Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review

Gaudet-Blavignac, Christophe; Foufi, Vasiliki; Bjelogrlic, Mina; Lovis, Christian

How to cite

GAUDET-BLAVIGNAC, Christophe et al. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review. In: Journal of Medical Internet Research, 2021, vol. 23, n° 1, p. e24594. doi: 10.2196/24594

This publication URL: https://archive-ouverte.unige.ch/unige:152330
Publication DOI: 10.2196/24594

Review

Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review

Christophe Gaudet-Blavignac[1,2], BSc, MSc; Vasiliki Foufi[1,2], PhD; Mina Bjelogrlic[1,2], PhD; Christian Lovis[1,2], MPH, MD, FACMI

[1]Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland

[2]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

Corresponding Author:
Christophe Gaudet-Blavignac, BSc, MSc
Division of Medical Information Sciences
Geneva University Hospitals
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 22 372 62 01
Email: christophe.gaudet-blavignac@hcuge.ch

Abstract

Background: Interoperability and secondary use of data is a challenge in health care. Specifically, the reuse of clinical free text remains an unresolved problem. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) has become the universal language of health care and presents characteristics of a natural language. Its use to represent clinical free text could constitute a solution to improve interoperability.

Objective: Although the use of SNOMED and SNOMED CT has already been reviewed, its specific use in processing and representing unstructured data such as clinical free text has not. This review aims to better understand SNOMED CT's use for representing free text in medicine.

Methods: A scoping review was performed on the topic by searching MEDLINE, Embase, and Web of Science for publications featuring free-text processing and SNOMED CT. A recursive reference review was conducted to broaden the scope of research. The review covered the type of processed data, the targeted language, the goal of the terminology binding, the method used and, when appropriate, the specific software used.

Results: In total, 76 publications were selected for an extensive study. The language targeted by publications was 91% (n=69) English. The most frequent types of documents for which the terminology was used are complementary exam reports (n=18, 24%) and narrative notes (n=16, 21%). Mapping to SNOMED CT was the final goal of the research in 21% (n=16) of publications and a part of the final goal in 33% (n=25). The main objectives of mapping are information extraction (n=44, 39%), feature in a classification task (n=26, 23%), and data normalization (n=23, 20%). The method used was rule-based in 70% (n=53) of publications, hybrid in 11% (n=8), and machine learning in 5% (n=4). In total, 12 different software packages were used to map text to SNOMED CT concepts, the most frequent being Medtex, Mayo Clinic Vocabulary Server, and Medical Text Extraction Reasoning and Mapping System. Full terminology was used in 64% (n=49) of publications, whereas only a subset was used in 30% (n=23) of publications. Postcoordination was proposed in 17% (n=13) of publications, and only 5% (n=4) of publications specifically mentioned the use of the compositional grammar.

Conclusions: SNOMED CT has been largely used to represent free-text data, most frequently with rule-based approaches, in English. However, currently, there is no easy solution for mapping free text to this terminology and to perform automatic postcoordination. Most solutions conceive SNOMED CT as a simple terminology rather than as a compositional bag of ontologies. Since 2012, the number of publications on this subject per year has decreased. However, the need for formal semantic representation of free text in health care is high, and automatic encoding into a compositional ontology could be a solution.

(J Med Internet Res 2021;23(1):e24594) doi: 10.2196/24594

KEYWORDS

SNOMED CT; natural language processing; scoping review; terminology

Introduction

Background

The ability to meaningfully exchange and process data is of utmost importance in health care, whether it is inside a hospital setting either among different health structures or among health systems in different countries [1-3]. The use of a common terminology is a way to improve both interoperability and the secondary use of data [4].

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) was created in 1999 by the fusion of 2 important health care terminologies—SNOMED reference terminology (SNOMED RT) and Clinical Terms Version 3. It was first released in 2002. SNOMED CT is currently considered as the most comprehensive, multilingual, clinical health care terminology in the world, with more than 350,000 concepts and a million relationships [5-7]. It is maintained and published by SNOMED International, a nonprofit organization comprising 39 member countries [8]. In the last 18 years, SNOMED CT has grown in size and coverage and has been included as a standard vocabulary in the meaningful use program [9]. This is an important step for any electronic health record willing to attain interoperability.

With 3 components, namely concepts, descriptions, and relationships, SNOMED CT can be observed as both a complex ontology and a graph containing vertices and labeled edges. This structure allows interesting features such as compositional grammar, expression constraint queries, or postcoordination. It is therefore possible to create postcoordinated concepts that represent new meanings not present in the terminology. These postcoordinated concepts can then be queried and processed with the rest of the terminology [5,10,11].

These characteristics, similar to those of a natural language, make SNOMED CT a candidate for representing clinical free text in a semantically rich, machine-readable manner. Although encoding free text into SNOMED CT can be done manually, it is costly and not scalable for large data sets. Therefore, it is often accomplished by natural language processing (NLP). NLP is an active research branch in the biomedical field and has been broadly applied in the scientific literature and clinical text for diverse tasks [12-14]. However, NLP applications on clinical documents are less frequent. Among the reasons explaining this disparity are the limited access to corpora of clinical documents and the lack of publicly available annotated corpora [15]. These barriers are even more important for languages other than English.

Objectives

SNOMED CT has already been the subject of many studies and evaluations of its coverage, ability to represent complex concepts, or usability in a clinical setting [16-19]. Its usage has already been a subject of reviews; however, those publications are older than 10 years [13,20] or focus on its general use without focusing on its usage to process and represent unstructured data such as clinical free text [7]. Therefore, this work aims to better understand the use of SNOMED CT for representing free text in medicine via a scoping systematic review. It also aims to decipher the use of this terminology across fields, languages, and countries and how it is used from an analytical point of view, such as terminology source up to exploiting its advanced features, that is, postcoordination and compositional grammar.

Methods

Article Selection Process

An exploratory research performed using text-based queries on MEDLINE and Google Scholar helped in defining the queries, topics, and objectives of this study. This work led to the selection of 3 databases for the review based on previous reviews addressing similar topics [7,20,21]. This choice was made to increase coverage. Purely engineering-related databases, such as the Institute of Electrical and Electronics Engineers Xplore or the Association for Computing Machinery digital library, were not selected because of the technical content of their publications, which was often not related to real clinical settings.

In this work, clinical free text is considered as any text written in a natural language about a patient, which does not come from a finite value set. Free-text fields in structured forms and problem lists have been included to broaden the scope.

The selected databases were PubMed [22], Embase [23], and Web of Science [24]. The final query used was as follows: (“SNOMED-CT” OR “SNOMED CT”) AND (“free-text” OR “free text” OR “narrative”). These keywords were defined during the preliminary research. The bottleneck was the presence of the term “SNOMED CT,” and no other synonyms of narrative or free text were added as they did not change the results. The final query was made on August 9, 2019.

To be selected, an article must meet the following inclusion criteria:

- It should be published in scientific journals or conference proceedings after 2002.
- It should include the usage of SNOMED CT to represent or process clinical free text.

The limitation on the date was set to avoid publications that focused on the previous versions of SNOMED.

Although the selection was voluntarily broad, white papers, editor papers, posters, or conference abstracts were excluded. Articles not available in English were also excluded. The Unified Medical Language System (UMLS) [25] developed by the National Library of Medicine (NLM) combines biomedical terminologies in a single resource. Since the release of the UMLS-labeled 2004AA [26,27], it contains SNOMED CT. In this work, publications focusing on the usage of UMLS were included only if they specifically mentioned the usage of SNOMED CT.

To be as inclusive as possible on the chosen topic, the references in every publication were also reviewed to include new publications. The recursive reference review was stopped when no additional publications were added to the set. This has been done with the aim of reducing the impact of the query on the final selection of articles. Moreover, 3 review articles about information extraction from clinical free text were included in the selection. Despite not meeting the inclusion criteria, they were considered as a source of reference to other publications meeting the criteria. Obviously, they were not the target of the topic review described below.

Topics Reviewed

The articles were then studied to extract some specific topics in a systematic manner. The first topic reviewed was the type of document used as a free-text source. To better detect which data were used in these publications, we defined the categories described in Textbox 1.

Textbox 1. Categories of documents.

- History and physical examinations: this category includes documents summarizing the situation of a patient admitted in a health care structure, and his or her physical examination such as admission notes
- Clinical summaries: this category includes any document summarizing a care episode such as a discharge summary
- Death certificates
- Problem lists: this category regroups documents listing the problems of a patient admitted in a health care structure
- Autopsy reports
- Incident reports
- Allergy reports
- Complementary exam reports: this category regroups any document related to a complementary exam, including but not limited to radiology, pathology, and genomic reports
- Narrative notes: this category includes progress notes, nurse notes, and clinical notes not further specified
- Various: this category was selected when a publication used more than one type of document according to this classification

The publications were then classified according to the language they targeted in their work. All the selected publications included a part where the free text was mapped to SNOMED CT concepts. This terminology binding step was classified depending on its justification and whether it was the final goal of the research or a step toward another goal. Textbox 2 defines the types of reasons. These reasons have been defined empirically to fully cover the possibilities encountered in publications. For each type, a point was added if it was present in the publication.

The method used for the terminology binding to SNOMED CT was classified as “manual,” “rule-based,” “machine learning,” or “hybrid” for each article. The definitions used for these categories are listed in Textbox 3. When mapping was accomplished using a specific software, it was reviewed.

The general usage of SNOMED CT was reviewed on 2 specific topics: whether the full terminology or a subset of concepts was used and whether more advanced features of SNOMED CT were included in the study.

Textbox 2. Categories classifying the reason for the terminology binding to Systematized Nomenclature of Medicine Clinical Terms.

- Information extraction: Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is used to extract meaningful information from free text. The focus must be aimed at extracting information, not structuring or encoding it. Publications using the terminology binding to extract clinical information from documents that fall under this category
- Data normalization: SNOMED CT is used to encode existing data. This category is different from information extraction because it focuses on adding semantics to the data while keeping it intact. It includes publications where SNOMED CT is used to define a template or to support information entry
- Synonym resource: SNOMED CT includes synonyms for a large number of its concepts. In this category, SNOMED CT is used as a source for synonyms
- Quality evaluation: SNOMED CT is used to evaluate the quality of care or documentation
- Coverage evaluation: The focus is aimed at evaluating the coverage of SNOMED CT for a specific task by mapping it to free text
- Similarity evaluation: SNOMED CT is used to evaluate similarity among data. It is usually made by using the relationships present in SNOMED CT to compute the semantic distance between concepts
- Gold standard creation: SNOMED CT is used to create a gold standard data set
- Feature in a classification task: SNOMED CT mapping is used as a feature in a classification task
- Value set creation: SNOMED CT is used to define a specific value set
- Mapping to other terminologies: SNOMED CT is used as a bridge to other terminologies

Textbox 3. Definition of the categories used to classify the mapping method.

- Manual: the mapping is made by manually reading the text and assigning the correct concept [28,29]
- Rule-based: the mapping is made using rule-based methods such as text search, regular expressions, finite state machines, or a tool that is defined as rule-based [30,31]
- Machine learning: the mapping is made using probabilistic algorithms based on a learning mechanism such as support vector machine, conditional random fields [32], or naïve Bayes [33]
- Hybrid: the mapping is made using both rule-based and machine learning methods, whether it is simultaneously combined or sequentially [34]

Results

Article Selection

After 3 rounds of recursive reference review, the final selection included 76 publications and 3 reviews. Complete list of the publications is provided in Multimedia Appendix 1 [14,16,28-101]. Those reviews [13,102,103] will be excluded from the rest of the analysis, as they were only studied to broaden the scope of this review. The flow diagram according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [104] is shown in Figure 1.

Among the 76 selected articles, 42 (55%) publications were journal articles and 34 (45%) were conference proceedings. The number of publications published per year is shown in Figure 2. The 76 publications were issued from 37 journals and conference proceedings, with 10 journals or proceedings appearing in more than one publication in the selection (Table 1).

Overall, 238 unique authors were credited in the selection. More prolific authors (more than one authorship) are displayed in Figure 3.

Figure 1. Flow diagram of the selection process. SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

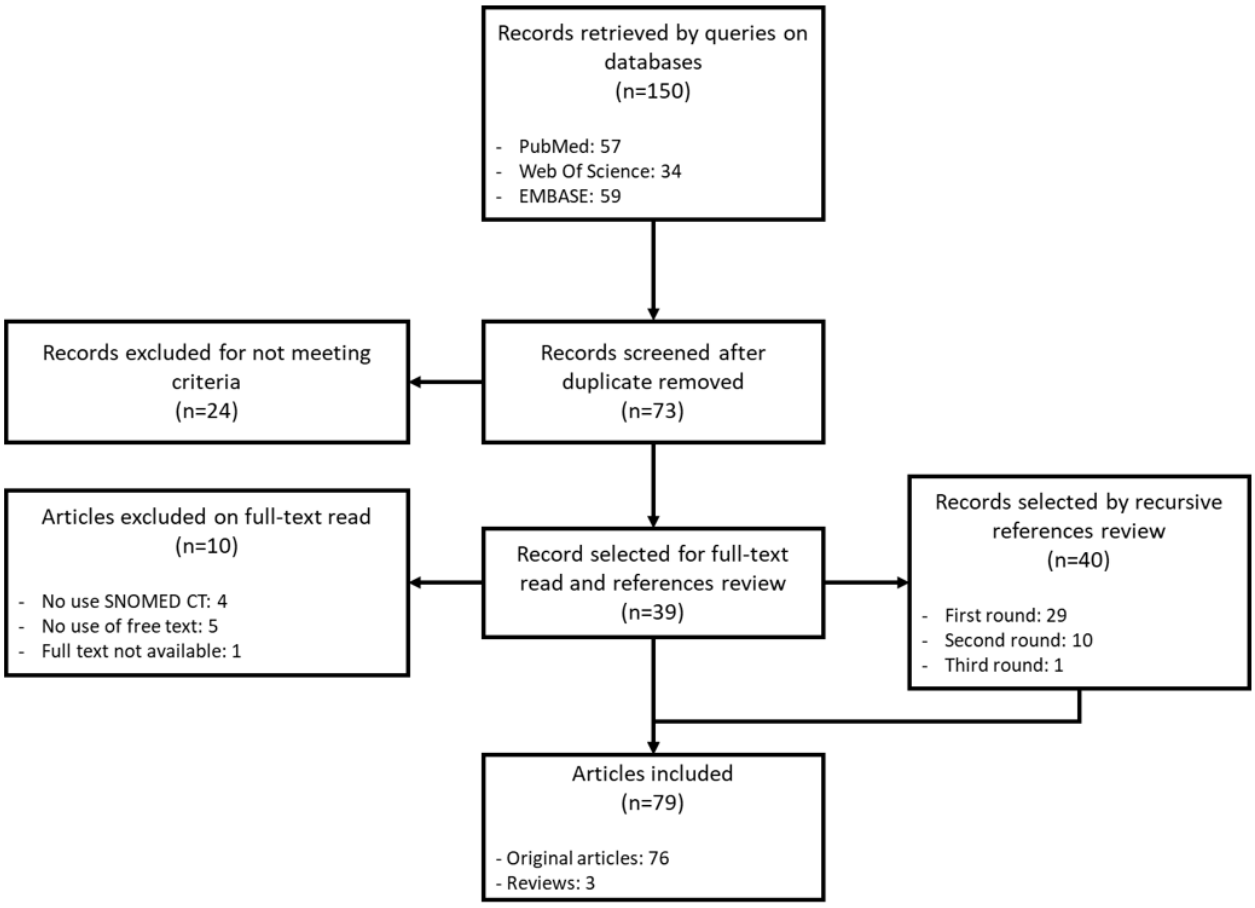


Figure 2. Number of publications per year of publication.

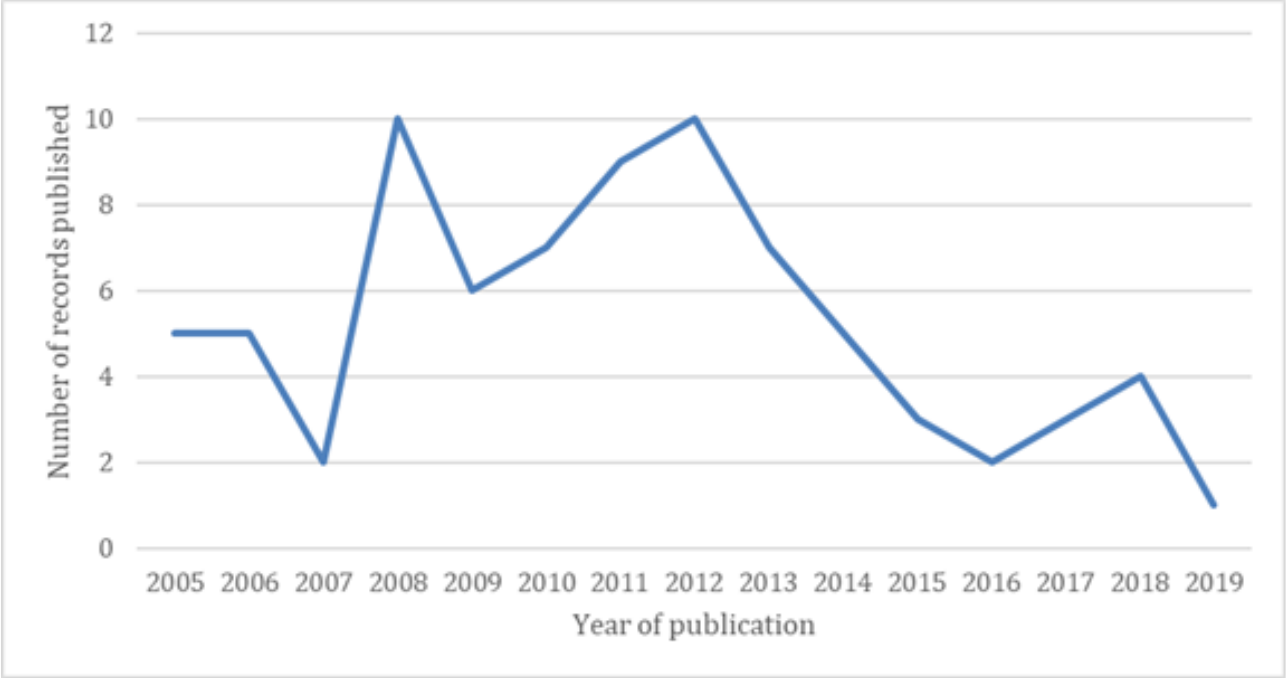


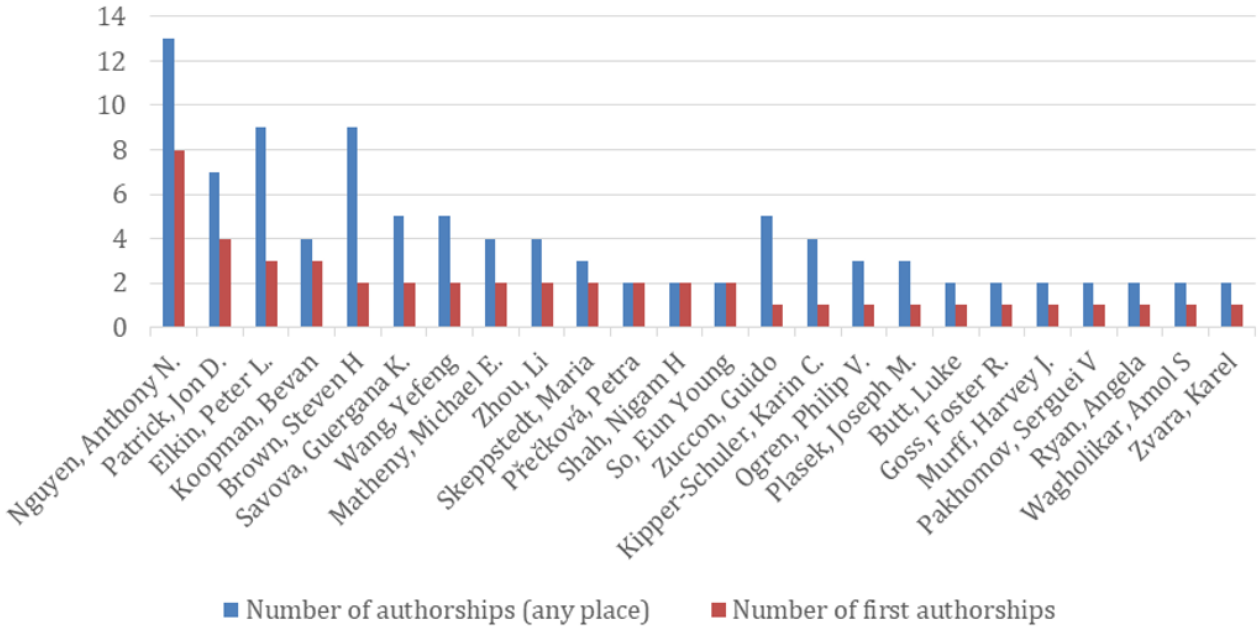
Table 1. Journals and conferences having more than 1 article in the selection.

Name of journal or conference	Publications, n
AMIA[a] Annual Symposium proceedings	15
Journal of Biomedical Informatics	8
BMC[b] Medical Informatics and Decision Making	7
Journal of the American Medical Informatics Association	7
Studies in Health Technology and Informatics	3
Journal of Digital Imaging	2
AMIA Joint Summits on Translational Science proceedings	2
Mayo Clinic Proceedings	2
Electronic Journal of Health Informatics	2
International Journal of Medical Informatics	2

[a]AMIA: American Medical Informatics Association.

[b]BMC: BioMed Central.

Figure 3. Number of authorships for the most prolific authors in selection.



Type of Data

The types of documents used in each publication are summarized in Table 2. The most frequent types are complementary exam reports (18/76, 24%), followed by narrative notes (16/76, 21%) and publications using more than one type of document (14/76, 18%).

Table 2. Number of publications per type of document used for the mapping.

Document Type	Publications (N=76), n (%)
Complementary exam report	18 (24)
Narrative note	16 (21)
Various	14 (18)
History and physical examination	8 (11)
Clinical summary	6 (8)
Death certificate	5 (7)
Problem list	3 (4)
Not available	3 (4)
Incident report	1 (1)
Autopsy report	1 (1)
Allergy report	1 (1)

Language

The target languages in the publications are listed in Table 3. Most papers focused on English (69/76, 91%). The 3 other languages were Swedish, Czech, and Chinese (Table 3).

Table 3. Target language in publications.

Language	Publications (N=76), n (%)
English	69 (91)
Swedish	3 (4)
Czech	3 (4)
Chinese	1 (1)

Reason for the Terminology Binding to SNOMED CT

As the focus of this work is to depict how the research community uses SNOMED CT to process clinical free text, selected articles had to include a part in which free-text data were mapped to SNOMED CT concepts. However, the mapping part was only a step toward another goal in many cases (eg, classification task [35,36], similarity measures [29,37], etc; Table 4).

Table 4. Role of the Systematized Nomenclature of Medicine Clinical Terms mapping in the publications.

Role of the SNOMED CT[a] mapping	Publications (N=76), n (%)
Final goal	16 (21)
Part of final goal	25 (33)
Step toward other goal	35 (46)

[a]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

The reasons for the SNOMED CT mapping in publications are displayed in Table 5. The most frequent reason is information extraction (44/76, 39%), followed by feature in a classification task (26/76, 23%) and data normalization (23/76, 20%). The remaining categories appear in 5 publications or less.

Table 5. Reason for the mapping in publications.

Reason for the SNOMED CT[a] mapping	Publications, n (%)
Information extraction	44 (39)
Feature in a classification task	26 (23)
Data normalization	23 (20)
Coverage evaluation	5 (4)
Similarity evaluation	4 (4)
Quality evaluation	3 (3)
Value set creation	3 (3)
Synonym resource	2 (2)
Terminology mapping	2 (2)
Gold standard creation	1 (1)
Total number of points given	113 (100)

[a]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

Mapping Method

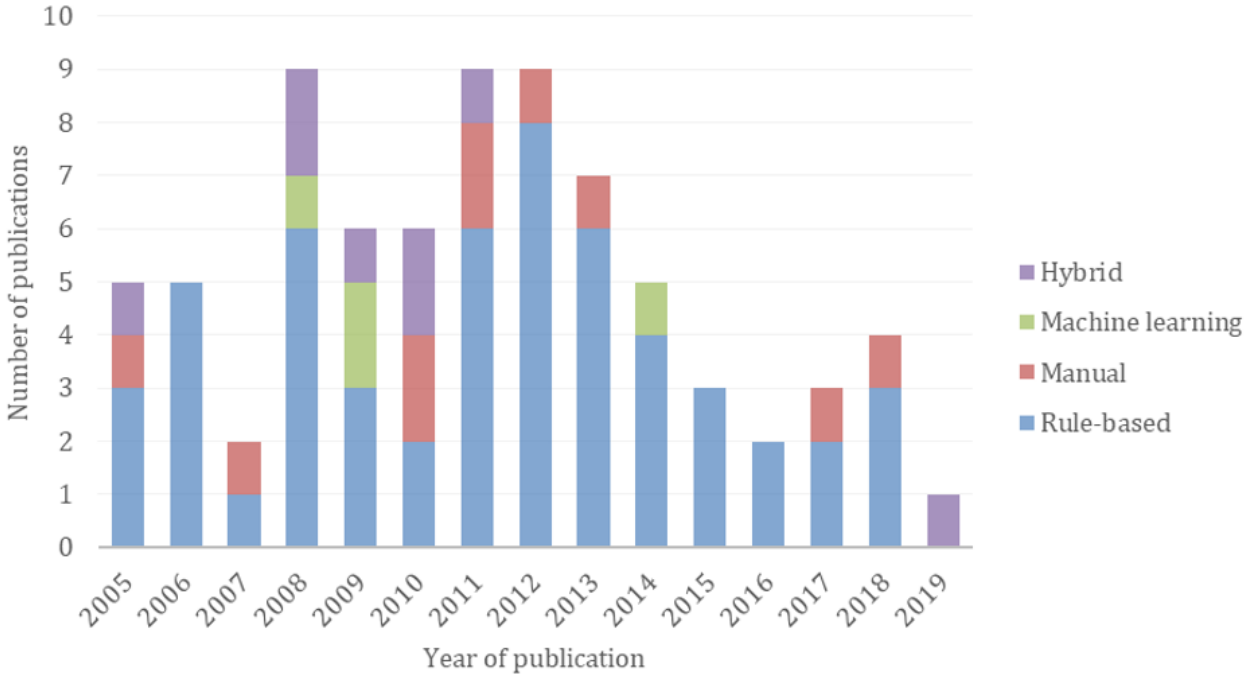
The type of method used for mapping according to the previously defined classification is presented in Table 6, and the methods used per year is displayed in Figure 4. The evolution of the methods shows that articles presenting machine learning approaches were published only in 2008, 2009, and 2014. Hybrid approaches are present during the period 2005 to 2010 and in 2019. With 70% (53/76) of publications, rule-based approaches were the most common method used to perform this task, although the number of publications per year is reducing overall.

Table 6. Method used for mapping free-text data to Systematized Nomenclature of Medicine Clinical Terms.

Method for SNOMED CT[a] mapping	Publications (N=76), n (%)
Rule-based	53 (70)
Manual	11 (14)
Hybrid	8 (11)
Machine learning	4 (5)

[a]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

Figure 4. Number of articles applying a specific method for Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) mapping when available.



Software Used for Terminology Binding

Table 7 shows the software used to specifically map free-text data to SNOMED CT concepts, the number of publications in which they appear, and whether they are publicly available. Only software used to produce a mapping into SNOMED CT are considered. Software used only for a step of the NLP pipeline such as negation detection or tokenization and not resulting in a concept-mapping output are not listed.

Table 7. Tools used for mapping free text to Systematized Nomenclature of Medicine Clinical Terms concepts.

Name of tool	Publications, n	Availability for public use
Medtex	12	No
MCVS[a]	8	No
MTERMS[b]	4	No
MetaMap	3	Yes
MetaMap transfer	3	Yes
Open biomedical annotator	2	Yes
MedLEE[c]	2	No
cTAKES[d]	2	Yes
Lingoengine	1	Yes
Snapper	1	No
iSCOUT	1	No
RapTAT[e]	1	No

[a]MCVS: Mayo Clinic Vocabulary Server.

[b]MTERMS: Medical Text Extraction Reasoning and Mapping System.

[c]MedLEE: Medical Language Extraction and Encoding System.

[d]cTAKES: clinical Text Analysis and Knowledge Extraction System.

[e]RapTAT: Rapid Text Annotation Tool.

Although all the software aim to detect concepts in free text, the wide disparities in methods and evaluation metrics, the

subsets of concepts used, and the output terminologies prevent strict comparison. Therefore, the following review focuses only on the systems themselves and their published evaluation.

Medtex [38]**,** developed by the Australian eHealth research center, is built based on other existing tools (GATE [105], metamap transfer MMTx [106], and NegEx [107]) and can annotate free text with SNOMED CT concepts and negation marks. Although it is used in 12 publications, to the best of our knowledge, no strict evaluation of the mapping has been published.

The *Mayo Clinic Vocabulary Server* (MCVS) [16], also called Multi-threaded Clinical Vocabulary Server [39], is able to map free text to codes in various classifications, among which, SNOMED CT codes. It is the subject of an evaluation of over 4996 problem statements, which resulted in a sensitivity of 99.7% and a specificity of 97.9%. It is linked to *LingoEngine* [40], which is described as a commercially available product linked to MCVS.

The *Medical Text Extraction Reasoning and Mapping System (MTERMS)* [41] is a system that uses shallow and deep parsers to extract and structure information from free text by using local and standard terminologies. The system also proposes mappings between the terminologies. It has been used to extract medication information, allergens, allergic reactions, [42,43] and family relatives [44]. Each of these uses required specific customization, such as adding ad hoc dictionaries. Evaluations proposed in publications about MTERMS cover the encoding of information in multiple terminologies and are restricted to a specific subject. The evaluation of allergy data shows a precision of 84.4%, a recall of 91.0%, and an F-measure of 87.6%. Moreover, the evaluation of family relatives showed a precision of 100%, a recall of 97.4%, and an F-measure of 98.7% over 291 occurrences.

MetaMap [106]**,** and its Java implementation *MMTx*, was developed by the NLM. Its goal is to map the biomedical text to the UMLS Metathesaurus [108]. Since 2004, the UMLS Metathesaurus contains SNOMED CT. Although MetaMap only maps free text to the UMLS concept unique identifier(CUI), the link between a CUI and a SNOMED CT concept is present in the Metathesaurus and it is possible to specify vocabulary sources used for mapping. Therefore, in this work, MetaMap is considered as a tool that can map free text to SNOMED CT concepts. A realistic evaluation of the performance of this software has never been performed [109]. However, specific task evaluations and comparisons with other software have been published [110-112]. They showed a performance of 88% in recall, 89% in precision, and 88% in F-score on clinical notes; a precision of 85% and a recall of 78% on concepts extracted from medical curriculum documents [110]; and finally, a precision between 33% and 76% on multiple web-based biomedical resources for the mapping of biological processes, depending on the data sources [111]. However, no specific evaluation of the SNOMED CT mapping has been published.

The *Open Biomedical Annotator* (*OBA*) [113] is an ontology-based web service that can annotate free text with a variety of ontologies. It uses and improves the annotations of a concept recognizer called Mgrep [114] and is developed by the National Center for Integrative Biomedical Informatics at the University of Michigan. Publications using OBA in the selection did not propose an evaluation of the SNOMED CT mapping. However, a comparison of Mgrep with Metamap showed a precision between 58% and 93% for biological processes depending on the data source [111]. However, these evaluations are not focused on SNOMED CT.

The *Medical Language Extraction and Encoding System* [115] developed in Columbia University aims to transform clinical data into controlled vocabularies. It has been specifically adapted for UMLS and evaluated on 300 random sentences with a precision of 89% and a recall of 83% [116]. However, this evaluation does not mention SNOMED CT or the UMLS version used.

The *clinical Text Analysis and Knowledge Extraction System* (*cTAKES*) [45], developed in the Mayo Clinic, is an open-source NLP software aimed at information extraction. It includes a dictionary lookup component able to map the free-text data to UMLS concepts. The named entity recognition component has been evaluated on a corpus of 160 notes manually annotated with UMLS concepts including SNOMED CT, and shows an F-score of 71.5% for exact and 82.4% for overlapping spans [46].

Snapper [117] by the Australian eHealth research center is a software with the ability to input free-text data and perform the mapping from a terminology to SNOMED CT. To the best of our knowledge, no strict evaluation of the software has been performed. *Snapper* has been used in the selection to classify narratives into symptom groups [47].

ISCOUT appears in only one publication in the selection. This software, developed at the Brigham and Women's Hospital in Boston, is used internally for document retrieval according to a list of terms from a terminology [48]. In the publication, it is used with a list of concepts from various terminologies, including SNOMED CT, to retrieve documents. However, no evaluation of concept detection is proposed.

The *Rapid Text Annotation Tool* (*RapTAT*) [33] is a token order–specific naïve Bayes–based machine learning system designed to predict an association between phrases and concepts. It has been evaluated on the manually annotated 2010 i2b2 shared task data [118] and compared with the MCVS output, defined as the gold standard on 2860 discharge summaries. On the manual data set, *RapTAT* reached a precision of 95%, a recall of 96%, and an F-measure of 95%. To reproduce the MCVS output, *RapTAT* achieved a precision of 92%, a recall of 85%, and an F-measure of 89%.

Among all software, 5 are available, either as a web-based interface or as an installer for public usage. For example, Metamap, MMTx, and cTAKES are open source, OBA is available as a web-based interface, and LingoEngine is commercially available.

Subset Usage

As SNOMED CT includes more than 340,000 concepts, the research studies described in publications often restrict their usage to a subset of the terminology (Table 8). The complete

SNOMED CT terminology was used in 64% (49/76) of the publications. A subset of the terminology was used in 30% (23/76). The size of these subsets could vary from less than 10 concepts [47] to several thousand [37].

Table 8. Subset of Systematized Nomenclature of Medicine Clinical Terms used in publications.

Subset of SNOMED CT[a] used	Publications (N=76), n (%)
Full terminology	49 (64)
Subset	23 (30)
Not available	4 (5)

[a]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

Advanced Functionalities Used

SNOMED CT includes a large set of functionalities atop the classical ontology usage, among which the most interesting are the combinatorial possibilities that offer postcoordination. Table 9 shows whether a publication performed postcoordination to a certain extent. Among the 13 publications using this feature, 4 of them (5%) [30,35,36,49], all by the same first author, specifically mentioned the compositional grammar published by SNOMED CT [10]; however, the others do not elaborate nor propose simple postcoordination such as combining concepts with a "+" sign.

Table 9. Use of postcoordination.

Usage of postcoordination	Publications (N=76), n (%)
No	61 (80)
Yes	13 (17)
Not available	2 (3)

Discussion

Principal Findings

SNOMED CT is mostly used to represent information found in the complementary exam reports (18/796, 24%). This is potentially influenced by an important number of studies focusing on radiology [119] and pathology, as complementary exam reports are often produced by those divisions. Moreover, pathology being historically the field of SNOMED CT, it could have influenced its application in this domain. In addition, these types of reports are usually focused on specific clinical questions and arguably convey more specific informational content.

The second type of free text represented in our results is narrative notes (16/76, 21%). Potentially, this can be explained by the large conceptual span of SNOMED CT, which allows good informational coverage on textual data.

Finally, a large set of articles do not filter data for specific types. This is explained by publications focusing more on providing a solution to map SNOMED CT concepts to text in general, without targeting a specific type of document. This is supported by the fact that those publications have the mapping to SNOMED CT concepts as their final goal in 9 out of 14 (64%) publications, which is significantly higher than the rest of the selection (16/76, 21%).

In the selection, only 7 out of 76 (9%) publications focused on a language other than English. Multiple reasons can explain this predominance of the English language in research studies. First, NLP is known to be dependent on language. Work performed in a language cannot easily be transferred to other languages. Therefore, the overhead to begin NLP research in another language is substantial and brings few rewards in the first stages, as the breakthrough has already been published in another language.

Second, SNOMED CT—like most international classifications and ontologies—was first published only in English. Rule-based methods, which are the most frequently used methods for SNOMED CT mapping, rely on the assumption that the description of a concept can be directly mapped to free text, which is not possible when the language of the text is not the language of the classification. However, translations of SNOMED CT exist for Spanish, Swedish, and recently French [120]. Therefore, there is hope for new developments as the barriers to the language start to be overcome.

Finally, several publications use public data sets such as the i2b2-shared task data sets [33,34,41,50,51] or the MIMIC II [52] data set as the sources of narrative documents. These public data sets are valuable for promoting research in NLP on clinical free text and are the subject of many publications. The availability of such resources in languages other than English is scarce.

Unsurprisingly, the most frequent reason for mapping to SNOMED CT is information extraction (44/76, 39%), as the ability of SNOMED CT to represent medical knowledge is the core feature of this terminology. Nonetheless, 26 articles (34.21%) used the resulting mapping as a feature in a classification task, usually using a learning algorithm such as support vector machines or conditional random fields [53,54]. SNOMED CT is used in these cases as a proxy for the semantic content of the data, between free text and structured data, to simplify the task of classification and improve results.

Similarity evaluation is the goal in 4 publications (5.26%). Whether it is to compare cases [55], documents [29,37], or concepts [56], the similarity is computed using the SNOMED CT concepts. Both the polyhierarchy and the defining relationships can be used to compute the semantic distance between concepts. However, only 3 of the publications used them. This is an example of the added value SNOMED CT can bring to the secondary use of medical data.

Only 21% (16/76) of the publications mapped free text to SNOMED CT as a final objective. This is explained by the large number of publications reusing a mapping tool developed in a previous publication for new goals. To illustrate this phenomenon, Nguyen et al [38] reuse the software Medtex presented their study in multiple publications [14,30,35,36,49]. This is also true for large publicly available tools such as MCVS [16,17,57] or MTERMS [41,42].

The 3 most represented software in the selection—Medtex, MTERMS, and MCVS—are not available for public use. They mainly appear in publications by teams that have developed them. However, 2 software packages are available under an open-source license and can be freely used to map free text to SNOMED CT concepts, Metamap (and MMTx), and cTAKES. These tools are available to perform automatic annotation with SNOMED CT; however, none of them are specifically aimed at this ontology nor do they include features such as postcoordination or multiple language support. There is currently no clear solution for mapping free text to SNOMED CT concepts out of the box with a specific focus on this ontology and its features. This could explain the overall small number of publications in the selection.

Rule-based methods are largely used to perform mapping (53/76, 70%). This tends to show that they are more suited for this task. This phenomenon could be due to the large number of concepts in SNOMED CT. The amount of annotated data needed to automatically map free text with more than 340,000 classes is enormous and would require an important investment.

The evaluations of the automatic mapping found in publications show that this is not a trivial task. Most solutions for mapping lack a clear and definitive evaluation, and when available, they usually focus on a small set of documents; they use a subset of the terminology or do not rely on a gold standard. This gap in research could be explained by several reasons.

The number of concepts in SNOMED CT is large, and all granularities coexist. To express a simple concept such as *Tuberculous pneumonia*, a single concept can be used: 80003002 (Tuberculous pneumonia [disorder]) or any combination of less granular concepts (233604007 | Pneumonia [disorder], 233618000 |Mycobacterial pneumonia [disorder], 56717001 | Tuberculosis [disorder], 113858008 |Mycobacterium tuberculosis complex [organism], etc). However, all these representations can be equally correct from a semantic point of view. Therefore, it is difficult to compute the recall as a gold standard, which usually represents only one of these representations. Moreover, SNOMED CT contains 18 subhierarchies focusing on different thematics (clinical findings, body structure, etc), which make the decision of which concept to use even more difficult. For example, the hierarchy of the observable entities defines what can be observed in a patient, but the clinical finding hierarchy contains the results of those observations. The choice between a finding and an observable entity is not always clear and can heavily depend on the context. Finally, the usage of postcoordinated terms increases the set of expressions that can be used to represent the same concept. Overall, the task of evaluating the automatic mapping of natural language to a SNOMED CT concept lacks a pragmatic and applicable method; therefore, it is often limited to small-scale evaluations or manual validations.

The version of SNOMED used in publications (SNOMED, SNOMED CT, or SNOMED RT) is not always specified, especially when the usage of this terminology is not the main goal of the research. Moreover, the usage of SNOMED CT is implicit when UMLS is used. This remark, as well as the small number of publications mentioning postcoordination, emphasizes the fact that SNOMED CT is often seen as a simple terminology, without the need to use its advanced features. This phenomenon is also shown by the fact that only a subset of the terminology is used in 64% (49/76) of the publications. Using a subset simplifies the mapping task by reducing complexity but also prevents from benefiting from the power of the polyhierarchy and the relationships among concepts.

As clinical free text is written in natural language and since SNOMED CT is designed as a formal language, it is surprising that very few papers use this functionality when mapping to free text. Although this can be explained by the fact that even if SNOMED International provides compositional grammar, there is, to the best of our knowledge, no explicit roadmap to use it for such a task. Postcoordination requires deep knowledge of the terminology and access to a terminology server that handles the resulting data. As SNOMED International is not a software provider, this has to be achieved either using the open-source server Snowstorm [121], for which SNOMED International does not provide technical support, or by relying on a private company software.

This work shows that although SNOMED CT is widely used in health care, its use to represent free-text data still remains a challenge. Polyhierarchy and compositional grammar are at the core of SNOMED CT and they can bring significant value to data; however, when it comes to mapping concepts to free text, there seems to be a margin for approaches that take advantage of those features. The same can be observed on the usage of SNOMED CT to process free text in languages other than English.

Although machine learning is clearly on the rise in multiple fields of medical informatics and scientific research in general, it is rarely used to map free text to SNOMED CT, most probably because of the size of the corpus needed to train on such a large set of classes. In contrast, rule-based symbolic approaches seem more suited and are used to map large terminologies to free-text data. A combination of the strengths of both hybrid approaches could be a way to improve performance.

Finally, an openly available tool that would process free texts and map them to SNOMED CT concepts is yet to be created.

Limitations

Although the review has been conducted following a systematic approach, this work has some limitations.

The last publication research was conducted in August 2019. It is possible that new publications have been published since then. As we have observed, the number of publications selected per year is reducing; therefore, we consider the impact of this gap to be arguably small. Although the recursive reference review has been performed with the aim of broadening the scope of the included papers, it is possible that some studies that have not yet been cited by other papers have not been considered. For example, the high-throughput phenotyping NLP system described by Schlegel et al. [122] did not appear in the search nor during the recursive reference review. This system uses a series of linguistic and semantic indexes to process clinical data and characterizes it using ontologies such as SNOMED CT and the International Classification of Diseases 10.

In the selection, a large number of publications are published by the same groups of authors and propose similar works. This could result in an overestimation of the impact of those publications on a complete selection.

Finally, it is possible that because of the choice to focus on biomedical databases to gather publications, some articles published on more engineering-oriented databases have not been included.

Conclusions

In conclusion, clinical free-text processing and SNOMED CT have been an important subject for research, but the number of publications has been diminishing in recent years. Most of the publications that we found mapped free text to SNOMED CT to obtain a semantic representation of the data and used it as a first step toward other goals such as document classification or information retrieval.

Almost none of the publications used advanced features of SNOMED CT, such as the polyhierarchy or postcoordination. Most publications conceive SNOMED CT only as a terminology, a dictionary, or a resource for synonyms.

Publications focusing on languages other than English are rare and, if software exists for mapping English free text to SNOMED CT, most of them are not available for public use or focus on UMLS and not strictly on SNOMED CT. There is currently no easy solution for mapping free-text data into the SNOMED CT concepts, especially if the source language is different from English or if postcoordination is needed.

However, the need for formal semantic representation of health care data and the secondary use of free-text data is high, and automatic encoding into a compositional ontology could be a way to achieve interoperability.

Acknowledgments

This research was funded by the Language and Communication Network of the University of Geneva.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of 76 articles included in the review process.
[PDF File (Adobe PDF File), 507 KB-Multimedia Appendix 1]

References

1. Kierkegaard P. E-prescription across Europe. Health Technol 2012 Dec 20;3(3):205-219. [doi: 10.1007/s12553-012-0037-0]
2. Lau L, Shakib S. Towards Data Interoperability: Practical Issues in Terminology Implementation and Mapping. In: HIC 2005 and HINZ 2005: Proceedings. HIC 2005 and HINZ 2005: Proceedings Health Informatics Society of Australia; 2005 Presented at: HIC 2005: Thirteenth National Health Informatics Conference ; HINZ 2005: Fourth Health Informatics Conference; 2005; Brunswick East, Vic.
3. Grimson J, Murphy J. The Jupiter approach to interoperability with healthcare legacy systems. Medinfo 1995;8 Pt 1:367-371. [Medline: 8591200]
4. Randorff Højen A, Rosenbeck Gøeg K. Snomed CT implementation. Mapping guidelines facilitating reuse of data. Methods Inf Med 2012;51(6):529-538. [doi: 10.3414/ME11-02-0023] [Medline: 23038162]
5. SNOMED CT Starter Guide. SNOMED Confluence. URL: https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide?preview=/28742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf [accessed 2019-06-14]
6. SNOMED. URL: http://www.snomed.org/ [accessed 2019-06-14] [WebCite Cache ID http://www.snomed.org/]
7. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014 Feb;21(e1):e11-e19 [FREE Full text] [doi: 10.1136/amiajnl-2013-001636] [Medline: 23828173]
8. Members. SNOMED. URL: https://www.snomed.org/our-customers/members [accessed 2020-09-01]

9. Agrawal A, He Z, Perl Y, Wei D, Halper M, Elhanan G, et al. The readiness of SNOMED problem list concepts for meaningful use of electronic health records. Artif Intell Med 2013 Jun;58(2):73-80. [doi: 10.1016/j.artmed.2013.03.008] [Medline: 23602702]
10. SNOMED CT Compositional Grammar. SNOMED Confluence. URL: https://confluence.ihtsdotools.org/display/SLPG/SNOMED+CT+Compositional+Grammar [accessed 2020-09-01]
11. SNOMED CT concept model. SNOMED. URL: https://confluence.ihtsdotools.org/display/DOCGLOSS/SNOMED+CT+concept+model [accessed 2020-09-01]
12. Cohen K, Demner-Fushman D. Biomedical Natural Language Processing. Amsterdam: John Benjamins Publishing Company; 2014:978-990.
13. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
14. Zuccon G, Wagholikar AS, Nguyen AN, Butt L, Chu K, Martin S, et al. Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology. AMIA Jt Summits Transl Sci Proc 2013;2013:300-304 [FREE Full text] [Medline: 24303284]
15. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011 Oct;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]
16. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006 Jun;81(6):741-748. [doi: 10.4065/81.6.741] [Medline: 16770974]
17. Garvin JH, Elkin PL, Shen S, Brown S, Trusko B, Wang E, et al. Automated quality measurement in Department of the Veterans Affairs discharge instructions for patients with congestive heart failure. J Healthc Qual 2013;35(4):16-24. [doi: 10.1111/j.1945-1474.2011.195.x] [Medline: 23819743]
18. Bakhshi-Raiez F, de Keizer NF, Cornet R, Dorrepaal M, Dongelmans D, Jaspers MW. A usability evaluation of a SNOMED CT based compositional interface terminology for intensive care. Int J Med Inform 2012 May;81(5):351-362. [doi: 10.1016/j.ijmedinf.2011.09.010] [Medline: 22030036]
19. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc 2003:699-703 [FREE Full text] [Medline: 14728263]
20. Cornet R, de KN. Forty years of SNOMED: a literature review. BMC Med Inform Decis Mak 2008 Oct 27;8 Suppl 1:S2 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S2] [Medline: 19007439]
21. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Feb 05;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180]
22. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/ [accessed 2020-09-01]
23. Embase. URL: https://www.embase.com/#search [accessed 2020-09-01]
24. Web of Science. URL: https://apps.webofknowledge.com [accessed 2019-06-14]
25. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Yearb Med Inform 2018 Mar 05;02(01):41-51. [doi: 10.1055/s-0038-1637976]
26. 2004AA UMLS Documentation. US National Library of Medicine. URL: https://www.nlm.nih.gov/archive/20080407/research/umls/archive/2004AA/UMLSDOC.html [accessed 2020-09-01]
27. US National Library of Medicine. National Institutes of Health (NIH). URL: https://www.nlm.nih.gov/research/umls/Snomed/snomed_represented.html [accessed 2020-09-01]
28. So E, Park H. Mapping medical records of gastrectomy patients to SNOMED CT. Stud Health Technol Inform 2011;169:764-768. [Medline: 21893850]
29. Přečková P, Zvárová J, Zvára K. Measuring diversity in medical reports based on categorized attributes and international classification systems. BMC Med Inform Decis Mak 2012 Apr 12;12:31 [FREE Full text] [doi: 10.1186/1472-6947-12-31] [Medline: 22498343]
30. Nguyen A, Lawley M, Hansen D, Colquist S. Structured pathology reporting for cancer from free text: Lung cancer case study. In: HIC 2010: Proceedings. 2012 Presented at: 18th Annual Health Informatics Conference: Informing the Business of Healthcare; August 24-26, 2010; Melbourne URL: https://search.informit.com.au/documentSummary;dn=429807729626920;res=IELHEA;type=pdf
31. Patrick J, Wang Y, Budd P. An automated system for conversion of clinical notes into SNOMED clinical terminology. In: Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68. 2007 Presented at: Australasian symposium on ACSW frontiers; January 30 - February 2, 2007; Ballarat, Victoria, Australia p. 226. [doi: 10.1007/978-1-4471-2801-4_13]
32. Wang Y, Patrick JD. Cascading Classifiers for Named Entity Recognition in Clinical Notes. In: WBIE '09: Proceedings of the Workshop on Biomedical Information Extraction. 2009 Presented at: WBIE: Workshop on Biomedical Information Extraction; September 2009; Bulgaria p. 42 URL: https://www.aclweb.org/anthology/W09-4507

33. Gobbel GT, Reeves R, Jayaramaraja S, Giuse D, Speroff T, Brown SH, et al. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. J Biomed Inform 2014 Apr;48:54-65 [FREE Full text] [doi: 10.1016/j.jbi.2013.11.008] [Medline: 24316051]
34. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010 Oct;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
35. Nguyen A, Moore J, Zuccon G, Lawley M, Colquist S. Classification of pathology reports for cancer registry notifications. Stud Health Technol Inform 2012;178:150-156. [Medline: 22797034]
36. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010;17(4):440-445 [FREE Full text] [doi: 10.1136/jamia.2010.003707] [Medline: 20595312]
37. Mabotuwana T, Lee MC, Cohen-Solal EV. An ontology-based similarity measure for biomedical data-application to radiology reports. J Biomed Inform 2013 Oct;46(5):857-868 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.013] [Medline: 23850839]
38. Nguyen A, Lawley M, Hansen D, Colquist S. A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text. In: HIC 2009: Proceedings; Frontiers of Health Informatics - Redefining Healthcare. 2009 Presented at: HIC 2009: Frontiers of Health Informatics; August 19-21, 2009; National Convention Centre Canberra, Australia p. 196. [doi: 10.1007/978-1-84882-803-2_11]
39. Elkin A, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc 2008 Nov 06:172-176 [FREE Full text] [Medline: 18998791]
40. Brown SH, Elkin PL, Rosenbloom ST, Fielstein E, Speroff T. eQuality for all: extending automated quality measurement of free text clinical narratives. AMIA Annu Symp Proc 2008:71-75 [FREE Full text] [Medline: 18999230]
41. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annu Symp Proc 2011;2011:1639-1648 [FREE Full text] [Medline: 22195230]
42. Goss FR, Plasek JM, Lau JJ, Seger DL, Chang FY, Zhou L. An evaluation of a natural language processing tool for identifying and encoding allergy information in emergency department clinical notes. AMIA Annu Symp Proc 2014;2014:580-588 [FREE Full text] [Medline: 25954363]
43. Plasek JM, Goss FR, Lai KH, Lau JJ, Seger DL, Blumenthal KG, et al. Food entries in a large allergy data repository. J Am Med Inform Assoc 2016 Apr;23(e1):e79-e87 [FREE Full text] [doi: 10.1093/jamia/ocv128] [Medline: 26384406]
44. Zhou L, Lu Y, Vitale CJ, Mar PL, Chang F, Dhopeshwarkar N, et al. Representation of information about family relatives as structured data in electronic health records. Appl Clin Inform 2014;5(2):349-367 [FREE Full text] [doi: 10.4338/ACI-2013-10-RA-0080] [Medline: 25024754]
45. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
46. Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. In: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems. Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems IOS Press; 2007 Presented at: Medinfo 2007; 2007; Brisbane, Australia p. 2325 URL: https://www.aclweb.org/anthology/L08-1366/
47. Wagholikar A, Lawley MJ, Hansen DP, Chu K. Identifying symptom groups from Emergency Department presenting complaint free text using SNOMED CT. AMIA Annu Symp Proc 2011;2011:1446-1453 [FREE Full text] [Medline: 22195208]
48. Warden GI, Lacson R, Khorasani R. Leveraging terminologies for retrieval of radiology reports with critical imaging findings. AMIA Annu Symp Proc 2011;2011:1481-1488 [FREE Full text] [Medline: 22195212]
49. Nguyen A, Moore J, Lawley M, Hansen D, Colquist S. Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. Stud Health Technol Inform 2011;168:117-124. [Medline: 21893919]
50. Patrick JD, Nguyen DH, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. J Am Med Inform Assoc 2011;18(5):574-579 [FREE Full text] [doi: 10.1136/amiajnl-2011-000302] [Medline: 21737844]
51. Jindal P, Roth D. Extraction of events and temporal expressions from clinical narratives. J Biomed Inform 2013 Dec;46 Suppl:S13-S19 [FREE Full text] [doi: 10.1016/j.jbi.2013.08.010] [Medline: 24022023]
52. Henriksson A, Conway M, Duneld M, Chapman WW. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. AMIA Annu Symp Proc 2013;2013:600-609 [FREE Full text] [Medline: 24551362]
53. Li D, Savova G, Schuler K, Kipper-Schuler K, Savova G, Schuler K. Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. In: Proceedings of the workshop on current trends in biomedical natural language processing. 2008 Presented at: BioNLP 2008: Current Trends in Biomedical Natural Language Processing; 2008; Columbus, Ohio p. 94. [doi: 10.3115/1572306.1572326]

54. Aseervatham S, Bennani Y. Semi-structured document categorization with a semantic kernel. Pattern Recognition 2009 Sep;42(9):2067-2076. [doi: 10.1016/j.patcog.2008.10.024]
55. Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 2006 Dec;39(6):697-705 [FREE Full text] [doi: 10.1016/j.jbi.2006.01.004] [Medline: 16554186]
56. Pivovarov R, Elhadad N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. J Biomed Inform 2012 Jun;45(3):471-481 [FREE Full text] [doi: 10.1016/j.jbi.2012.01.002] [Medline: 22289420]
57. Matheny ME, FitzHenry F, Speroff T, Green JK, Griffith ML, Vasilevskis EE, et al. Detection of infectious symptoms from VA emergency department and primary care clinical documentation. Int J Med Inform 2012 Mar;81(3):143-156. [doi: 10.1016/j.ijmedinf.2011.11.005]
58. Tahmasebi AM, Zhu H, Mankovich G, Prinsen P, Klassen P, Pilato S, et al. Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. J Digit Imaging 2019 Feb;32(1):6-18 [FREE Full text] [doi: 10.1007/s10278-018-0116-5] [Medline: 30076490]
59. Jackson R, Patel R, Velupillai S, Gkotsis G, Hoyle D, Stewart R. Knowledge discovery for deep phenotyping serious mental illness from electronic mental health records. F1000Res 2018;7:210 [FREE Full text] [doi: 10.12688/f1000research.13830.2] [Medline: 29899974]
60. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Classification of forensic autopsy reports through conceptual graph-based document representation model. J Biomed Inform 2018 Jun;82:88-105 [FREE Full text] [doi: 10.1016/j.jbi.2018.04.013] [Medline: 29738820]
61. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Extracting cancer mortality statistics from death certificates: a hybrid machine learning and rule-based approach for common and rare cancers. Artif Intell Med 2018 Jul;89:1-9. [doi: 10.1016/j.artmed.2018.04.011] [Medline: 29754799]
62. Nguyen AN, Truran D, Kemp M, Koopman B, Conlan D, O'Dwyer J, et al. Computer-assisted diagnostic coding: effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. AMIA Annu Symp Proc 2018;2018:807-816 [FREE Full text] [Medline: 30815123]
63. Zvára K, Tomečková M, Peleška J, Svátek V, Zvárová J. Tool-supported interactive correction and semantic annotation of narrative clinical reports. Methods Inf Med 2017 May 18;56(3):217-229. [doi: 10.3414/ME16-01-0083] [Medline: 28451691]
64. Zhang R, Liu J, Huang Y, Wang M, Shi Q, Chen J, et al. Enriching the international clinical nomenclature with Chinese daily used synonyms and concept recognition in physician notes. BMC Med Inform Decis Mak 2017 May 02;17(1):54 [FREE Full text] [doi: 10.1186/s12911-017-0455-z] [Medline: 28464923]
65. Lin C, Hsu CJ, Lou YS, Yeh SJ, Lee CC, Su SL, et al. Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. J Med Internet Res 2017 Nov 06;19(11):e380 [FREE Full text] [doi: 10.2196/jmir.8344] [Medline: 29109070]
66. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Automated cancer registry notifications: validation of a medical text analytics system for identifying patients with cancer from a state-wide pathology repository. AMIA Annu Symp Proc 2016;2016:964-973 [FREE Full text] [Medline: 28269893]
67. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. AMIA Annu Symp Proc 2015;2015:953-962 [FREE Full text] [Medline: 26958232]
68. Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Med Inform Decis Mak 2015 Jul 15;15:53 [FREE Full text] [doi: 10.1186/s12911-015-0174-2] [Medline: 26174442]
69. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 classification of cancers from free-text death certificates. Int J Med Inform 2015 Nov;84(11):956-965. [doi: 10.1016/j.ijmedinf.2015.08.004] [Medline: 26323193]
70. Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. J Biomed Inform 2014 Jun;49:148-158 [FREE Full text] [doi: 10.1016/j.jbi.2014.01.012] [Medline: 24508177]
71. Ou Y, Patrick J. Automatic structured reporting from narrative cancer pathology reports. Electronic Journal of Health Informatics 2014;8(2).
72. Hong Y, Kahn CE. Content analysis of reporting templates and free-text radiology reports. J Digit Imaging 2013 Oct;26(5):843-849 [FREE Full text] [doi: 10.1007/s10278-013-9597-4] [Medline: 23553231]
73. So EY, Park HA. Exploring the possibility of information sharing between the medical and nursing domains by mapping medical records to SNOMED CT and ICNP. Healthc Inform Res 2011 Sep;17(3):156-161 [FREE Full text] [doi: 10.4258/hir.2011.17.3.156] [Medline: 22084810]
74. Butt L, Zuccon G, Nguyen A, Bergheim A, Grayson N. Classification of cancer-related death certificates using machine learning. Australas Med J 2013;6(5):292-299 [FREE Full text] [doi: 10.4066/AMJ.2013.1654] [Medline: 23745151]
75. Skeppstedt M, Kvist M, Dalianis H. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. International Renewable Energy Conference. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/521_Paper.pdf [accessed 2020-12-28]

76. ul Muntaha S, Skeppstedt M, Kvist M, Dalianis H. Entity Recognition of Pharmaceutical Drugs in Swedish Clinical Text. 2012 Presented at: The Fourth Swedish Language Technology Conference; 2012; Lund University, Sweden p. 77-78 URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.5525&rep=rep1&type=pdf
77. Davis K, Staes C, Duncan J, Igo S, Facelli JC. Identification of pneumonia and influenza deaths using the Death Certificate Pipeline. BMC Med Inform Decis Mak 2012 May 08;12:37 [FREE Full text] [doi: 10.1186/1472-6947-12-37] [Medline: 22569097]
78. Liu H, Wagholikar K, Wu ST. Using SNOMED-CT to encode summary level data - a corpus analysis. AMIA Jt Summits Transl Sci Proc 2012;2012:30-37 [FREE Full text] [Medline: 22779045]
79. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. J Am Med Assoc 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]
80. Martiniz D, Li Y. Information extraction from pathology reports in a hospital setting. In: Proceedings of the 20th ACM international conference on Information and knowledge management. 2011 Presented at: 20th ACM international conference on Information and knowledge management; 2011; Glasgow, Scotland. [doi: 10.1145/2063576.2063846]
81. Fung KW, Xu J, Rosenbloom ST, Mohr D, Maram N, Suther T. Testing Three Problem List Terminologies in a simulated data entry environment. AMIA Annu Symp Proc 2011;2011:445-454 [FREE Full text] [Medline: 22195098]
82. Lee DH, Lau FY, Quan H. A method for encoding clinical datasets with SNOMED CT. BMC Med Inform Decis Mak 2010 Sep 17;10:53 [FREE Full text] [doi: 10.1186/1472-6947-10-53] [Medline: 20849611]
83. Přečková P. Language of Czech Medical Reports and Classification Systems in Medicine. European Journal of Biomedical Informatics. URL: https://www.ejbi.org/scholarly-articles/language-of-czech-medical-reports-and-classification-systems-in-medicine.pdf
84. Arnot-Smith J, Smith AF. Patient safety incidents involving neuromuscular blockade: analysis of the UK National Reporting and Learning System data from 2006 to 2008. Anaesthesia 2010 Nov;65(11):1106-1113 [FREE Full text] [doi: 10.1111/j.1365-2044.2010.06509.x] [Medline: 20840604]
85. Wang Y. Annotating and Recognising Named Entities in Clinical Notes. 2006 Presented at: Proceedings of the ACL-IJCNLP 2009 Student Research Workshop USA: Association for Computational Linguistics; 2006; The University of Sydney URL: https://www.aclweb.org/anthology/P09-3003.pdf [doi: 10.3115/1667884.1667888]
86. Matheny ME, Fitzhenry F, Speroff T, Hathaway J, Murff HJ, Brown SH, et al. Detection of blood culture bacterial contamination using natural language processing. AMIA Annu Symp Proc 2009 Nov 14;2009:411-415 [FREE Full text] [Medline: 20351890]
87. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics 2009 Feb 05;10 Suppl 2:S1 [FREE Full text] [doi: 10.1186/1471-2105-10-S2-S1] [Medline: 19208184]
88. Wang Y. UIMA-based clinical information extraction system. Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP. URL: http://www.lrec-conf.org/lrec2008/IMG/ws/programme/W16.pdf [accessed 2020-12-28]
89. Schuler K, Kaggal V, Masanz J, Ogren P, Savova G. System Evaluation on a Named Entity Corpus from Clinical Notes. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). 2008 Presented at: LREC'08; May 2008; Marrakech, Morocco URL: https://www.aclweb.org/anthology/L08-1365/
90. Ruch P, Gobeill J, Lovis C, Geissbühler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak 2008 Oct 27;8 Suppl 1:S6 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S6] [Medline: 19007443]
91. Ryan A, Patrick J, Herkes R. Introduction of enhancement technologies into the intensive care service, Royal Prince Alfred Hospital, Sydney. Health Inf Manag 2008;37(1):40-45. [doi: 10.1177/183335830803700105] [Medline: 18245864]
92. Patrick J, Wang Y, Budd P, Brandt S, Rogers B, Herkes R, et al. Developing SNOMED CT Subsets from Clinical Notes for Intensive Care Service. Health Care and Informatics Review Online. URL: https://cdn.ymaws.com/www.hinz.org.nz/resource/collection/0F09C2E4-7A05-49FB-8324-709F1AB2AA2F/F38_Patrick.pdf [accessed 2020-12-28]
93. Johnson SB, Bakken S, Dine D, Hyun S, Mendonça E, Morrison F, et al. An electronic health record based on structured narrative. J Am Med Inform Assoc 2008;15(1):54-64 [FREE Full text] [doi: 10.1197/jamia.M2131] [Medline: 17947628]
94. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 2006 Dec;39(6):589-599 [FREE Full text] [doi: 10.1016/j.jbi.2005.11.004] [Medline: 16359928]
95. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, Greevy R, et al. eQuality: electronic quality assessment from narrative clinical reports. Mayo Clin Proc 2006 Nov;81(11):1472-1481. [doi: 10.4065/81.11.1472] [Medline: 17120403]
96. Shah NH, Rubin DL, Supekar KS, Musen MA. Ontology-based annotation and query of tissue microarray data. AMIA Annu Symp Proc 2006:709-713 [FREE Full text] [Medline: 17238433]
97. Pakhomov S, Buntrock J, Duffy P. High throughput modularized NLP system for clinical text. In: Proceedings of the ACL 2005 on Interactive poster and demonstration sessions. 2005 Presented at: ACLdemo '05; June 26, 2005; Ann Arbor, Michigan. [doi: 10.3115/1225753.1225760]
98. Long W. Extracting diagnoses from discharge summaries. AMIA Annu Symp Proc 2005:470-474 [FREE Full text] [Medline: 16779084]

99. Rindflesch TC, Pakhomov SV, Fiszman M, Kilicoglu H, Sanchez VR. Medical facts to support inferencing in natural language processing. AMIA Annu Symp Proc 2005:634-638 [FREE Full text] [Medline: 16779117]
100. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005 May 05;5:13 [FREE Full text] [doi: 10.1186/1472-6947-5-13] [Medline: 15876352]
101. Burkhart L, Konicek R, Moorhead S, Androwich I. Mapping parish nurse documentation into the nursing interventions classification: a research method. Comput Inform Nurs 2005;23(4):220-229. [doi: 10.1097/00024665-200507000-00010] [Medline: 16027538]
102. Cornet R, Van Eldik A, De Keizer N. Inventory of tools for Dutch clinical language processing. Stud Health Technol Inform 2012;180:245-249. [Medline: 22874189]
103. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc 2010;17(6):646-651 [FREE Full text] [doi: 10.1136/jamia.2009.001024] [Medline: 20962126]
104. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
105. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput Biol 2013 Feb;9(2):e1002854 [FREE Full text] [doi: 10.1371/journal.pcbi.1002854] [Medline: 23408875]
106. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
107. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
108. MetaMap. URL: https://metamap.nlm.nih.gov/ [accessed 2020-09-01]
109. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
110. Denny JC, Smithers JD, Miller RA, Spickard A. 'Understanding' medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc 2003;10(4):351-362 [FREE Full text] [doi: 10.1197/jamia.M1176] [Medline: 12668688]
111. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics 2009 Sep 17;10(S9). [doi: 10.1186/1471-2105-10-s9-s14]
112. Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak 2018 Sep 14;18(S3):-. [doi: 10.1186/s12911-018-0654-2]
113. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on Translat Bioinforma 2009;2009:56-60 [FREE Full text] [Medline: 21347171]
114. Dai M, Shah NH, Xuan W, Musen MA, Watson SJ, Athey BD, et al. An efficient solution for mapping free text to ontology terms. In: AMIA summit on translational bioinformatics. 2008 Presented at: AMIA summit on translational bioinformatics; 2008; San Francisco URL: https://knowledge.amia.org/amia-55142-tbi2008a-1.650887/t-002-1.985042/f-001-1.985043/a-041-1.985157/an-041-1.985158?qr=1
115. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-174 [FREE Full text] [Medline: 7719797]
116. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392-402 [FREE Full text] [doi: 10.1197/jamia.M1552] [Medline: 15187068]
117. Vickers DM, Lawley MJ. Mapping Existing Medical Terminologies to SNOMED CT: An Investigation of the Novice User's Experience. In: HIC 2009: Proceedings; Frontiers of Health Informatics - Redefining Healthcare, National. 2009 Presented at: HIC 2009: Frontiers of Health Informatics - Redefining Healthcare; August 19-21, 2009; National Convention Centre Canberra, Australia p. 46 URL: https://cdn.ymaws.com/hisa.site-ym.com/resource/resmgr/hic2009/DVickers.pdf
118. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011 Sep 01;18(5):552-556. [doi: 10.1136/amiajnl-2011-000203]
119. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
120. An exemplar of collaboration: The first release of the SNOMED CT common French translation Internet. SNOMED. URL: http://www.snomed.org/news-and-events/articles/first-release-common-french-translation [accessed 2020-09-01]
121. IHTSDO/snowstorm. GitHub. 2020. URL: https://github.com/IHTSDO/snowstorm [accessed 2020-09-01]
122. Schlegel DR, Crowner C, Lehoullier F, Elkin PL. HTP-NLP: a new NLP system for high throughput phenotyping. Stud Health Technol Inform 2017;235:276-280. [Medline: 28423797]

Abbreviations

cTAKES: clinical Text Analysis and Knowledge Extraction System
CUI: concept unique identifier
MCVS: Mayo Clinic Vocabulary Server
MMTx: Metamap transfer
MTERMS: medical Text Extraction Reasoning and Mapping System
NLM: National Library of Medicine
NLP: natural language processing
OBA: Open Biomedical Annotator
RapTAT: Rapid Text Annotation Tool
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SNOMED RT: SNOMED reference terminology
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 28.09.20; peer-reviewed by S Madani, P Elkin; comments to author 14.10.20; revised version received 24.11.20; accepted 30.11.20; published 26.01.21

Please cite as:
Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C
Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review
J Med Internet Res 2021;23(1):e24594
URL: http://www.jmir.org/2021/1/e24594/
doi: 10.2196/24594
PMID: