

Diagnosis clarification by generalization to patient-friendly terms and definitions: Validation study

Hugo J.T. van Mens^{a,b,*}, Savine S.M. Martens^c, Elisabeth H.M. Paiman^c,
Alexander C. Mertens^c, Remko Nienhuis^b, Nicolette F. de Keizer^a, Ronald Cornet^a

^a Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health Research Institute, Meibergdreef 9, Amsterdam, Netherlands

^b Department of Research & Development, ChipSoft B.V., Amsterdam, Netherlands

^c SNOMED CT Netherlands National Release Center, Nictiz, The Hague, Netherlands

ARTICLE INFO

Keywords:

Diagnoses
Health literacy
Patient access to records
Patient-friendly terminology
Personal health records
SNOMED CT

ABSTRACT

Background: Now that patients increasingly get access to their healthcare records, its contents require clarification. The use of patient-friendly terms and definitions can help patients and their significant others understand their medical data. However, it is costly to make patient-friendly descriptions for the myriad of terms used in the medical domain. Furthermore, a description in more general terms, leaving out some of the details, might already be sufficient for a layperson. We developed an algorithm that employs the SNOMED CT hierarchy to generalize diagnoses to a limited set of concepts with patient-friendly terms for this purpose. However, generalization essentially implies loss of detail and might result in errors, hence these generalizations remain to be validated by clinicians. We aim to assess the medical validity of diagnosis clarification by generalization to concepts with patient-friendly terms and definitions in SNOMED CT. Furthermore, we aim to identify the characteristics that render clarifications invalid.

Results: Two raters identified errors in 12.7% (95% confidence interval – CI: 10.7–14.6%) of a random sample of 1,131 clarifications and they considered 14.3% (CI: 12.3–16.4%) of clarifications to be unacceptable to show to a patient. The intraclass correlation coefficient of the interrater reliability was 0.34 for correctness and 0.43 for acceptability. Errors were mostly related to the patient-friendly terms and definitions used in the clarifications themselves, but also to terminology mappings, terminology modelling, and the clarification algorithm. Clarifications considered to be most unacceptable were those that provide wrong information and might cause unnecessary worry.

Conclusions: We have identified problems in generalizing diagnoses to concepts with patient-friendly terms. Diagnosis generalization can be used to create a large amount of correct and acceptable clarifications, reusing patient-friendly terms and definitions across many medical concepts. However, the correctness and acceptability have a strong dependency on terminology mappings and modelling quality, as well as the quality of the terms and definitions themselves. Therefore, validation and quality improvement are required to prevent incorrect and unacceptable clarifications, before using the generalizations in practice.

1. Background

1.1. Scientific background

Patients can increasingly access their healthcare data in patient portals and personal health records. However, these records are full of jargon with which most persons without a medical background are

unfamiliar. [1,2] Therefore, patient-friendly terms and definitions might help patients and their significant others understand the medical concepts in their records. [3–6] There are, however, myriads of terms that are used in the medical domain. SNOMED CT, a comprehensive medical terminology system in healthcare, contains terms for 352,568 concepts (January 2020 release). It is thus not practical, and very costly, to manually develop patient-friendly terms and definitions for each concept. Furthermore, specialists use very detailed terms that a general

* Corresponding author at: Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health Research Institute, Room J1B-109, Meibergdreef 9, PO Box 22700, Amsterdam, 1100 DE, Netherlands.

E-mail addresses: h.j.vanmens@amsterdamumc.nl (H.J.T. van Mens), savine@live.nl (S.S.M. Martens), paiman@nictiz.nl (E.H.M. Paiman), mail@sandermertens.nl (A.C. Mertens), r.nienhuis@live.nl (R. Nienhuis), n.f.keizer@amsterdamumc.nl (N.F. de Keizer), r.cornet@amsterdamumc.nl (R. Cornet).

<https://doi.org/10.1016/j.jbi.2022.104071>

Received 2 December 2021; Received in revised form 12 March 2022; Accepted 5 April 2022

Available online 13 April 2022

1532-0464/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Abbreviation

CI	95% Confidence Interval
DT	Diagnosethesaurus (“Diagnosis Thesaurus” in English)
EHR	Electronic Health Record
ICC	Intraclass Correlation Coefficient
PFE	SNOMED CT Netherlands Patient-Friendly Extension Release
SCT-Int	SNOMED CT International Release
SCT-NL	SNOMED CT Netherlands Release

Table 1

Examples of types of clarifications with supertypes, synonyms and definitions. The supertypes are bolded.

Clarification type	Example diagnosis	Example clarification
Synonym and definition	Phlebitis	Another word for “phlebitis” is inflammation of the vein: Inflammation of a vein that makes it red, swollen, and painful.
Definition	Blepharospasm	This involves the involuntary contraction of the eyelids as a cause of a disorder in muscle tension.
Supertype synonym	Reactive perforating collagenosis	A skin disorder .
Supertype synonym and definition	Familial periodic paralysis	A type of myopathy . Myopathy: This is an abnormality in the structure of the muscle tissue.
Multiple supertypes	High altitude retinopathy	A type of disorder of retina and barotrauma . Barotrauma: Damage or pain, mainly in the middle ear, because of changes in pressure.
Multiple supertypes with text filter	Congenital cyst of adrenal gland	A type of inborn abnormality and hormonal disorder . Cyst: cavities in the body filled with liquid.
Multiple supertypes with seven supertypes	Lowe syndrome	A type of inborn abnormality , mental disorder , and disorder of brain , kidney , eye , and metabolism . It is hereditary .

practitioner might not use. Laypersons may also require fewer details than specialists, general practitioners, or chronic patients. Hence, a higher level, general description might already be sufficient to clarify the meaning to a layperson.

In a former study, we showed that more than 70% of 12,453 diagnoses in the Dutch Diagnosethesaurus (“Diagnosis Thesaurus” in English, referred to as DT hereafter) can be generalized to a small set of 211 concepts with patient-friendly synonyms from the SNOMED CT Netherlands Patient-Friendly Extension (PFE), by employing the SNOMED CT “is-a” hierarchy.[7] The PFE language reference set ‘states which descriptions are appropriate to show to patients, caregivers and other stakeholders who have not received care-related training’ (definition of concept ‘15551000146102 [Patient-friendly Dutch language reference set]’). For example, ‘reactive perforating collagenosis’ can be generalized to the concept of ‘skin disorder’, which we consider more comprehensible to laypersons. Similarly, there were 1316 different types of other skin disorders in the DT that can be clarified in this manner. The PFE also contains text definitions that can be used to further explain the diagnoses with a textual description. For example, ‘familial periodic paralysis’ can be generalized to ‘myopathy’. Myopathy is not a common term, but the definition of myopathy can be provided to further explain the concept. See Table 1 for examples of diagnosis clarifications based on SNOMED CT.

Previous efforts had focused on developing and enriching lists of

Table 2

Terminology statistics: the number of definitions, preferred and acceptable synonyms in the SNOMED CT International Release (SCT-Int) and SCT Netherlands Release (SCT-NL), SNOMED CT Netherlands Patient-Friendly Extension Release (PFE), the total number of diagnoses and the number of diagnoses with terms or definitions in PFE from the Diagnosethesaurus (DT).

Terminology	Statistic	n
SCT-Int	Active concepts	352,568
SCT-NL	Active concepts	361,835
PFE	Preferred synonyms	1,409
PFE	Acceptable synonyms	138
PFE	Definitions	700
DT	Diagnoses	24,966
DT	Diagnoses with SNOMED CT mapping	18,579
DT	Diagnoses with PFE term	714
DT	Diagnoses with PFE definition	361

patient-friendly terms, by text mining consumer vocabulary and using frequency measures and term similarity to find easier terms. [8–12] However, few studies have focused on generating explanations. They were primarily focused on increasing coverage for lexical simplification. Zeng et al[13] used relationships in the UMLS to generate an explanation, such as ‘a type of lung disease’ as the explanation of ‘pulmonary emboli’, derived from the “is-a” relationships. They found some problems with hierarchical relationships, for instance, that a cyst should not be generalized to a type of tumor, because it ‘may falsely alarm or unnecessarily alienate the reader’. In a later study elaborating upon these results, Kandula et al[6] reported 32% of hierarchical explanations to be incorrect and therefore evaluated an alternative method, using other types of semantic relationships from the UMLS, such as the ‘part of’ and ‘a device used in’ relationship. In a similar application that linked medical texts to existing lay synonyms and definitions, Chen et al[14] found generalizing drugs to higher level drug class definitions was sometimes not accurate to define a drug, because drugs may be used in several treatments that were not always applicable to subclasses. However, to our knowledge, no extensive evaluations have been published about the issues involved in generalization to concepts with patient-friendly terms and definitions. Moreover, these prior studies were primarily focused at lexical simplification of texts rather than clarification of concepts with generated clarifications.

1.2. Rationale for the study

Generalization essentially implies a loss of detail, and thus a loss of information. The technique we have developed of generalization to concepts with patient-friendly terms and definitions can also lead to errors. For example, errors arise if there are mistakes in the mapping of patient-friendly synonyms and text definitions to SNOMED CT concepts, the mapping of diagnoses from the DT to SNOMED CT, or in SNOMED CT “is-a” relationships. Our technique has not been validated yet by medical doctors and terminologists. Systematic validation is necessary before implementing the generalizations in patient portals and personal health records. It can also shed light on for which terms, concepts and clarifications this technique works and for which it does not, and how it can be improved.

1.3. Objectives of the study

We aim to assess the medical validity of diagnosis clarification by generalization to supertype concepts with patient-friendly terms and definitions in SNOMED CT. Furthermore, we want to identify the characteristics of the clarifications that render them invalid.

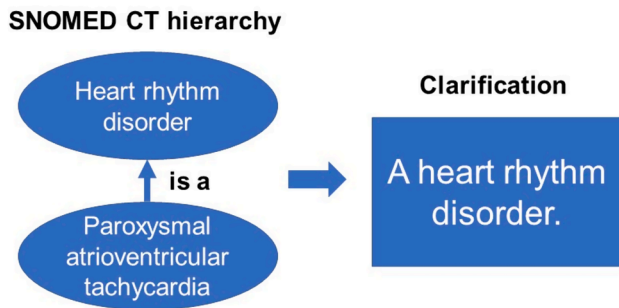


Fig. 1. The algorithm generates a clarification by generalizing a medical concept to a more general, supertype concept with a patient-friendly description from the PFE.

2. Study context

2.1. Terminology systems

We used the SNOMED CT Netherlands Edition Release (SCT-NL) together with the PFE from March 2020 (SNOMED CT Netherlands National Release Center, Nictiz, The Hague, Netherlands; referred to as NRC hereafter), and the DT version 3.10 from April 2020 (DHD, Utrecht, Netherlands). The SCT-NL and PFE are dependent upon the SNOMED CT International Release from January 2020. Table 2 shows the number of synonyms, definitions, and diagnoses in these terminologies. Only 714 diagnoses in the DT had synonyms or definitions in the PFE.

2.2. System details

For diagnoses from the DT that did not have a definition in the PFE, we generated clarifications by generalizing them to supertypes in SNOMED CT with a patient-friendly synonym and definition. See Fig. 1 for an illustration of the algorithm. Examples of clarifications are provided in Table 1. To find the supertypes, we traversed the SNOMED CT “is-a” hierarchy.[7] Because SNOMED CT is polyhierarchical, this method can produce generalizations to a single supertype (e.g., ‘Reactive perforating collagenosis’ in Table 1) or to multiple supertypes (e.g., ‘High altitude retinopathy’ in Table 1). The algorithm removed all redundant supertypes that were also supertypes of other supertypes of the concept to be clarified. For example, disease is a supertype of cardiovascular disorder, which is a supertype of heart disorder, which is a supertype of heart valve disorder, which is a supertype of aortic valve stenosis. All these supertypes have PFE terms. The clarification of aortic valve stenosis, however, only included the direct supertypes (that it is a heart valve disorder), leaving out the redundant supertypes of supertypes. The synonyms and definitions of the supertypes of the concept were concatenated into a short description that served as the clarification. When available, the clarification consisted of the synonym, an enumeration of supertypes and an enumeration of the definitions of the supertypes. For diagnoses with multiple supertypes that were disorders of body parts, the body parts were aggregated to remove redundancy (e.g., ‘Lowe syndrome’ in Table 1). If a supertype term was already part of the medical description, the algorithm did not use the term as a clarification, but did add its definition (e.g., ‘Congenital cyst of adrenal gland’ in Table 1). We developed an ASP Core.NET (Microsoft, Redmond, WA, USA) web application with a Neo4j graph database (Neo4j, San Mateo, CA, USA) to generate the clarifications.

3. Materials and methods

3.1. Study design

We performed a validation study with a randomized sample of

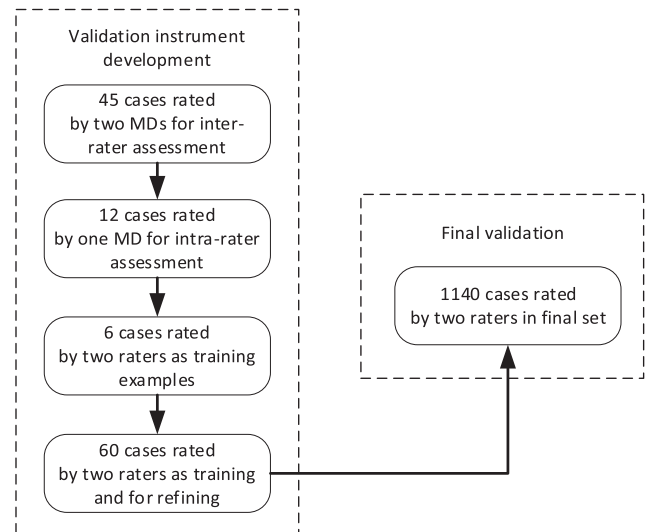


Fig. 2. Study flow.

diagnosis clarifications that was evaluated by two raters (SMA and EP). The raters have a medical background and had been involved in the translation of SNOMED CT concepts to Dutch for the SCT-NL release, at the NRC.

3.2. Study flow

3.2.1. Validation instrument development

We developed a validation instrument to evaluate the quality of the diagnosis clarifications. The validation questions were developed in three iterations. Firstly, with two medical doctors (MDs) other than the raters in the current study, we tested on 45 cases whether the instructions and validation questions were clear. Secondly, with one of the MDs other than the raters of the current study, we measured the intra-rater reliability. In this second iteration, the MD completed twelve cases after receiving the full instructions. Subsequently, he completed the same cases again a few days later. The MDs and the first author (HvM) discussed the differences in ratings and HvM improved the validation questions accordingly to disambiguate them. Additionally, the validation questions were discussed and refined after the first 60 cases with the two raters in the current study. Three cases were repeated in these first 60 cases to assess the intra-rater reliability. We sampled them randomly from the first 30 cases. We included the three repeated cases again and repeated six extra cases in the final 1140 cases (1131 unique cases).

3.2.2. Sampling and validation

We took a random sample of 1200 clarifications by supertypes. This was constrained by the estimated amount of time of 40 h of validation work made available by the NRC. For each patient-friendly term, we randomly selected one subtype that only had that concept as supertype in its clarification (when available), and one subtype that also had other supertypes that were used in the clarification (when available). In other words, in this way, we could validate all concepts with patient-friendly terms or definitions, both as a single supertype in the clarification, as well as one of multiple supertypes (see Table 1). To get a representative set of combinations of supertypes, we sampled by unique combinations of supertypes, covering up to one-third of all unique combinations of supertypes. Additionally, we sampled from concepts with patient-friendly synonyms that had no patient-friendly definitions, to see whether clarifications with supertypes can be useful for these cases. Two raters evaluated the clinical validity and provided feedback from the perspective of how they would explain the diagnosis to a patient. The raters were first provided with instructions on carrying out the

Table 3

Outcome measures with a description and the scale used.

Outcome measure	Description	Scale
Error	Whether the clarification of the diagnosis contains any mistakes.	Boolean (0 or 1)
Completeness	The extent to which important information is present that you would use to provide a short clarification of the diagnosis.	Likert (1 – 5)
Relevance	The extent to which the information that is provided in the clarification is relevant to clarify a diagnosis to a patient.	Likert (1 – 5)
Clarity	The extent to which the clarification is clear, not vague, not confusing and not difficult to understand the diagnosis.	Likert (1 – 5)
Acceptability	The extent to which it is acceptable to show the clarification to a patient in the diagnosis list of his or her patient portal.	Likert (1 – 5)

validation and were provided with six example clarifications and how to validate them. They then validated 60 clarifications independently and discussed the results with the first author (HvM) to clarify the approach of the study and to achieve consensus. Where necessary, the validation questions were further refined and made explicit. After that, the raters continued with the rest of the clarifications independently. Both raters validated all of the clarifications. Finally, we discussed the clarifications in general. See Fig. 2 for the steps in the study flow, Appendix A for the sampling script, Appendix B for the sampling procedure, Appendix C for the rating guide including the validation questions, and Appendix D for the data dictionary.

3.3. Outcome measures and evaluation criteria

The primary outcome measures were the percentage of clarifications with identified errors and the percentage of clarifications that were considered unacceptable by the raters. Secondary outcomes were the completeness, relevance, and clarity of the clarifications and the interrater reliability of correctness and acceptability. See Table 3 for the description and scale of the outcome measures. The acceptability was based on the errors, completeness, relevance and clarity. A clarification should not contain errors (correctness), generalization should not leave out important information (completeness), the information in the clarification should be relevant (relevance) and it should be clear (clarity). The raters motivated their ratings in their own words in free text. Clarifications could be considered unacceptable without containing any mistakes, for example, if they provided irrelevant information. Conversely, mistakes could be considered acceptable, if they were considered minor issues to patients. Additionally, we categorized problems found and feedback for improvement.

3.4. Data entry forms

A case report form developed by SMe used for the mapping of code systems to SNOMED CT was extended and used to validate the diagnoses. It was developed using Python (Python Software Foundation, Delaware, DE, USA), Django (Django Software Foundation, Atlanta, GA, USA), and PostgreSQL (The PostgreSQL Global Development Group, Berkeley, CA, USA). The results were exported into a spreadsheet.

3.5. Methods for data analysis

3.5.1. Quantitative analysis

The intra-rater and interrater reliability were calculated using a two-way, agreement, single-measures intraclass correlation coefficient (ICC) [15]. We calculated the ICCs based on the Boolean and Likert rating scales used to score the outcome measures, to assess the agreement between the raters on those scales. The percentages of clarifications that

Table 4

Inclusion reasons and the number of included cases n with totals m per inclusion type, for clarifications consisting of a single supertype, those consisting out of multiple superotypes, and those with synonyms. If there was only one clarification with a particular supertype (only single) we took that one, if there were more than one, we sampled one of them. We also took extra samples to have a total of 1,188 cases.

Inclusion reason	n	Inclusion type	m
Only single	115	Total single	543
Sampled single	416		
Extra single	12		
Only multiple	108	Total multiple	522
Sampled multiple	260		
Sampled unique	144		
Extra multiple	10		
Sampled synonyms	123	Total synonyms	123
Total			1,188

were considered to contain errors and to be unacceptable were reported with their 95-percent confidence intervals (CI), similarly for their completeness, relevance, and clarity. We dichotomized the Likert scales taking a score of 1 or 2 as falsified (unacceptable, incomplete, irrelevant or unclear, respectively). We carried out a subgroup analysis by testing whether there were significant differences between clarifications that had single or multiple superotypes and compared clarifications with patient-friendly synonyms with those that did not have a patient-friendly synonym. We used the CIs and the Mann-Whitney test to determine whether the differences were significant. For this quantitative analysis, we excluded the first 60 cases, because they were still used to refine the validation instrument. Statistical data analysis was carried out in R (R Foundation for Statistical Computing, Vienna, Austria), using RStudio (RStudio Inc., Boston, MA, USA). The script can be found in Appendix E.

3.5.2. Qualitative analysis

Problems found regarding correctness, completeness, relevance, clarity, and acceptability, as well as suggestions for improvement were thematically analyzed, categorized and summarized narratively by the first author (HvM). A qualitative subgroup analysis of clarifications was carried out by analyzing patterns in the correctness, completeness, relevance, clarity, and acceptability. For this qualitative analysis, we included all cases (also the first 60). The qualitatively derived themes were based on the feedback provided during the validation. The themes were discussed with the raters to achieve consensus.

4. Results

4.1. Sampling

We generated clarifications with 2,690 unique combinations of 620 superotypes for the 16,124 diagnoses from the DT for which a SNOMED CT mapping was available. From these diagnoses, we randomly sampled 1,188 cases for validation (7.4% of all cases) with 1,103 unique combinations of superotypes (41.0% of all unique combinations). With this sample, we covered all unique 620 superotypes, both as single (531) and multiple (512) superotypes. See Table 4 for the inclusion reasons and types of cases from the random sample.

4.2. Ratings

4.2.1. Errors

The raters identified errors in 12.7% (CI: 10.7–14.6%) of the clarifications. The raters agreed on the correctness of 90.7% of the cases and the ICC of the interrater reliability was 0.34 for correctness. Most errors were related to the PFE synonyms (36.8%) and definitions (51.3%), but also some were related to the mappings between DT and SNOMED CT

Table 5

Ratings of dichotomous measures, as a percentage with 95% confidence interval (CI), and the interclass coefficient (ICC). * Correctness (whether a clarification contained errors) was only measured dichotomously. For the other measures, the dichotomous percentage was understood as a score of 1 or 2 (falsified).

Measure	ICC	Measure	Percentage %	CI
Correctness	0.34	Incorrect	12.7	10.7–14.6
Completeness	0.45	Incomplete	14.2	12.1–16.2
Relevance	0.39	Irrelevant	16.7	14.6–18.9
Clarity	0.45	Unclear	18.0	15.7–20.2
Acceptability	0.43	Unacceptable	14.3	12.3–16.4

(5.9%) and the modelling of SNOMED CT (16.5%), or were inherent to the algorithm (9.2%). Splitting into supertypes by the algorithm introduced errors with metastases and sometimes the ‘type of’-explanation was considered to lead to erroneous clarifications. The categorization of errors with examples is listed in Table 6 and Appendix F.

4.2.2. Acceptability

The raters found 14.3% (CI: 12.3–16.4%) of the clarifications to be unacceptable for a patient portal diagnosis list (defined as one or both raters providing a score of 1 or 2 for acceptability). The raters agreed on 48.3% of the Likert scale acceptability ratings with an ICC of 0.43 and 87.9% of the dichotomized acceptability. Table 5 shows the percentage of cases that were rated with a certain score. Fig. 3 shows the majority of the clarifications were considered complete, relevant, clear and acceptable. A categorization of issues with examples is listed in Table 7.

4.2.3. Subgroup analysis

Clarifications with a single supertype were rated significantly higher than those with multiple supertypes on correctness, relevance, and acceptability. For completeness, this difference was not significant. Clarifications with patient-friendly synonyms were rated significantly higher on the completeness, but there was no significant difference on the other measures, see Appendix G. However, in 59% of these cases, a synonym itself, without any supertypes, would have been sufficient in clarifying the diagnosis and providing synonyms or definitions of supertypes did not add any value. In the other 41%, the synonym itself was too difficult and needed further clarification.

Table 6

Error category, with the number and percentage of clarifications with errors that had an error of that category, an example case, a description of the error in the example, and a description of the error category explaining what the category means. One clarification can contain multiple errors.

Error category	n	%	Example case	Example error	Error category description
DT-SNOMED CT mapping	9	5.9	<u>‘Duchenne female carrier’</u> A type of birth defect and limb-girdle muscular dystrophy. It’s hereditary. ‘limb-girdle’ muscular dystrophy: This is a hereditary condition in which the muscles of the shoulders, upper arms, hips and thighs in particular do not function or function insufficiently.’	‘Duchenne female carrier’ from the DT was mapped on ‘240050008 Manifesting female carrier of X-linked muscular dystrophy ’ and thereby being clarified with several manifestations.	DT diagnoses that are not mapped correctly to SNOMED CT concepts
SNOMED CT modelling	25	16.3	<u>‘40425004 Postconcussion syndrome </u> A type of dementia’	‘40425004 Postconcussion syndrome ’ was modelled as being a ‘230282000 Post-traumatic dementia ’	Defining relationships that are modelled incorrectly in SNOMED CT
PFE synonym	56	37.3	<u>‘32644009 Braxton Hicks contractions </u> A type of childbirth’	Childbirth was the patient-friendly synonym of ‘118216002 Labor finding ’	PFE synonyms that are not correctly mapped to SNOMED CT concepts
PFE definition	78	51.6	<u>‘20751000146109 Recurrent nasopharyngeal carcinoma </u> A type of chronic illness and cancer in the throat. Chronic illness: Conditions with no prospect of full recovery and with a relatively long duration of illness, often requiring long-term care.’	A recurrent disease is a chronic disease in SNOMED CT. However, the definition of chronic disease ‘with no prospect of recovery’ does not apply to this recurrent disease.	PFE definitions that are not correctly mapped or not applicable to subtypes
Algorithm	14	9.2	<u>‘94360002 Secondary malignant neoplasm of kidney </u> A type of cancer in kidney and metastases’	It seems like kidney cancer with metastases, but it rather is a metastasis in the kidney.	Splitting up into supertypes and word order by the algorithm changes the meaning incorrectly.

4.2.4. General remarks

Raters suggested adding lexical clarifications, e.g., what malignant or benign means. Definitions should not use enumeration with examples, because they are only applicable to some of the subtypes and may cause unnecessary worry, e.g., the definition pancreatic disorder contains the example of pancreatic cancer. Additionally, raters noted some synonyms and definitions were too difficult to be used as a clarification.

5. Discussion

5.1. Principal findings

We validated diagnosis clarifications that were automatically generated by generalizing the diagnoses to patient-friendly synonyms

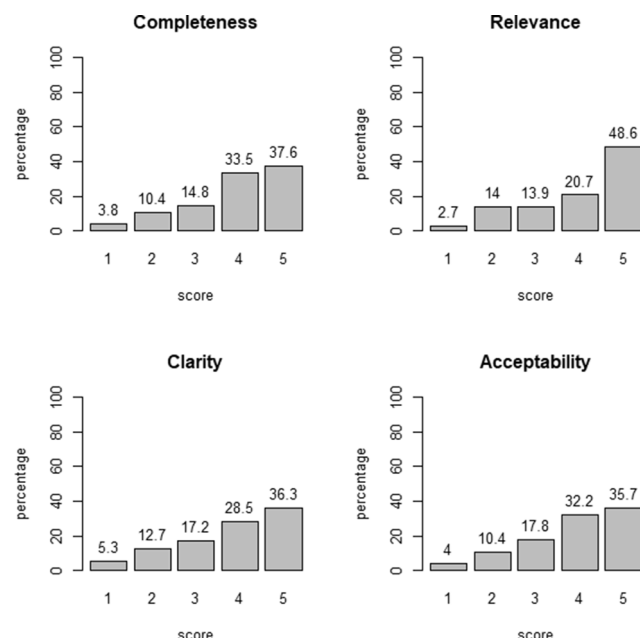


Fig. 3. The percentage of clarifications with scores for completeness, relevance, clarity, and acceptability. The strictest ratings given by the raters are shown.

Table 7

Examples of issues concerning acceptability with the number and percentage of clarifications with that issue that were considered unacceptable, an example case, and a description of the issue in the particular example and a description of the issue category explaining what the category means. One clarification can have multiple issues.

Issue	n	%	Example case	Example issue	Issue category description
Error	93	54.1	See Table 6 for examples of errors		One or both raters found an error. Note that a clarification may be considered unacceptable, even if it does not contain an error (e.g., the example below).
Incomplete	102	59.3	<u>'296495003 Methocarbamol overdose </u> A type of disease. Disease: This is a deviation of the healthy state of the body or mind, which manifests itself in symptoms and disorders.'	There is no clarification of the disease itself, but rather the term disease is defined, which is too general.	Ranked as incomplete by one or both raters.
Irrelevant	147	85.5	<u>'20781000146104 Recurrent supratentorial primitive neuroectodermal tumor </u> A type of chronic disease, glioma, brain tumor and cancer. Chronic disease: Disorders with no prospect of full recovery and with a relatively long duration of illness, which often means that care is required for a long time. Glioma: This is a malignant tumor that develops in the supporting cells (glia) of the brain.'	The explanation of this recurrent disease as a chronic disease is incorrect, not relevant and unacceptable.	Ranked as irrelevant by one or both raters.
Unclear	154	89.5	<u>'232357009 Post-surgical epistaxis </u> A type of bleeding, complication and disease of the heart and vessels and nose. Bleeding: Loss of blood, such as bruises, arterial bleeding, etc. Complication: Unintended and unwanted outcomes during or following the actions of a healthcare provider, which are so harmful to the health of the patient that adjustment of the medical treatment is necessary is, or that there is irreparable damage; the cause of a complication may be unavoidable and lie in the underlying disease, in comorbidity or characteristics of the patient, or be avoidable and lie in the (non) action of a care provider and/or in the care system.'	This case contains a very complex clarification for a nosebleed after surgery.	Ranked as unclear by one or both raters.
Algorithm	12	7.0	<u>'254486004 Benign neoplasm of nasal vestibule </u> A kind of benign tumor of airways and nasal disease.'	Airways is redundant when nose is mentioned already and now it might seem like multiple types of cancer	Splitting up into supertypes and word order by the algorithm changes the meaning unacceptably.
Examples in definition	7	4.1	<u>'262830005 Pancreatic duct injury </u> A pancreatic disease. Pancreatic disease: Diseases of the pancreas, such as pancreatic infection and pancreatic cancer.'	The example of pancreatic cancer is not applicable and may cause unnecessary worry.	Examples that might cause worry.
PFE definition missing	21	12.2	<u>'267611002 Degeneration of macula and posterior pole </u> A type of macular degeneration.'	Macular degeneration is a difficult synonym that needs further clarification.	The synonym is not patient-friendly and requires further clarification

and definitions of more general concepts in the PFE. We found the majority of clarifications were considered correct, acceptable, complete, relevant, and clear by both raters. However, we identified errors and unacceptable results in some clarifications that need to be addressed. Clarifications that were considered to be the most unacceptable were incorrect and seemed to cause worry. The interrater reliability of correctness (ICC: 0.342) was poor and of acceptability (ICC: 0.429) was fair [16], however, which shows that raters had different interpretations. Most problems that we found were associated with the patient-friendly synonyms and definitions from the PFE themselves, SNOMED CT modelling, and terminology mappings. Some identified problems related to the algorithm itself, creating unnecessarily complicated clarifications or changing the semantics by splitting the concept into its supertypes.

5.2. Results in relation to other studies

Zeng et al [13] used the Unified Medical Language System to provide generalizations for concepts and encountered similar problems: too general explanations that were not helpful 'nulligravida (a type of finding)' and some incorrect explanations. Zeng et al used a readability

score to assess whether a term needed to be translated, while we assumed that terms and definitions in the PFE were at the right difficulty level. Their system was used to clarify free-text documents and was evaluated with a sample of EHR notes. They didn't address the validity of the clarifications on the level of the terminology but rather assessed the performance on the level of the text. Research on text simplification [6,13,14,17,18], with a focus on free-text documents and records, has to address named entity recognition and medical concept normalization, to determine which concepts are represented in the terms that might contain ambiguity and spelling errors. We avoided this by starting from diagnosis data that are coded with terminology systems. This way, we could focus on the task of clarifying the specific medical concept, using knowledge already available in SNOMED CT ontology. However, we still encountered issues in the underlying terminologies our approach was based on.

Studies [19–21] on ontology verbalization have aimed to convert SNOMED CT ontology into text. This can be used to enable users to understand the formal SNOMED CT representations for terminology auditing and to access definitions of concepts. These studies have not yet provided information on a patient-friendly level of detail. Our approach can also be considered a type of ontology verbalization, by retrieving

knowledge represented in the SNOMED CT hierarchy and terminology mappings. Additionally to the challenges of verbalization, we found issues in the SNOMED CT hierarchy and the level of detail required to clarify concepts to laypersons.

5.3. Implications of the study and recommendations

SNOMED CT logically specifies disorders well that are fully defined, for example by morphology and finding site. However, some domains are underspecified, such as genetic and metabolic disorders, because SNOMED CT does not contain formal specifications of genes and proteins. Furthermore, concepts of syndromes and uncertainty are not fully defined in SNOMED CT. The algorithm splits concepts along the anatomic hierarchy, which can lead to unexpected results. These problems imply that the ontology verbalization approach will work better with fully defined concepts, rather than the generalization approach. We will investigate a combination of these approaches in further research. In our study, we focused on Dutch diagnoses and the PFE, but the algorithm can also be applied to other parts of SNOMED CT, and other languages and terminology systems. We expect that one would encounter similar issues.

Most problems we identified can be resolved by improvements in the algorithm, terms and definitions. They should be taken into account and improvements should be made accordingly where possible before implementing the clarifications in a patient portal. Most of the errors related to PFE synonyms and definitions arose because they did not apply to subtypes. We therefore recommend that, when terminologists add synonyms and definitions to SNOMED CT concepts in the PFE, the applicability to subtypes is taken into account to ensure the synonym or definition is also applicable to its subtypes. This implies excluding examples in definitions because those are not always applicable to subtypes. Additionally, the PFE should provide terms and definitions at a patient-friendly, low reading difficulty level; currently, the PFE still contains some difficult definitions that are not easy to read. A readability measure could also be considered, for example using frequency scores, to determine whether a term needs a definition. In focus groups with several Dutch terminology organizations[22], there was indeed disagreement on whether the terms and definitions in the PFE should be considered “patient-friendly”. Additionally, “patient-friendliness” itself was a controversial concept. In one of the focus groups it was agreed that terms and definitions should be written at B1 level (in the European Reference Framework, i.e., using more common, plain language terms, and more simple formulations) or should be tailored to the user. In future additions and modifications of terms and definitions in the PFE this will be taken into account. In our study, we found that the experts did not always identify the same errors. The PFE, and results from future similar studies, should therefore be validated by more than one evaluator, preferably with a background in the particular medical specialty in which the concept is used. Although this was not the initial objective of our study, the problems with terminology mappings and modelling we found show that this method can be used to evaluate the quality of the underlying terminologies.

5.4. Strengths and weaknesses

The selection of clarifications covers all patient-friendly synonyms and definitions from the PFE that are supertypes of diagnoses from the DT and a large part of the unique combinations of supertypes used in the diagnosis clarifications. We, therefore, consider it to be representative of other diagnoses. However, the diagnosis clarifications that were not

included due to the limited sample size still have to be validated. Additionally, clarifications should be revalidated when new terminology versions are released that change the relationships, terms and definitions used to generate the clarifications. We assumed the patient-friendly terms and definitions were included in SNOMED CT on the right level of detail and written in a patient-friendly form, but some terms and definitions were too difficult. Additionally, the definitions were not created to generalize more specific concepts to more general ones. We expect that results can be improved by taking this into account when improving or adding terms and definitions to the PFE.

The poor interrater reliability shows that experts identify different errors and differ in their opinion on whether patient-friendly clarifications are acceptable, which indicates that the validation instrument could be improved and emphasizes that more than one rater is needed to identify issues in the clarifications. One rater systematically identified more errors and provided lower ratings than the other rater. While the agreement between raters on correctness was high (91%), the ICC was affected by the skewed distribution of the correctness: there was a high a priori probability of agreement because 87% of clarifications were considered correct by both raters, which led to a higher penalty by the ICC in case of disagreement. The divergent interpretations seemed to be related to having a different medical specialization and clinical experience. For example, ‘steroid responder’ is an ophthalmologic disorder in the DT, but may seem not to be a diagnosis from the point of view of another medical specialty.

5.5. Future work

The clarifications have only been medically evaluated in this validation study. Ultimately, patients and their significant others are supposed to benefit from clarifications and further evaluation by these end-users is planned. However, a medical evaluation is a necessary first step to assess the validity of the patient-friendly terms, as patients are incapable of this assessment. The future validation instrument should be improved to increase the reliability and to specifically gauge the experience and understanding of the patients. While generalization increases the content coverage and reuse of terms and definitions, the dependence upon the quality of underlying terminology systems introduces uncertainty about the correctness and acceptability of the clarifications. In further research, we thus have to address how this approach should be used in practice: whether the automatic clarifications can be used in practice directly, with some disclaimer or warning, or whether it should only be used as a first draft clarification to develop patient-friendly terminology.

6. Conclusion

We found the majority of diagnosis clarifications to be considered correct, complete, relevant, clear and acceptable but also identified some problems in diagnosis generalization to patient-friendly synonyms and definitions. Most of the found problems can be resolved, as they were related to the underlying terminologies. This shows generalization using the SNOMED CT hierarchy can be a useful and adequate method to increase the content coverage of the PFE. Before implementing the clarifications in practice some modifications have to be made to the terms and definitions, and this needs to be updated when new terminology versions are released. Further research should focus on improving the quality of the underlying terminologies, the relevance and clarity of the clarifications, and on evaluation with end-users.

Consent to participate and for publication.

Written informed consent to participate and informed consent for publication was obtained from the authors SMA and EP who rated the clarifications.

Availability of data and materials.

Data are available from the corresponding author on reasonable request.

Funding.

This research did not receive any specific grant from funding

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jbi.2022.104071>.

Appendix G

Table 8.

Table 8

Comparison of outcomes between rater 1 and 2, between concepts with a single supertype or multiple superotypes, and between concepts with a patient-friendly synonym and those that did not have a patient-friendly synonym. Percentages with lower and upper value of the confidence interval, total number of cases, whether the CI was different and the p-value of the Mann-Whitney test, with column s containing an asterisk where p-value < 0.05.

from	measure	%	lower	upper	total	from	measure	%	lower	upper	total	p	s
Rater 1	Error	11.1	9.3	12.9	1125	Rater 2	Error	4.5	3.3	5.8	1123	0	*
	Not Complete	9.2	7.5	10.8	1125		Not Complete	8.5	6.8	10.1	1123	0.29	
	Not Relevant	14.0	12.0	16.1	1125		Not Relevant	6.3	4.9	7.7	1123	0	*
	Not Clear	12.1	10.2	14.0	1125		Not Clear	9.7	8.0	11.4	1123	0	*
	Not Accept	11.2	9.4	13.0	1125		Not Accept	5.3	3.9	6.6	1123	0	*
Singles	Error	7.4	5.1	9.7	502	Multiples	Error	18.9	15.4	22.4	486	0	*
	Not Complete	16.3	13.1	19.6	502		Not Complete	14.0	10.9	17.1	486	0.83	
	Not Relevant	13.3	10.4	16.3	502		Not Relevant	21.4	17.8	25.0	486	0	*
	Not Clear	16.9	13.7	20.2	502		Not Clear	21.2	17.6	24.8	486	0	*
	Not Accept	12.0	9.1	14.8	502		Not Accept	18.7	15.3	22.2	486	< 0.01	*
Friendly	Error	10.9	5.3	16.5	119	Unfriendly	Error	12.9	10.8	14.9	1010	0.55	
	Not Complete	3.4	0.1	6.6	119		Not Complete	15.4	13.2	17.7	1010	< 0.01	*
	Not Relevant	12.6	6.6	18.6	119		Not Relevant	17.2	14.9	19.6	1010	0.87	
	Not Clear	10.1	4.7	15.5	119		Not Clear	18.9	16.5	21.3	1010	0.05	
	Not Accept	7.6	2.8	12.3	119		Not Accept	15.1	12.9	17.4	1010	0.15	

agencies in the public, commercial, or not-for-profit sectors.

Author's contributions.

HvM drafted and organized the study with input from RC, NdK, and RN. SMe developed the case report forms. SMA and EP carried out the validation. HvM carried out the analyses and wrote the paper with input from RC, NdK, RN, SMe, SMA, and EP. All authors approved the manuscript and the submission of the manuscript.

CRedit authorship contribution statement

Hugo J.T. van Mens: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization, Project administration. **Savine S. M. Martens:** Investigation, Writing – review & editing. **Elisabeth H.M. Paiman:** Investigation, Writing – review & editing. **Alexander C. Mertens:** Software, Validation, Resources, Writing – review & editing. **Remko Nienhuis:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing. **Nicolette F. de Keizer:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Ronald Cornet:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Thanks to Pim Volkert, Annemarieke Homan, Feikje Hielkema, Natasha Krul, Eef Verbeek, Daniël Huliselan, Theo van Mens and Frank van Oosterhout, for their contributions to this study.

References

- [1] T. Irizarry, A. DeVito Dabbs, C.R. Curran, Patient portals and patient engagement: a state of the science review, *J. Med. Internet Res.* 17 (6) (2015) e148, <https://doi.org/10.2196/jmir.4255>.
- [2] A. Keselman, C.A. Smith, A classification of errors in lay comprehension of medical documents, *J. Biomed. Inform.* 45 (6) (2012) 1151–1163.
- [3] M. Wernick, P. Hale, N. Anticich, S. Busch, L. Merriman, B. King, T. Pegg, A randomised crossover trial of minimising medical terminology in secondary care correspondence in patients with chronic health conditions: impact on understanding and patient reported outcomes, *Intern. Med. J.* 46 (5) (2016) 596–601.
- [4] J.P. Lalor, B. Woolf, H. Yu, Improving Electronic Health Record Note Comprehension With NoteAid: Randomized Trial of Electronic Health Record Note Comprehension Interventions With Crowdsourced Workers, *J. Med. Internet Res.* 21 (1) (2019) e10793, <https://doi.org/10.2196/10793>.
- [5] Y.K. Choi, K. Kirchhoff, A.M. Turner, editors. Medical Text Simplification by Medical Trainees: A Feasibility Study, in: 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016, IEEE.
- [6] S. Kandula, D. Curtis, Q. Zeng-Treitler, A Semantic and Syntactic Text Simplification Tool for Health Content, *AMIA Annu. Symp. Proc.* 2010 (2010) 366–370.
- [7] H.J.T. van Mens, N.F. de Keizer, R. Nienhuis, R. Cornet, Clarifying Diagnoses to Laymen by Employing the SNOMED CT Hierarchy, *Stud. Health Technol. Inform.* 247 (2018) 900–904.
- [8] Q.T. Zeng, T. Tse, G. Divita, A. Keselman, J. Crowell, A.C. Browne, S. Goryachev, L. Ngo, Term Identification Methods for Consumer Health Vocabulary Development, *J. Med. Internet Res.* 9 (1) (2007) e4, <https://doi.org/10.2196/jmir.9.1.e4>.
- [9] A. Keselman, T. Tse, J. Crowell, A. Browne, L. Ngo, Q. Zeng, Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study, *J. Med. Internet Res.* 9 (1) (2007) e5, <https://doi.org/10.2196/jmir.9.1.e5>.
- [10] J. Chen, A.N. Jagannatha, S.J. Fodeh, H. Yu, Ranking Medical Terms to Support Expansion of Lay Language Resources for Patient Comprehension of Electronic Health Record Notes: Adapted Distant Supervision Approach, *JMIR Med. Inform.* 5 (4) (2017) e42, <https://doi.org/10.2196/medinform.8531>.
- [11] Z. He, Z. Chen, S. Oh, J. Hou, J. Bian, Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach, *J. Biomed. Inform.* 69 (2017) 75–85.
- [12] M. Ibrahim, S. Gauch, O. Salman, M. Alqahtani, An automated method to enrich consumer health vocabularies using GloVe word embeddings and an auxiliary lexical resource, *PeerJ Comput. Sci.* 7 (2021) e668.
- [13] Q. Zeng-Treitler, S. Goryachev, H. Kim, A. Keselman, D. Rosendale, Making texts in electronic health records comprehensible to consumers: a prototype translator, in: *Amia 2007/Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 846–850.

- [14] J. Chen, E. Druhl, B. Polepalli Ramesh, T.K. Houston, C.A. Brandt, D.M. Zulman, V. G. Vimalananda, S. Malkani, H. Yu, A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notes to Lay Definitions: System Development Using Physician Reviews, *J. Med. Internet Res.* 20 (1) (2018) e26, <https://doi.org/10.2196/jmir.8669>.
- [15] K.A. Hallgren, Computing inter-rater reliability for observational data: an overview and tutorial, *Tutor. Quant. Methods Psychol.* 8 (1) (2012) 23–34.
- [16] D.V. Cicchetti, Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology, *Psychol. Assess.* 6 (4) (1994) 284–290.
- [17] N. Kloehn, G. Leroy, D. Kauchak, Y. Gu, S. Colina, N.P. Yuan, D. Revere, Improving Consumer Understanding of Medical Text: Development and Validation of a New SubSimplify Algorithm to Automatically Generate Term Explanations in English and Spanish, *J. Med. Internet Res.* 20 (8) (2018) e10779, <https://doi.org/10.2196/10779>.
- [18] L. Ramadier, M. Lafourcade, editors. Radiological Text Simplification Using a General Knowledge Base, 2018; Cham: Springer International Publishing.
- [19] S.F. Liang, R. Stevens, D. Scott, A. Rector, editors. Automatic verbalisation of SNOMED classes using OntoVerbal, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2011.
- [20] M. Kanhov, X. Feng, H. Dalianis, editors. Natural Language Generation from SNOMED Specifications. CLEF (Online Working Notes/Labs/Workshop), 2012, Citeseer.
- [21] O. Lyudovyyk, C. Weng, SNOMEDtxt: Natural Language Generation from SNOMED Ontology, *Stud Health Technol. Inform.* 264 (2019) 1263–1267.
- [22] Nictiz. Taal en techniek ontmoeten elkaar bij brainstorm over patiëntvriendelijke termen (in Dutch) 2019 [cited 2022 January, 24]. Available from: <http://purl.org/hjtvanmens/generalization/nictizblog>.