already in used across EU member states?

All use cases are committed to the overall goal of semantic interoperability, i.e. the meaningful exchange of clinical data within and across linguistic and institutional borders. The leading hypothesis is that the more meaningful content is maintained in this exchange process, the better patient safety and cost-effectiveness in health care delivery and preventive medicine is assured. This is closely coupled to the second hypothesis, namely that this requires the formalization and standardization of meaning across languages, countries, and medical domains. As a result, semantic artefacts are required to introduce language-independent meaningful units (commonly referred to as concepts) in a precision and granularity sufficient for clinical documentation purposes across clinical disciplines and specialties. These concepts should ultimately be unambiguous by means of formal or textual definitions, as well as due to fully specified names. On the other hand, they should also be compatible with synonyms frequently applied by clinicians.

Workpackage 2 addresses three use cases:

- Use of SNOMED CT vs. other terminologies for manual annotation of clinical texts in different languages. This is mainly justified by the fact that natural language documents contain the terms clinicians use in their daily practice. The more easily these terms can be linked to concepts in a terminology, the higher is its ease of use, which is an important quality criterion of a clinical terminology. This depends on two aspects, *viz.* (i) the granularity of content provided by the concepts in the terminology

(concept coverage) and (ii) the wealth of clinically relevant synonyms or entry terms in the terminology (term coverage). Another quality criterion is inter-annotator agreement. Inter-annotator agreement measures the ease of selecting terminology content consistently: the more the annotation results coincide between two or more annotators, the more precisely defined and/or self-explaining is the terminology.

- Use of SNOMED CT vs. other terminologies for providing textual values for structured data entry forms. Despite the predominance of text, structured data entry is increasingly important in clinical documentation, especially for clinical research, quality monitoring, disease registries, health management and billing. The structuring of clinical information is provided by binding the meaning of the data elements of information models to external terminologies and by constraining value sets for coded data elements. Here also, inter-annotator agreement is an important characteristic to track as it will resemble the actions of documenting clinicians when they select or co-ordinate terms into structured templates, for which consistency of choice leads to semantic interoperability between them.
- Use of SNOMED CT vs. other terminologies for machine annotation of clinical text in different languages. The main rationale is the fact that natural language continues being the main carrier of clinical information, in original clinical documents like findings reports as well as free-text patient summaries. The ongoing adoption of Electronic Health Record (EHR) systems, is substituting paper charts by computer-based charts, but often with no change of content structure, which contains highly compacted text, often with idiosyncratic and error-laden terms and passages. Natural Language Processing (NLP) has developed powerful tools and techniques to analyse human language and map its content to controlled terminologies. This use case uses an off-the-shelves text processing pipeline tailored to several languages.

All use cases provide indicators for SNOMED CT's theoretical fitness for use. As technical fitness for use is a prerequisite for clinical fitness for use, and samples of clinical data are used for the studies, clinical fitness for use can be indirectly assessed. The evidence created by the studies proposed in WP2 is assumed to disseminate knowledge about the current state of SNOMED CT, in order to inform policy dialogues and strategic planning processes that are necessary to set the course for EU-wide clinical reference terminologies.

This deliverable abstracts from the use cases and focuses primarily on the measurement of terminology coverage and quality as study endpoints. The experiments needed for this are, however, intertwined with the manual annotation / coding use cases.

3 Customized resources

3.1 Terminologies

3.1.1 Manual annotation of free text experiment

In order to respond to the overall requirements of ASSESS CT, *viz.* comparing SNOMED CT to alternative terminologies, two custom terminology settings are described in the following. All of them were filtered by selected UMLS Semantic groups[1]. These groups constitute pairwise disjoint divisions of all concepts in the UMLS Metathesaurus. Via SNOMED CT – UMLS mappings, the same semantic groups are also used to partition SNOMED CT.

- SNOMED CT, international version (English) August 2015, Swedish version, as well as Dutch and French fragments provided by the Belgian government where the terminology is currently being localised, were included. Only concepts from selected UMLS semantic groups were used, in order to exclude terminology content that is outside of the clinical domain proper. We also excluded the SNOMED CT "Situation" hierarchy, which provides pre-coordinated concepts to express context like negation, certainty, time etc. The reason for this is that in the context of this study we consider this as belonging to information models, not terminologies, with the "Situation" hierarchy constituting an information model inside SNOMED CT.
  This terminology setting is named SCT_ONLY.
- In order to address alternative settings, we created an alternative hybrid terminology, including terminologies already in use. Starting point was the UMLS Metathesaurus, a repository of over 160 biomedical terminologies in different languages, with all of their content linked to unique identifiers (CUIs). Criteria for inclusion in the ASSESS CT alternative setting were sources that are actively updated[2]. From these, the following sources were excluded: (i) sources the use of which makes only sense in an U.S. context, such as U.S: drugs, (ii) sources in languages other than English, Dutch, French, German, Swedish, Finnish, (iii) SNOMED versions and Read code versions, (iv) sources out of scope regarding our data (nursing, dentistry).
  Additional localised terminologies were added, for which the English version was part of the UMLS selection, e.g. MeSH, ATC, ICD in several languages.
  This terminology setting is termed UMLS_EXT.

The third terminology scenario LOCAL, is not considered in this interim deliverable.

[1] ANAT = Anatomy, CHEM = Chemicals & Drugs, CONC = Concepts & Ideas, DEVI = Devices, DISO = Disorders, GENE = Genes & Molecular Sequences, LIVB = Living Beings, OBJC = Objects, PROC = Procedures

[2] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/active_release.html

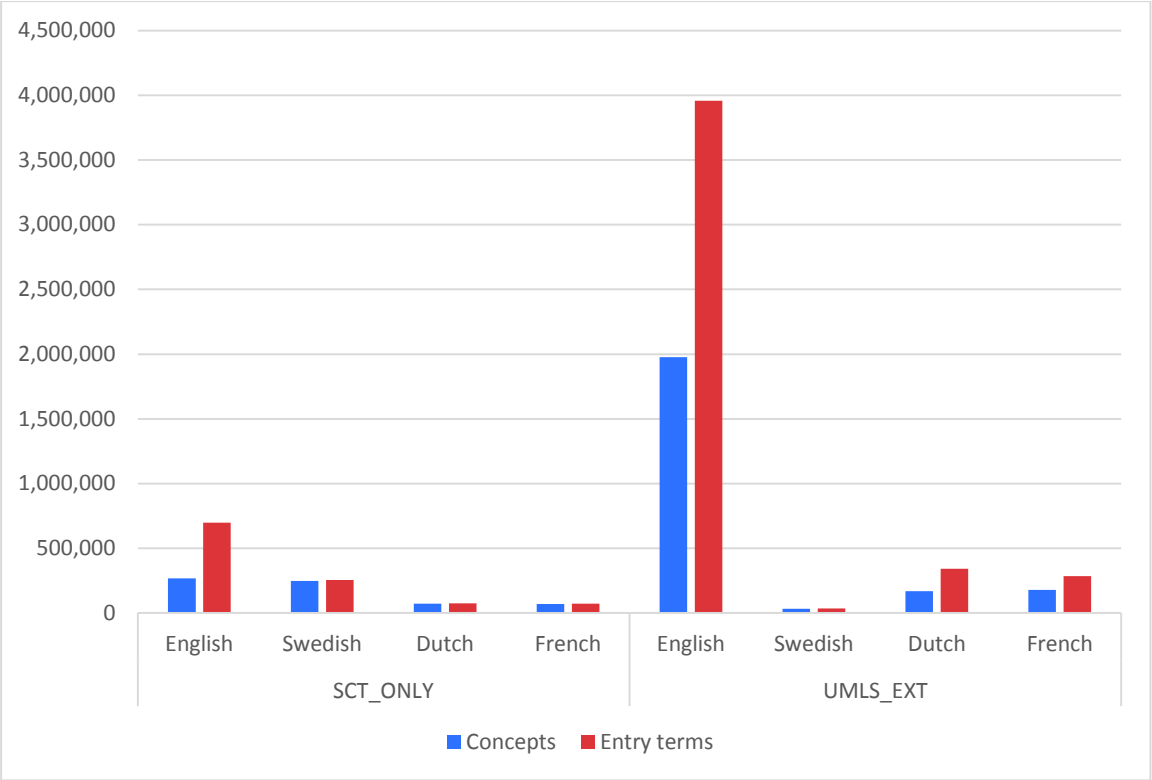

Fig. 1: Number of terms and concepts in the experimental settings and scenarios. Concept numbers ranging from 31,918 (Swedish UMLS) to 1,977,437 (English UMLS)

3.1.2 Terminology binding experiment

For the structured annotation task, a smaller set of terminologies was selected, without reference to UMLS. The complete International Release of SNOMED CT in July 2015 was used for the SNOMED CT setting and the complete current international or English versions of ATC, ICD-10, LOINC, and MeSH were used for the alternative setting. The terminologies of the alternative settings were prioritized so that annotators should look for codes in the first three before considering the last one. The first three of the alternative setting terminologies are all widely used in healthcare while the last one, MeSH, is not. However, MeSH is in comparison a fairly large terminology with larger coverage in domains not covered by the other three.

3.2 Corpora for free text annotation

For both the manual and machine annotation tasks, a multilingual corpus was necessary. To this end, clinical texts in six languages (Dutch, English, French, Finnish, German, and Swedish) were collected by the consortium partners. The acquisition of corpora was done in a way supposed to approximate representativeness in terms of clinical domains, document sections, and document types. Finally, 60 document snippets (400 – 600 characters), 10 for each language were selected.

Apart from the language, the snippets were characterized by document type (3 autopsy reports, 1 death certificate, 30 discharge summaries, 2 microscopy reports, 1 outpatient summary, 3 pathology reports, 5 referral reports, 4 finding reports, 1 toxicology report, 10 visit reports), and document sections (characterized as conclusions (3), diagnosis (2), evolution (7), findings (22), history (10), history & diagnosis (1), imaging (1), indication (1), lab (4), lab/medication (1), medication (2), order (1), plan & finding (1), recommendation (1), summary (3)). The clinical disciplines were represented as follows: Anaesthesiology (1), Dermatology (3), Gynaecology (2), Internal Medicine (17), Neurology (3), Ophthalmology (1), Paediatrics (3), Pathology (12), Surgery (14), Urology (4).

Table 1. One snippet example in six languages (original language Dutch)

Dutch	English	Finnish	French	German	Swedish
Echo nieren **Enige dilatatie van pyelum en calyces linker nier, passend bij milde hydronefrose links. Linker ureter niet evident verwijd.** **Rechter nier: geen bijzonderheden.** **X-buikoverzicht** **Foto goed beoordeelbaar, goede belichting. Wat gas in de darmen. Linker nier is licht vergroot.** **Concrement van 1.5 x 0.5 cm mediodistaal van linker nier t.h.v. L3 passend bij proximale (ureter)steen links.** **IVP (negatieve score indien aangevraagd)** **Foto goed beoordeelbaar. Goede belichting.** **Geen afwijkingen aan bot of weke delen.**	Kidney ultrasound Certain dilatation of renal pelvis and calices of the left kidney, matching mild hydronephrosis on the left side. Left ureter not clearly dilated. Right kidney: normal. Abdominal x-ray Image easily assessable, good lighting. Some gas in the bowels. Left kidney is slightly enlarged. Concrement to 1.5 x 0.5 cm mediodistal of the left kidney on level L3 matching proximal (ureteral) calculus on the left. IVP (negative experience when requested) Image easily assessable. Good lighting. No abnormalities of bone or soft tissue.	Munuaisten ultraääni Vasemman munuaisen pyelonin ja maljan tietty dilataatio, mikä sopii vasemman puolen lievään hydronefroosiin. Vasen virtsanjohdin ei ilmeisesti laajentunut. Oikea munuainen: ei mitään huomiota herättävää.Röntgen-katsaus vatsasta Kuva helposti arvosteltavissa, hyvä valaistus. Vähän kaasua suolessa. Vasen munuainen on hieman suurentunut. 1.5 x 0.5 cm kokoinen konkrementti mediodistaalisesti vasempaan munuaiseen nähden korkeudella L3 sopien vasemman puoleiseen proksimaaliseen (ureter-)kiveen. IVP (negatiivinen tulos, jos vaaditaan) Kuva helposti arvosteltavissa. Hyvä valaistus. Ei poikkeamia luissa eikä pehmytosissa.	Ultrason des reins Certaine dilatation de pyélon et calices du rein gauche, assortie à la légère hydronéphrose à gauche. Uretère gauche n'est pas manifestement élargi. Rein droit: aucune particularité. Vue d'ensemble du ventre aux rayons X Image bien interprétable, bon éclairage. Un peu de gaz dans l'intestin. Rein gauche légèrement agrandi. Concrétion de 1.5 x 0.5 cm médiodistal du rein gauche à la hauteur de L3 assortie au calcul proximal (d'uretère) à gauche. UIV (résultat négatif, sur demande) Image bien interprétable, bon éclairage. Pas de divergences des os ou tissus mous.	Ultraschall der Nieren Gewisse Dilatation von Pyelon und Kelchen der linken Niere, passend zur milden Hydronephrose links. Linker Harnleiter nicht offensichtlich erweitert. Rechte Niere: keine Auffälligkeiten.Röntgen-Abdomenübersicht Bild gut beurteilbar, gute Beleuchtung. Etwas Gas im Darm. Linke Niere ist leicht vergrößert. Konkrement zu 1.5 x 0.5 cm mediodistal der linken Niere auf Höhe von L3 passend zu proximalem (Ureter-)Stein links. IVP (negatives Ergebnis, wenn angefordert) Bild gut beurteilbar. Gute Beleuchtung. Keine Abweichungen an Knochen oder Weichteilen.	Ultraljudsundersökning av njurarna Viss dilatation av pyelon och kalkar på vänster njure, verkar vara mild vänstersidig hydronefros. Vänster urinledare är inte synbart vidgad. Höger njure: inget anmärkningsvärt. Röngten av buk Bild är lätt att bedöma, bra belysning. Lite gas i tarmen. Vänster njure är en aning förstorad. Konkrement på 1,5 x 0,5 cm mediodistal på vänster njure med höjden L3 vilket tyder på proximal vänstersidig (ureter-)sten. IVP (negativt resultat, om det krävs) Bild lätt att bedöma. Bra belysning. Inga avvikelser vad gäller ben eller mjukdelar.

Each snippet was translated into all other languages by professional translators. Due to the mediocre performance of this translation service, all translations were reviewed and corrected, especially requiring fixes of the translation of acronyms and abbreviations, and the normalization of (non-translatable) drug names to active ingredient names. The output, a parallel corpus consisting of 60 text snippets per language, was tokenized in order to generate the input for the manual and machine annotation experiments.

For the experiments described in this document, only the English, French, Dutch, and Swedish texts were used, as no SNOMED CT translations are available for German and Finnish.

3.3 Clinical models for terminology binding

For the purposes of terminology binding experiment, elements of clinical information model extracts (for coded data only) were bound to terminologies. Two kinds of terminology binding are distinguished, *viz.* (i) the binding of single nodes of an information model to terminology concepts, and (ii) the binding of sets of allowed values to terminology concepts. Criteria for selecting clinical information models were set up: clinical models should be in routine use in healthcare, encompass information generated by different professions, cover both primary and secondary (e.g. health registries) information use cases, cover a range of health specialties and different levels of granularity, and cover both common and rare cases. In addition, the models should be sourced from a variety of member states. Finally, they should cover different technical aspects of binding such as binding to attributes as well as values. The following set of information models collectively, but not always individually, address all these criteria:

- The SemanticHealthNet Heart Failure Summary Smoking status and Heart failure symptoms
- The epSOS patient summary Allergy kinds
- Trauma registry observables (DE)
- Medications from a Heart failure registry (SE)
- COPD PROMs (SE)
- Allergens from Intermountain Healthcare (US)
- Biological relations for patient transfer data, based on HL7 RoleCode value set (NL)
- Spirometry observables from a COPD examination program (DK)
- Headache location (anatomy) from a structured data entry form (US)
- Blood pressure measurement details from the Detailed Clinical Models repository (NL)
- Diseases under surveillance by the ECDC

All elements in the information model extracts were translated to English.

4 Study endpoints and protocol

The endpoints of this study are the following indicators for terminology coverage and quality. In particular, we compute:

- Concept coverage: the degree of successful representation of the content of structured or unstructured samples. In their totality, these samples are supposed to be typical and representative for the clinical content against which the fitness for purpose of the terminology settings is assessed.
- Term coverage: Given that conceptual coverage is present, term coverage measures the degree by which the language used in the source to represent that content shows a (close) match with the terms used in the terminology scenario under scrutiny.
- Inter-annotator agreement (on concepts): The more two annotators propose the same codes for representing the same resource; the less ambiguous is the terminology setting. We can therefore take the inter-annotator agreement as an indicator for the ease of selecting concepts consistently, as an important quality criterion of a terminology (setting).

We expect several interdependencies between these endpoints as well as with the overall size of each terminology scenario. E.g. with a low term coverage we expect a negative impact on the measurement of concept coverage, as the correct concepts may be missed due to poor retrieval within the browsing tool. Furthermore we expect that inter-annotator disagreement grows with the size of a terminology setting, just because the choice of similar concepts will be more. As Fig. 1 demonstrates, the size of the terminology settings used ranges over two orders of magnitude.

A study protocol was elaborated among the WP2 group. It describes the subsequent processes, from the collection of material, the definition of study purposes and endpoints, the tooling and training of the data acquisition process, and finally, the data analysis. The study protocol is available as annex to this document.

5 Tooling

5.1 Terminology browsers

Terminology browsers constitute the interface between human coders/annotators and the terminology. *For the free text annotation experiment* a customized terminology browser (Averbis TermBrowser) was developed. Its web-based user interface allows for selection of the terminology setting (SNOMED CT vs. other settings, see below) in the language of interest. The browser supports several search options, e.g. wildcard search. It is used for term selection for the manual annotation of free text. We chose this customized browser to ensure that the two settings appeared as similar as possible to the annotators.

In the terminology binding experiment, external browsers were used, *viz.* the IHTSDO SNOMED CT browser, the WHO browser for ICD-10 and ATC, the MeSH browser by the U.S. National Library of Medicine, and the LOINC browser by Regenstrief Institute, Inc. These native terminology browsers gave annotators the complete view of the terminology as intended by the terminology developers, including the native hierarchy with terminological parents, children and siblings.

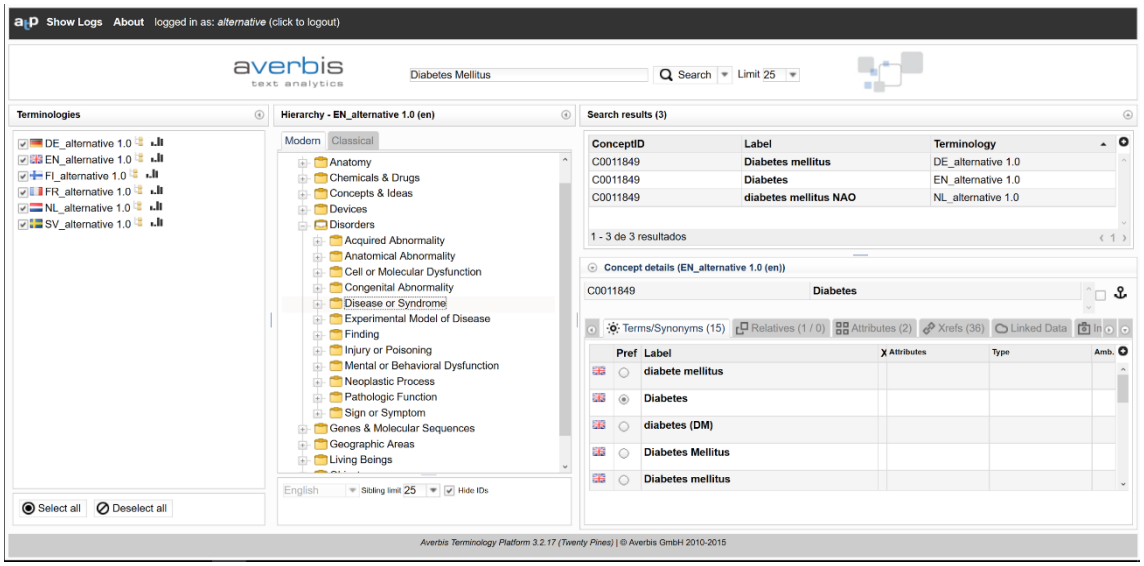

Fig. 2 – Averbis TermBrowser, used for the manual semantic annotation of clinical text

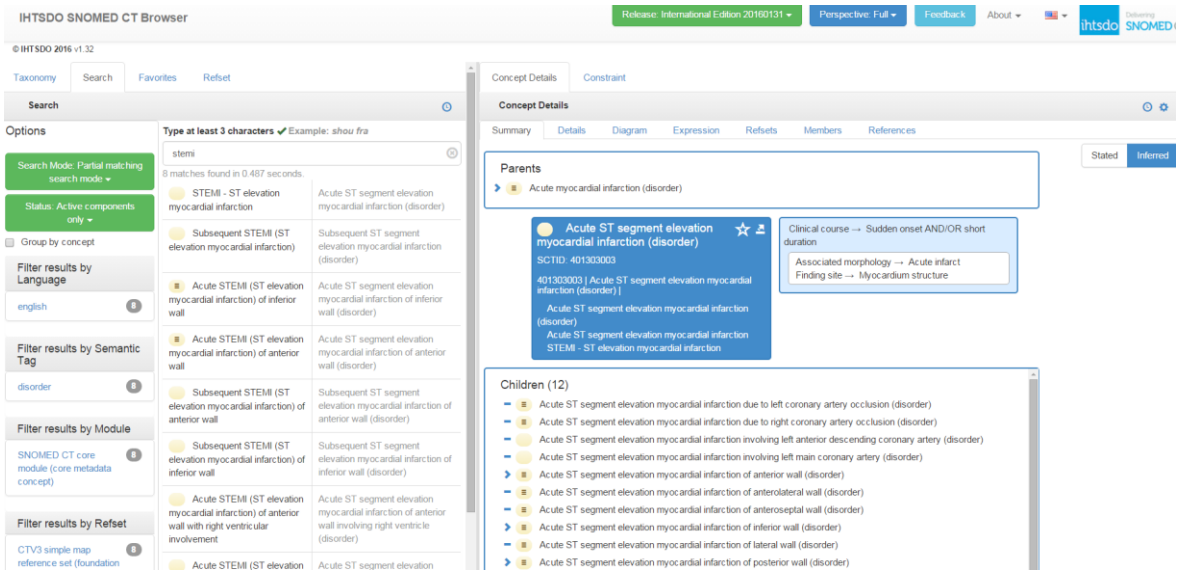

Fig. 3 – IHTSDO browser, used for the annotation of structured information models with clinical content

	TOKENS	CHUNK	CODE SNOMED ID	CONCEPT COVERAGE SCORE	TERM COVERAGE Y/N	CODE UMLS CUI	CONCEPT COVERAGE SCORE	TERM COVERAGE Y/N
3								
4	A							
5	40							
6	years							
7	old							
8	female							
9	with							
10	history							
11	of							
12	non-ST-elevation	1	40131400	Partial cov	no	C1561921	Full cov	no
13	myocardial	1	40131400	Partial cov	yes	C1561922	Full cov	no
14	infarction	1	40131400	Partial cov	yes	C1561923	Full cov	no
15	in							
16	2016-09-30							

Fig. 4: Fragment of an annotation spreadsheet. The leftmost column contains the tokens of the text snippets. The chunk ID (here "1") is given by the annotator to identify clinically relevant, cohesive text fragments (mostly noun phrases), to which then a terminological representation is assigned, together with ratings of concept coverage and term coverage. The spreadsheet shows the coding with SCT_ONLY (centre) and UMLS_EXT (right)

5.2 Manual data entry

For the free text annotation experiment, data entry was supported by an Excel Spreadsheet template, which presents the text to be annotated with one line for each token. The users enter the codes(s) retrieved, together with scores for concept and term coverage. As the same sheet is used for both scenario and the completed sheets had to be handed over to the study coordinator at short notice, period effects could be avoided.

In the terminology binding experiment the annotators entered codes and scores via web-based forms. For each element of the information model extracts, the annotators entered the code corresponding to the meaning of the element and the terminology used, the assessment of the coverage of that meaning, and also any comments related to the annotation.

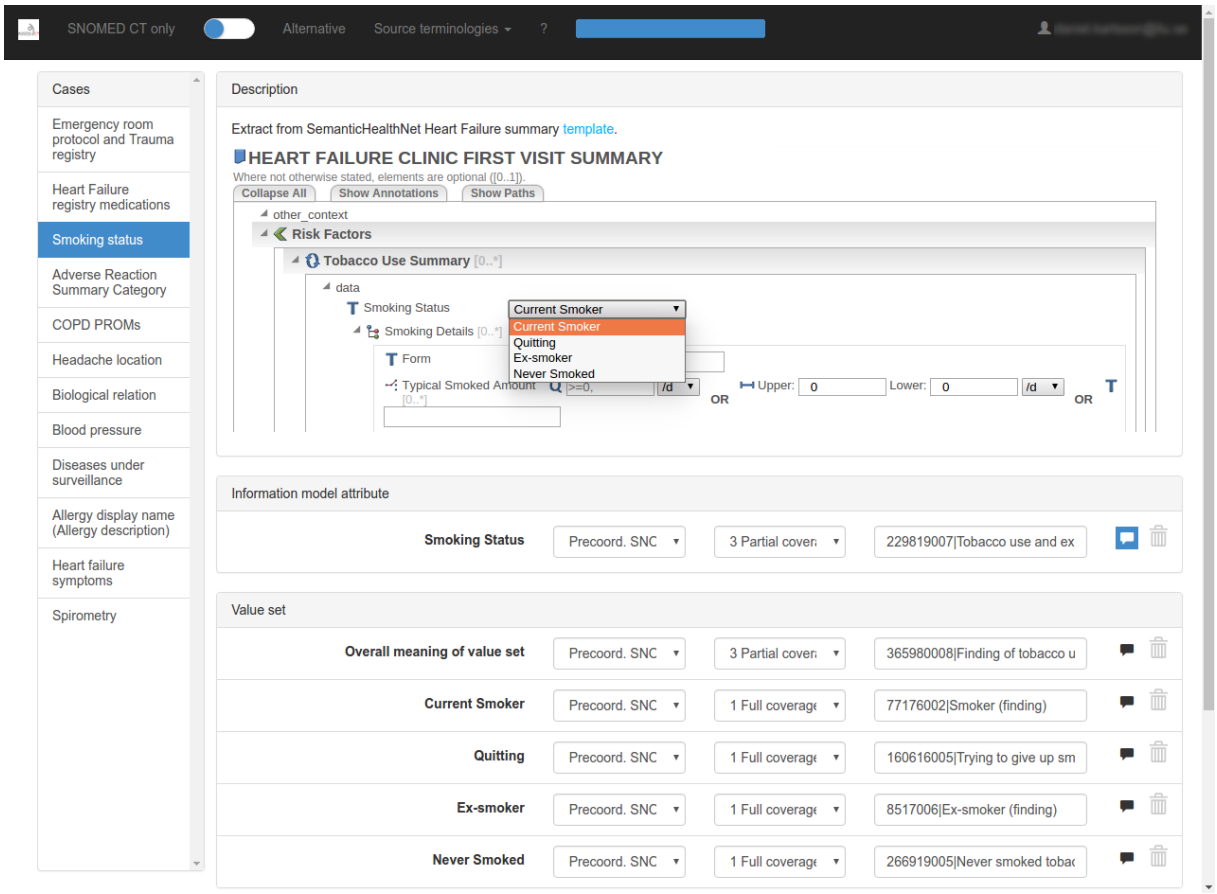

Fig. 5: Terminology binding data entry tool

5.3 Analysis of results

The analyses of the manual annotations include the description of the manual annotations, and the calculation of the concept coverage, term coverage and inter annotator agreements. These analyses were carried out by developing specific software in Java and R. The Java software, first, analyses the resulting annotation file to provide the statistics of its content; second, calculates the concept coverage; third, obtains the term coverage; and, finally, generates the input files that are required for the calculation of inter-annotator agreements. The R scripts make use of agreestat functions (http://agreestat.com/r_functions.html) to calculate the inter-annotator agreements with Krippendorff's alpha measure on sentence level agreement. In order to compare the results from different annotators, the annotations are grouped at sentence level into units; therefore, each unit gathers all annotations in a sentence. Percentage agreement between units was used as a weight function in the Alpha calculation. Then, agreements are calculated for each setting and language. IAA is measured in two modes, the strict mode, which takes into account the annotation with values "full" and "inferred" for concept coverage, and the loose mode which utilizes the annotation with the values "full", "inferred" and "partial" for concept coverage.

All materials and softwares to reproduce the results of this project are available in github (https://github.com/joseminya/ASSESSCT_WP2_T2-3). To compare results between terminology settings, the Wilcoxon signed rank test was used.

For the terminology binding experiment, data from the web-based tool were imported into a MySQL database for further processing. Analyses were carried out using R scripts with data sets prepared using Java software and MySQL database queries. Agreement was calculated using the agreestat R functions while other statistics were calculated using built in R functions. Krippendorff's alpha was used as a measure of agreement. Agreement measures for agreement on annotation were weighted by the semantic distance between codes for both the SNOMED CT and the alternative setting to allow for the structure of the terminology to influence agreement. The agreement for two annotators disagreeing with closely related codes should be higher than if the codes are not as closely related. The Lin semantic distance measure[3] was applied as a weighting factor in the agreement calculation. Krippendorff's alpha was also calculated for agreement on coverage assessment. A chi-square test was used to determine difference in coverage between terminology sets.

[3] D. Lin. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, Madison, WI (1998) p. 296–304

6 Methods

6.1 Recruiting and training of annotators

For the corpus annotation task, one person per language assumed the task to recruit domain experts. As a main condition, all annotators must have a medical background, i.e. at least advanced medical students. For each language, two or three annotators were found. All of them received annotation training in a Webinar, based on an annotation guideline (see annex). The annotation guideline specifies inclusion / exclusion criteria, the scoring system for term coverage and concept coverage, the mechanisms for assigning more than one code for a token, or one code to two and more codes, and the grouping of tokens into clinically meaningful noun phrases.

For the terminology binding experiment, data were acquired by consortium members from five different countries as well as one additional participant from the US in order to reflect the cross-border perspective. The participants were selected or chose to join the study based on their expertise in the field of terminology binding i.e. experience in the multidisciplinary field of health terminologies, information models, and medicine as such. However, given the broadness of such competences, terminology binding guidelines as well as documents explaining the use of the tool were available to the participants.

6.2 Assignment of tasks to annotators

Whereas for the terminology binding study all annotators performed the same tasks, i.e. each annotator annotated all clinical model extracts using both SNOMED CT and the alternative terminology set, the corpus annotation required custom data entry forms (spreadsheets) for each annotator. The goal was that 40 snippets were annotated by one annotator per language and 20 snippets (the same ones for each language) were annotated twice (for one language with three annotators (French), the 20 snippets were annotated three times).

6.3 Reference standard creation

The double annotations for English were the basis of the creation of a reference standard. To this end, all annotations in which the two annotators agreed, were included into the reference standard. In case of disagreements, two other domain experts worked on consensus annotations. The disagreement cases were also extracted for qualitative analysis.

6.4 Post-processing

The resulting annotations of the manual annotation experiment were post-processed in order to reduce errors due to missing information and trivial annotation mistakes. Missing information is managed depending on the type of missing data:

- If there is a code without coverage score, then the score is manually evaluated for each case.
- If there is a coverage score without code, then the coverage score is set to "None".
- If there is a concept coverage score without term coverage, the letter is set to “no”.
- If there is a term coverage value without concept coverage score, the term coverage value is removed.
- If a token in a chunk should be annotated in two settings but there was only an annotation in one, it was checked whether (i) the token is out of scope of the experiment and then the annotation is removed or (ii) the token is within the scope and a “None” coverage score is assigned to the non-annotated setting.

Trivial annotation mistakes are errors that can be automatically detected and fixed. The following methods were applied for checking the annotations:

- A code does not belong to its terminology setting. E.g., a UMLS CUI is used to annotate a token in the SCT_ONLY setting: Here, via the UMLS MRCONSO file the corresponding SNOMED CT code is identified. In case of no direct mapping, the codes are removed and the concept coverage score is set to "None".
- When a code does not belong to any terminology, we checked whether the value in the cell above is valid and the error can be explained by MS Excel's auto-increment mechanism. In such cases the wrong code is replaced by the preceding value of the same chunk. In all other cases, the code is removed and the concept coverage is set to “None”.

For the terminology binding experiment, first trivial coding mistakes were corrected, including one case of swapped clinical model elements and one case of a missing digit in a SNOMED CT code as well as formatting errors in groupings, e.g. using the wrong character to demarcate group members. Some groupings corresponded to pre-coordinated SNOMED CT concepts and were replaced with those pre-coordinated SNOMED CT concepts. After these initial corrections, a first data set was created.

In addition to the trivial coding mistakes there were non-trivial coding mistakes due to errors made with groupings. As per the study protocol, groupings should always conform to the guidelines of the terminology used. This meant that for SNOMED CT, groupings should conform to the SNOMED CT Editorial Guidelines, and for ICD-10, the groupings should conform to the dagger-asterisk system. Only about 30 % of the groupings did conform to any such guidelines. 35% of the groupings, although not conforming to guidelines, corresponded to reasonable information model or query constructs and were thus allowed. Another 15 % of the groupings were considered close enough according to agreed and mechanically applicable rules and were corrected. The rest of the groupings were removed. All such corrections were made to SNOMED CT groupings. Coverage assessment was changed accordingly, e.g. Full coverage of a non-conformant grouping was changed to Partial coverage. All corrections were considered to be normal parts of a terminology use quality assurance scheme. After these corrections, a second data set was created.

7 Results

7.1 Results of manual annotation of text samples

Text snippets are extracts from clinical documents. They consist of (complete or incomplete) sentences, many of which contain one or more clinically significant chunks (mostly noun phrases). Chunks consist of one or more tokens. The number of sentences per snippet ranged from 8.5 (French) to 8.9 (Dutch), the number of tokens from 95 (Swedish) to 111 (French). The number of relevant chunks per snippet, i.e. token sequences with clinically relevant content, as determined by the annotators, ranged from 12 (French) to 14 (Dutch). The number of tokens within relevant chunks ranged from 26 33 (Dutch) to 54 (English). The number of tokens with semantic annotations per snippet ranged from 26 (Dutch) to 38 (English) for the SCT_ONLY setting and from 25 (Dutch) to 38 (English) in the UMLS_EXT setting.

The concept coverage of the manual annotation experiment in Task 2.3 is calculated using the concept coverage scores of each annotator. Out of the four score values ("full", "inferred", "partial" and "none") we defined:

- Strict coverage: Only "full" and "inferred" are considered.
- Loose coverage: "full", "inferred", and "partial" are considered.

Table 2 and Fig. 6 show the concept coverage per language and terminology setting. For snippets annotated more than once, coverage valued were averaged by the number of annotations.

Table 1: Concept coverage of the manual annotation experiment for the three settings and for five languages. The table shows the average concept coverage and the confidence interval (CI) with 95% significance.

Language	Concept Coverage SCT_ONLY				Concept Coverage UMLS_EXT			
	Strict Coverage		Loose Coverage		Strict Coverage		Loose Coverage	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
English	0.86	[0.82; 0.88]	0.92	[0.88; 0.93]	0.88	[0.86; 0.91]	0.94	[0.93; 0.96]
Swedish	0.87	[0.84; 0.89]	0.91	[0.88; 0.93]	0.59	[0.55; 0.63]	0.65	[0.61; 0.69]
Dutch	0.43	[0.35; 0.44]	0.52	[0.45; 0.55]	0.60	[0.57; 0.65]	0.67	[0.64; 0.72]
French	0.45	[0.37; 0.47]	0.57	[0.49; 0.59]	0.64	[0.61; 0.70]	0.75	[0.73; 0.80]

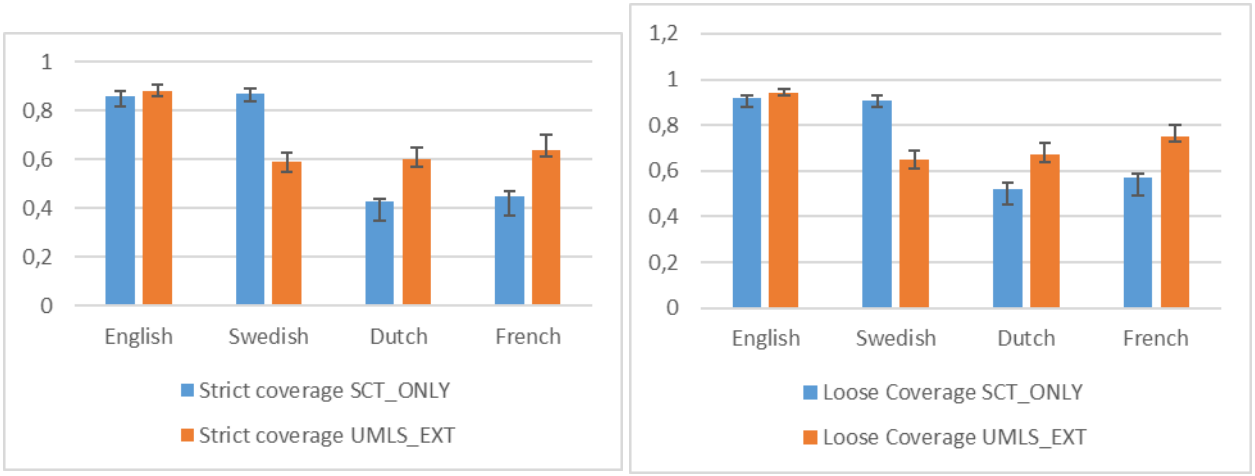

Figure 6: Concept coverage of the manual annotation experiment for the three settings, strict and loose modes, and English, Swedish, Dutch and French languages. The table shows the average concept coverage and the confidence interval (CI) with 95% significance

The results for each language are further grouped by UMLS Semantic Groups (see section on Terminologies). Table 3 provides the rate of annotations with full coverage, inferred coverage and partial coverage scores by semantic group, language and setting.

Table 2: Annotations ("Full" + "Inferred" + "Partial") for SCT_ONLY and UMLS_EXT settings and grouped by UMLS semantic groups.

	SCT_ONLY				UMLS_EXT			
UMLS Semantic Group	Dutch	English	French	Swedish	Dutch	English	French	Swedish
Absolute numbers	773	1,822	997	1,565	981	1,821	1,295	1,037
Anatomy	0.29	0.14	0.27	0.20	0.22	0.13	0.24	0.24
Chemicals & Drugs	0.08	0.09	0.06	0.14	0.14	0.09	0.16	0.24
Concepts & Ideas	0.20	0.28	0.29	0.20	0.01	0.28	0.03	0.03
Disorders	0.34	0.31	0.30	0.28	0.41	0.31	0.36	0.32
Procedures	0.07	0.16	0.06	0.17	0.22	0.19	0.20	0.16
Others	0.02	0.02	0.02	0.01	0.00	0.00	0.01	0.01

Term coverage of the manual annotations is indicated by the values "yes" and "no". Term coverage "yes" means that the terminology setting in use contains an entry term for a token or a token sequence that matches the related tokens in the snippet. This matching is flexible and allows variations in terms of word inflection and order, as determined by the annotation guidelines. Term coverage is calculated assuming that for each concept annotation there is one term coverage value. Table 4 and Fig.7 show the results for term coverage.

Table 3: Term coverage of the manual annotation experiment for "SCT_ONLY", "UMLS_EXT" and "LOCAL" settings and for English, Swedish, Dutch, French and German. The table shows the average term coverage and the confidence interval (CI) with 95% significance.

Language	SCT_ONLY		UMLS_EXT	
	Term Coverage	95% CI	Term Coverage	95% CI
English	0.68	[0.64; 0.70]	0.73	[0.69; 0.76]
Swedish	0.47	[0.44; 0.52]	0.35	[0.32; 0.40]
Dutch	0.35	[0.29; 0.36]	0.44	[0.41; 0.49]
French	0.39	[0.34; 0.43]	0.57	[0.55; 0.64]

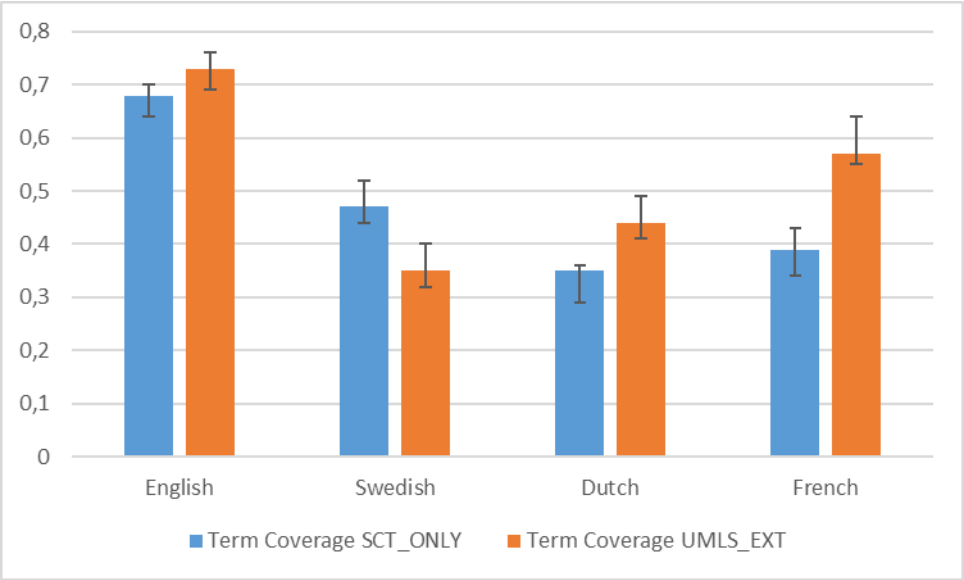

Figure 7: Term coverage of the manual annotation experiment for "SCT_ONLY", "UMLS_EXT" and "LOCAL" settings and for English, Swedish, Dutch, French and German. The table shows the average term coverage and the confidence interval (CI) with 95% significance.

Next table shows the IAA for each language and setting. The table contains the language, the Krippendorff's alpha measure for strict and loose modes and their corresponding confidence intervals with 95 % significance. Table 5 and Fig. 8 show the inter-annotator agreement results.

Table 4: Inter-annotator agreement between annotators of each language and settings. Krippendorff's alpha measure and confidence interval with 95% significance are provided for strict and loose modes.

Language	SCT_ONLY				UMLS_EXT			
	Strict Concept Cov.		Loose Concept Cov.		Strict Concept Cov.		Loose Concept Cov.	
	α	95% CI	α	95% CI	α	95% CI	α	95% CI
English	0.37	[0.33; 0.41]	0.64	[0.60; 0.69]	0.36	[0.32; 0.40]	0.64	[0.60; 0.68]
Swedish	0.30	[0.26; 0.34]	0.55	[0.51; 0.60]	0.49	[0.43; 0.54]	0.74	[0.70; 0.78]
Dutch	0.30	[0.25; 0.35]	0.55	[0.49; 0.62]	0.45	[0.40; 0.50]	0.70	[0.65; 0.75]
French	0.22	[0.17; 0.27]	0.40	[0.34; 0.47]	0.36	[0.30; 0.41]	0.57	[0.51; 0.62]

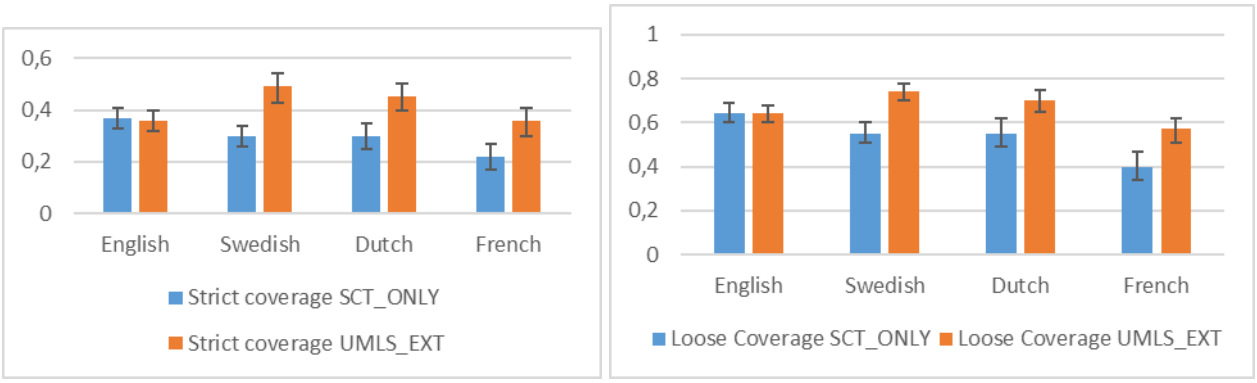

Figure 8: Inter annotator agreement between annotators of each language and settings. Krippendorff's alpha measure and confidence interval with 95% significance are provided for strict and loose modes

7.2 Results of terminology binding

There were in total 1,212 annotations made by six annotators, 606 for each of the two terminology settings. Information provided in the tables is with corrections applied (i.e. based on the second data set) if not otherwise stated.

Table 5 Use of terminologies, Alternative setting

Terminology	Number of annotations	Of which are groupings
ATC	59	7
ICD-10	140	5
LOINC	201	14
MeSH	82	1

Table 6 Use of terminology, SNOMED CT setting

Pre-coordination or Grouping	Number of annotations
Single, pre-coordinated code	562
Grouping	44

Table 7 Total concept coverage of SNOMED CT vs alternative settings

	SNOMED CT	Alternative
Full or inferred coverage	481	309
Not full or inferred coverage	125	297

The difference in concept coverage between SNOMED CT and alternative settings was significant, $X^2 = 103$, $p = 3.81*10^{-24}$. However, the agreement on the existence of a Full- or Inferred-coverage code was low (Krippendorff's alpha (95 % CI), SNOMED CT: 0.28 (0.16, 0.40), Alternative: 0.33 (0.24, 0.42), no significant difference between settings). There were many occasions where there was agreement on the code but disagreement on the coverage assessment, for example for the clinical information model element "Anaphylaxis", one annotator assessed the ICD code "T78.2 Anaphylactic shock, unspecified" to be Full coverage whereas another annotator assessed the same code as Partial coverage.

Table 8 Agreement of SNOMED CT vs alternative settings

Krippendorff's alpha (95 % CI)	SNOMED CT	Alternative
First data set, before non-trivial corrections	0.59 (0.53, 0.65)	0.47 (0.41, 0.54)
Second data set, after non-trivial corrections	0.61 (0.55, 0.67)	0.47 (0.41, 0.54)

Due to issues with groupings, the agreement for SNOMED CT coding differed between the first and the second data set, i.e. the data set before and after non-trivial corrections. After application of corrections, there was a significant difference in agreement between the SNOMED CT and the alternative setting.

8 Discussion and outlook

8.1 Main messages

8.1.1 Manual text annotation

The data obtained from manual annotation of text snippets using the two scenarios SCT_ONLY and UMLS_EXT showed slightly higher concept and term coverage for the UMLS_EXT setting, however not significant. Inter-annotator agreement for English, both for strict and loose coverage was practically the same. These results should be interpreted in the light of the fact that UMLS_EXT is about a factor of 7.5 larger than SCT_ONLY regarding the number of concepts, and about 5.5 times larger regarding terms.

Analysing absolute values, we have to acknowledge that roughly one out of ten concepts were not found, and that that the inter-annotator agreement was rather low for the strict coverage (0.37) but much higher for the loose coverage (0.64) assessment. This shows, on the one hand, the methodological problem of a rating scale for which the boundary decisions often depend from individual judgement. On the other hand, these values highlight the complexity of the task of free text annotation as such, independent of the quality of the terminology. The term coverage value shows a certain advantage (5 percentage points) of the much larger vocabulary of UMLS_EXT. This is not surprising given the much higher number of terms. For SNOMED CT, supporting further interface terms could improve results. However, concept coverage remains high regardless of lower term coverage, showing that most coders were able to make a distinction between term and concept.

Swedish is rather underserved by terminologies from the UMLS_EXT scenario, with about 32,000 concepts, and not many more terms. It is therefore not surprising that SCT_ONLY fares much better, and equals the values of concept coverage for English. That term coverage being much lower was to be expected, as the Swedish SNOMED CT version does not contain synonyms. The significantly higher agreement values for the Swedish UMLS_EXT scenario, compared to SCT_ONLY, can be explained by two factors, viz. (i) the abovementioned inverse dependency between terminology size and inter-annotator agreement, and (ii) the fact that the Swedish terminologies in UMLS_EXT are mainly ICD and MeSH, i.e. terminologies that have evolved through practical use for a much longer time than SNOMED CT.

The Dutch and French SNOMED CT versions (Belgian subsets) are characterized by the fact that their size is about one fourth of the English SNOMED CT, and that they have no synonyms. Compared to a UMLS_EXT scenario of about double the size they exhibited a much poorer behaviour in all aspects. This shows that the French and Dutch subsets are not yet fit for purpose as they fail in more than 40% of the cases covered by the English SCT_ONLY scenario. As a scrutiny by semantic groups reveals the weaknesses of the French and Dutch versions concern primarily procedures, but also drugs and modifiers.

8.1.2 Manual terminology binding

The results of the terminology binding experiment are much easier to interpret, because it relied on the International/English SNOMED CT version only and compared them not to a large hybrid terminology like UMLS_EXT, but to four widely used terminologies (ATC, ICD-10, LOINC, MeSH). Both coverage and agreement values are significantly higher for SNOMED CT. This is especially noteworthy because the alternative scenario contained LOINC which is not assumed to overlap much with SNOMED CT. In accordance with the annotation experiment, disagreement on the coverage assessment was frequent, which may question the raison d'être of the underlying ISO rating scale.

The results of the terminology binding experiment also show that the use SNOMED CT may be more complex. SNOMED CT is a more complex terminology, which is reflected in that the rules governing especially post-coordination are harder to grasp even for experienced terminology users. The alternative terminologies still has rules governing their use, but less complex than those of SNOMED CT. However, if rules are enforced, as they would be in a terminology use quality assurance program, SNOMED CT may enable more consistent terminology use. Such quality assurance measures would be strongly recommended in any terminology use project.

8.2 Limitations

The experiments exhibit several limitations:

- The translation results were of different quality dependent on the languages. Short forms and drug names remained not translated. Most but not all of this could be mitigated by the WP2 group. The decision to substitute all brand names in the sources by substance names excludes the (national) terminologies of pharmaceutical specialties from the scope of our investigations.
- The representativeness of clinical texts and clinical models is limited, due to impossibility of a good sampling approach. Consequently, both the selection of text snippets and the selection of clinical models were done manually, however by domain experts attempting to yield a high degree of representativeness.
- Due to resource constraints the amount of texts and clinical models was limited, which has an impact on the power of the study, especially when partitioning the annotation / binding results by semantic types.
- The recruitment of annotators yielded a rather heterogeneous group. Although they received on-line training and were instructed to adhere to guidelines, their performance varied which means comparisons between languages may be problematic.
- The terminology binding was partially performed by people with a medical informatics, but not a medical background, and vice versa. There was also evidence that the guidelines were only partly followed.

8.3 Additional results to be expected

Further results will be available in the final deliverable (D4.3) regarding the analysis of German and Finnish annotations.

9 Conclusion

Clinical text annotation experiments showed no superiority of a large-scale terminology hybrid constituted by existing non-SNOMED CT terminologies over SNOMED CT for languages in which a full translation of SNOMED CT is available. The advantage of SNOMED CT was especially visible for Swedish as a language with a limited coverage of clinical terminology of other sources. In contrast, the coverage of the translation of Belgian SNOMED CT subsets (French, Dutch) was insufficient compared to alternative settings.

The terminology binding experiment showed a better performance of SNOMED CT, compared to a set of four international standard terminologies.

The fact that SNOMED CT is a single-source product, with periodic releases, downwards compatibility, a uniform licence management, issued by an international non-for-profit organization, is already an advantage over hybrid terminology settings, such as constituted by the U.S. based Unified Medical Language System UMLS. Together with our experimental findings, we can testify fitness for purpose of SNOMED CT in terms of concept coverage in annotating free text at least as well as an alternative hybrid solution, and better in the case of clinical model binding.

However, term coverage in SNOMED CT was low, compared to that of alternative terminologies. Given that higher term coverage provides more usable terminologies, better term coverage would improve SNOMED CT's fitness for purpose. However, a focus must be set on terminology localisation, beyond the translation of preferred terms. The localisation of SNOMED CT subsets as an alternative to large-scale terminology translation and interface term acquisition must be carefully checked against the use cases for which the terminology is to be adjusted in order to avoid gaps and underspecification.

10 Annexes

Annex_A_ASSESS-CT-WP2_AnnotationGuidelines_.pdf

Annex_B_ASSESS-CT-WP2_StudyProtocol.pdf

Annex_C_ASSESS-CT-WP2_ResultAnnotationTablePostprocessed.xlsx

Annex_D_ASSESS-CT-WP2_ResultAnnotationTableWithoutPostprocessing.xlsx

Annex_E_ASSESS-CT-WP2_Results_Task_3.pdf