UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Thèse 2021

Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

Semantic interoperability of clinicaldata: a multi-dimensional approach

Gaudet-Blavignac, Christophe

How to cite

GAUDET-BLAVIGNAC, Christophe. Semantic interoperability of clinicaldata: a multi-dimensional approach. Doctoral Thesis, 2021. doi: 10.13097/archive-ouverte/unige:157668

This publication URL: https://archive-ouverte.unige.ch/unige:157668
Publication DOI: 10.13097/archive-ouverte/unige:157668

© The author(s). This work is licensed under a Other Open Access license
https://www.unige.ch/biblio/aou/fr/guide/info/references/licences/

2021

Semantic interoperability of clinical data: a multi-dimensional approach

DOCTORAL THESIS

Christophe Gaudet-Blavignac

Supervisor: Pr. Christian Lovis

Biomedical Sciences, Global Health

Faculty of Medicine, University of Geneva

1. Acknowledgements

Starting this thesis and finishing it was one of the greatest challenges of my life. Of course, as no one is an island, it would have never been possible without the numerous people that supported me and my work both directly and indirectly.

First, I would like to thank Professor Christian Lovis, my supervisor. Throughout those five years he pushed me to develop my skills, to learn how to lead a project and to become autonomous. By sharing his vision with me and infecting me with the semantic virus, he guided me, staying behind and putting me in the light every time he could. For this I am deeply thankful and realize how rare this kind of liberty and support is. Then I would like to thank my thesis committee members Professor Marcel Salathé and Professor Vincent Mooser. They followed my progress, giving me precious insights and asking questions that helped me improve my work. I also deeply thank Professor Eric Wehrli for his help in the adaptation of his incredible tool Fips to new horizons, Professor Stéfan Darmoni for providing so many precious SNOMED CT translations and Professor Antoine Geissbuhler for introducing me to the world of Medical Informatics.

I would also like to thank my colleagues from the Division of Medical Information Sciences. Throughout those five years numerous people worked in this unique division of the hospital. They all contributed to this unique environment and I would like to thank them for the luck of working with them and for the privilege of counting them among my friends. I would like to especially thank Vasiliki Foufi that helped me improve and correct my work, even in the smallest details and at the latest hours. And in no particular order I would like to thank Mina Bjelogrlic, Arnaud Robert, Lucien Troillet, Pascal Rémy, Jessica Rochat, David Issom, Cyrille Duret, Mirjam Mattei, Jean-Philippe Goldman, Dominique Guerin, Christelle Fournier, Pierre Gilquin, Raphaël Chevrier, Jason Chenaud, Dina Vishnyakova, Frédéric Ehrler, Philippe Bauman, Dominique Baillon, Marzia del Zotto, Sébastien Abegg, Gérome Pasquier, Joel Spaltenstein, Jean-Louis Raisaro, Andrea Rudaz and all the others that I can't name here but that contributed to make those years interesting, happy and fun.

I would like to thank the Institute of Global Health and particularly Professor Antoine Flahault, Nathalie Bot, Nadia Elia, and Lemlem Girmatsion for their help, and for always bringing solutions and insightful advices throughout this adventure.

I would like to thank my family for their constant support and for believing in me a lot more than myself. My parents Richard and Brigitte Gaudet-Blavignac, my sister Fabienne and my brother Olivier who although they didn't completely understand what I was doing and why this "SNOMED CT" interested me so much, still watched my presentations, read my articles and encouraged me. I would like to thank my closest friends, Loïc Serafin, Arthur Giroux, Paul-Elie Kupsc, Lou Richard and Pablo Venturelli for always being there when I needed them, whether it was for discussing my work or to blow some steam playing a videogame.

I also want to have a special thought for my late brother Marc that unfortunately won't be able to see me finish this work. I know that he wouldn't have been interested at all by my subject but I'm confident he would have found a way to speak of it in a funny way and to make us laugh in the process.

Finally, I would like to thank my wife Chloé and my two children, Lucien and his future little brother or sister. I would never have managed to finish this work without the thoughtful help of my wife and her fight against my procrastinating tendencies. I started this thesis unmarried and without children and I finish it married with two kids and happier than ever.

2. Academic output during the thesis period

2.1. Journal publications related to this thesis

1. Gaudet-Blavignac, C., Raisaro, JL., Touré, V., Österle, S., Crameri, K. & Lovis, C. *"A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study."* JMIR Medical Informatics 9, no. 6 (June 24, 2021): e27591.
2. Gaudet-Blavignac, C., Rudaz, A. & Lovis, C. *"One list to rule them all and many semantics to bind them: Building a shared, scalable and sustainable source for the problem oriented medical record."* Submitted to the Journal of Medical Internet Research in March 2021
3. 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."* Journal of Medical Internet Research (2021), 23(1), e24594.
4. Gaudet-Blavignac, C.Foufi, V., Timakum, T., , Lovis, C., & Song, M. *"Mining of Textual Health Information from Reddit : Analysis of Chronic Diseases With Extracted Entities and Their Relations."* Journal of Medical Internet Research (2019), 21(6), e12876.

2.2. Conference papers related to this thesis

1. Gaudet-Blavignac, C., Foufi, V., Wehrli, E., & Lovis, C. *"Automatic Annotation of French Medical Narratives with SNOMED CT Concepts."* Studies in Health Technology and Informatics (2018), 247, 710-714.
2. Gaudet-Blavignac, C., Foufi, V., Wehrli, E., & Lovis C. *"Reconnaissance et représentation automatiques de concepts médicaux français en SNOMED CT"* TALMED 2019: Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical, Lyon, France.

2.3. Posters and oral presentations related to this thesis

1. Oral presentation: *"Speak like a clinician: bridging the gap between controlled vocabularies and medical language"*, UNIGE Data Science Seminars, 2021
2. Oral presentation: *"Codification automatique des motifs de venue"*, PHAST Journée francophone SNOMED CT, 2021
3. Poster and oral presentation: *"CoviDB, donner du sens aux données, HUG journée de l'innovation"*, 2020
4. Oral presentation: *"Traduction automatique du Français médical en SNOMED CT"*, UNIGE Langage et Communication, journée des doctorants, 2019
5. Oral presentation: *"Usage of SNOMED CT to represent French medical data in the Electronic Health Record"*, Collogues du G6, Campus Biotech, 2019
6. Oral presentation: *"Traduction automatique du langage médical Français en SNOMED CT"*, PHAST Journée francophone SNOMED CT, 2019
7. Oral presentation: *"Automatic annotation of French medical narratives with SNOMED CT concepts"*, Swiss eHealth Summit, 2018 (Best Contribution Award)
8. Poster: *"Translating patient-related narratives into SNOMED CT to enable interoperability of healthcare data"*, Geneva Health Forum, 2018
9. Poster and oral presentation: *"Lexico-semantic resources and corpora for the automatic translation of Electronic Health Records into SNOMED CT"*, UNIGE Langage et Communication, journée des doctorants, 2017
10. Poster and oral presentation: *"MicMac : Automatic SNOMED CT translation"*, UNIGE Langage et Communication, journée des doctorants, 2016
11. Poster: *"Health Big Data Framework to Support Clinical Research"*, Geneva Health Forum, 2016

2.4. Other publications made during the thesis period

1. Turbé H., Bjelogrlic M., Robert A., Gaudet-Blavignac C., Goldman JP., Lovis C. *"Adaptive Time-Dependent Priors and Bayesian Inference to Evaluate SARS-CoV-2 Public Health Measures Validated on 31 Countries."* Front Public Health. 2021 Jan 21;8:583401.
2. Foufi V., Ing Lorenzini K., Goldman JP., Gaudet-Blavignac C., Lovis C., Samer C. *"Automatic Classification of Discharge Letters to Detect Adverse Drug Reactions."* Stud Health Technol Inform. 2020 Jun 16;270:48-52.
3. Rochat J., Gaudet-Blavignac C., Del Zotto M., Ruiz Garretas V., Foufi V., Issom D., Samer C., Hurst S., Lovis C. Citizens' *"Participation in Health and Scientific Research in Switzerland."* Stud Health Technol Inform. 2020 Jun 16;270:1098-1102.
4. Chevrier R., Foufi V., Gaudet-Blavignac C., Robert A., Lovis C. *"Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review."* J Med Internet Res. 2019 May 31;21(5):e13484.
5. Lovis C., Gaudet-Blavignac C., Chevrier R., Robert A., Issom D., Foufi V. *"BigData, intelligence artificielle, blockchain : guide pratique [Bigdata, artificial intelligence and blockchain for dummies]."* Rev Med Suisse. 2018 Sep 5;14(617):1559-1563. French. PMID: 30226672.
6. Foufi V., Gaudet-Blavignac C., Chevrier R., Lovis C. *"De-Identification of Medical Narrative Data."* Stud Health Technol Inform. 2017;244:23-27.
7. Walpoth BH., Meyer M., Gaudet-Blavignac C., Baumann P., Gilquin P., Lovis C. *"The International Hypothermia Registry (IHR): Dieter's ESAO Winter Schools and Beat's International Hypothermia Registry."* Int J Artif Organs. 2017 Jan 1;40(1):40-42
8. Vishnyakova D., Gaudet-Blavignac C., Baumann P., Lovis C. *"Clinical Data Models at University Hospitals of Geneva."* Stud Health Technol Inform. 2016;221:97-101.

3. Abstract (English)

Since the development of the first hospital information systems in late 1960s, digitalization has become a major driver of all aspects pertaining to health, leading to major paradigm shifts in the field, notably towards data-driven personalized medicine, among others. Large initiatives, such as the Meaningful Use in the United states, promoted the wide adoption of Electronic Health Records (EHR) both in ambulatory and hospital settings. Today, the vast majority of hospitals have adopted some sort of EHR.

With the growing production of data, in volume and coverage, expectations have raised sharply accompanied with new analytical means and approaches and an important need to promote data sharing and exchange. Consequently, data interoperability has become a major hurdle in all communities and for all usages, is it care, administrative support, research or public health, and especially in case of urgent public health needs such as experienced during the covid-19 pandemic.

While the definition of interoperability can vary, it is commonly accepted to define it in three frequently overlapped layers: technical, semantic and process interoperability. Technical interoperability relates to the technological means used to structure, send and receive data. Semantic interoperability is needed to understand its meaning. Process interoperability allows the coordination of systems and workflows in a harmonized manner.

The field of interoperability has been largely shaped by standardization organizations and the multiple standards they produce. Those standards have multiple purposes and granularities and cover almost every aspect of healthcare, from simple data transfer to the standardization of care processes in a hospital. They constitute the focus of industrial approaches as well as active research. Among them, controlled vocabularies, such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) or the International Statistical Classification of Diseases and Related Health Problems (ICD), aim at representing the information contained in the data and, therefore, are key components of semantic interoperability.

However, in this profusion of standards, interoperability remains an unresolved challenge. By restricting the scope of this work to clinical data interoperability, several observations can be made on its limitations.

Firstly, a standard can only bring interoperability to the extent of its adoption. Secondly, regardless of their type (semantic, technical, process), standards are not neutral. They are developed by organizations with a purpose and their adoption depends on it. Thirdly as clinical care is composed of many actors, roles, cultural habits and needs, enforcing a unique standard is neither possible nor desirable. Finally, in a connected, digitalized world, data must be shareable and understandable across domains and standards. However, this would require the creation of mappings from each standard to all the other ones, which is, to date, not done and would represent enormous work to maintain.

This work is based on three hypotheses:

SNOMED CT in conjunction with a limited number of other formal knowledge representations can be used as a formal interlingua to represent clinical information properly.

This hypothesis has been at the core of the Swiss Personalized Health Network (SPHN) initiative and is the focus of the first article of this thesis. It implies that a framework aiming to create interoperability for multiple communities such as research, healthcare and regulatory agencies needs to be strongly semantically driven. Hence, the first pillar of the strategy being a semantic framework represented by a set of compositional concepts encoded in SNOMED CT and other specifically relevant controlled

vocabularies. Then, to allow the storage and transfer of data, the second pillar proposes the usage of a formal language to describe the data without enforcing any data model. Finally, these semantically rich, formally described data can be transformed through conversion mechanics into various existing data models to be used by various communities without requiring mappings between every standard. The first article of this thesis describes the design of this strategy by the Clinical Data Semantics Interoperability Working Group of the SPHN and the results of its successful implementation in Switzerland which confirms this first hypothesis.

The second hypothesis is based on the observation that the combinatorial power of such an interlingua exceeds the effective needs for clinical activities and assumes that it can be reduced to a meaningful and manageable size.

This has been tested through the building and implementation of a common list to represent patients' problems in the Geneva University Hospitals (HUG). This list was built and is constantly enriched by gathering and manually curating a list of expressions encountered in real clinical documents written by clinicians. This list was specifically targeted at imitating clinicians' language and representing useful, clinically relevant concepts rather than being exhaustive in its coverage. In a second phase, each expression is encoded in a series of semantic dimensions, including SNOMED CT, to allow multiple uses of the data. After four years of usage and numerous updates and refinements, the use of the list has been evaluated and reported in the second article of this thesis. The results confirmed that the list has become in four years a central axis of the EHR. It proved to be usable, semantically interoperable and small enough to be manageable, proving the validity of this second hypothesis.

Finally, the last hypothesis of this work states that the SNOMED CT interlingua can be used to represent the information contained in clinical narratives framing the challenge as an automatic translation task.

This idea emerged from two observations: First, most concepts expressed in clinical settings cannot be expressed as single elements in a classification. As any language used by humans, there is a need to combine multiple concepts to fully express a meaning. Secondly, SNOMED CT presents similarities with a natural language. Indeed, with a compositional grammar, more than 350,000 concepts and 1,000,000 relations, SNOMED CT can be used and combined into complex post-coordinated sentences in a similar way words are combined into sentences in a natural language like French or English. Therefore, the challenge of representing narratives into SNOMED CT could be framed as a translation from French to SNOMED CT as a target language.

As a first step toward this goal, a scoping review has grounded the possibility to represent the content of clinical free text and narratives using SNOMED CT as a conceptual framework. However, the review showed that little work has been done at exploiting SNOMED CT's combinatorial power, expressing a sentence in a specific language using a sentence in SNOMED CT. Thus, the last part of this work has been devoted to (a) investigating the translation of SNOMED CT concepts into English, French and German by participating in the translation of the starter kit for Switzerland and extracting SNOMED CT concepts from new subtypes of the English language, such as social media content, (b) manually representing complex medical expressions found in French narratives in SNOMED CT and (c) implementing an attempt of an automatic French to SNOMED CT translator.

The experience gathered while translating concepts into multiple languages highlighted key specificities that were injected in the task of translating French medical text into SNOMED CT post-coordinated sentences. Finally, in this work a multilingual automatic translation tool was adapted to translate text written in French (source language) to SNOMED CT as a target language with encouraging results.

In this thesis, a multi-dimensional approach for semantic interoperability on various types of data and for multiple use cases is presented. This approach emphasizes the challenges of bringing interoperability in a domain which covers multiple communities, types of data, standards and information systems. It confirms that no single solution exists and that targeted, semantically-centered approaches are necessary. From large national frameworks to very specific documents written in any natural language, interoperability must penetrate every layer of healthcare. The proposed solution is based firstly on strong semantics by using compositional controlled vocabularies to create a computer-readable interlingua without enforcing a data model, it then restricts the representation complexity to a useful set of concepts encountered in practice and finally exploits the compositional capabilities of the SNOMED CT interlingua to represent complex narrative data into post-coordinated SNOMED CT sentences.

Our approach defines a new, semantically interoperable landscape for clinical data that can leverage new opportunities proposed by the growth of personalized medicine.

4. Abstract (French)

Depuis l'apparition des premiers systèmes d'information hospitaliers, à la fin des années soixante, la numérisation est devenue un moteur de développement majeur de la santé, menant à des changements de paradigme, notamment concernant la santé personnalisée. D'importantes initiatives nationales, telles que la Meaningful Use aux Etats-Unis, ont abouti à l'adoption massive de Dossiers Patients Informatisés (DPI) que ce soit dans le secteur hospitalier ou ambulatoire. De nos jours, la vaste majorité des hôpitaux disposent d'un DPI.

Avec la production grandissante de données en variété comme en volume, les attentes ont grandi rapidement, accompagnées par de nouvelles approches analytiques et d'importants besoins de partage et d'échange. De ce fait, le manque d'interopérabilité des données est devenu un obstacle majeur pour toutes les communautés, que ce soit dans les soins, la gestion administrative, la recherche ou la santé publique, et d'autant plus dans des situations de crise telles que la pandémie qui a frappé le monde en 2019.

Si la définition de ce qu'est l'interopérabilité peut varier selon les sources consultées, il est communément accepté de la séparer en trois couches qui se superposent : L'interopérabilité technique, qui couvre les moyens technologiques utilisés pour établir une connexion, structurer les données, les envoyer puis les recevoir ; l'interopérabilité sémantique, nécessaire à la compréhension et à l'utilisation de ces données une fois échangées ; l'interopérabilité des processus qui permet la coordination harmonieuse des activités des différents acteurs de la santé.

Le domaine de l'interopérabilité a été, pour une grande partie, forgé par les organisations de standardisation et les nombreux standards qu'elles produisent. Ceux-ci ont des buts, des couvertures et des granularités multiples. Il en existe pour à peu près tous les aspects de la santé, du simple transfert des données à la standardisation des processus d'un hôpital. Ils sont à la fois un élément important pris en compte par l'industrie et le sujet de recherches actives dans le monde académique. Parmi eux, les vocabulaires contrôlés comme la Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) ou la Classification Internationale des Maladies (ICD) visent à représenter l'information contenue dans les données et sont donc des composants majeurs de l'interopérabilité sémantique.

Cependant, malgré cette profusion de standards, l'interopérabilité reste un défi non résolu. En réduisant le champ de cette thèse à l'interopérabilité des données cliniques, plusieurs observations peuvent être faites.

Premièrement, un standard ne peut créer de l'interopérabilité que s'il est adopté. Deuxièmement, indépendamment de son type, un standard n'est jamais neutre. Il est toujours développé par une organisation dans un but précis qui conditionnera son adoption. Troisièmement, comme le monde des soins est composé de multiples acteurs, rôles, habitudes, cultures et besoins, forcer l'utilisation d'un standard unique n'est ni possible ni désirable. Finalement, dans un monde connecté et digitalisé, les données doivent pouvoir être partagées et réutilisées qu'importe le domaine ou le standard. Cependant, cela requiert la création de ponts entre chaque standard, ce qui représente, à ce jour, une charge de travail et de maintenance irréalisable.

Ce travail se base donc sur trois hypothèses :

SNOMED CT combiné à un nombre limité d'autres représentations de l'information peut être utilisé comme une interlingua formelle pour représenter adéquatement l'information clinique.

Cette première hypothèse est au cœur de l'approche de l'initiative Swiss Personalized Health Network (SPHN) et est le sujet du premier article de cette thèse. Elle postule que pour rendre interopérables les

données entre les multiples communautés que sont la santé, la recherche et les instances de régulation, il est nécessaire d'adopter une approche centrée sur la sémantique. De ce fait, le premier pilier de la stratégie est un cadre sémantique robuste, constitué d'un ensemble de concepts composables encodés en SNOMED CT et dans d'autres vocabulaires contrôlés pertinents. Dans un deuxième temps, afin de permettre le stockage et le transfert de ces données, le deuxième pilier propose l'utilisation d'un langage formel pour la description des données plutôt que de forcer l'adoption d'un modèle de données. Finalement, ces données sémantiquement définies et formellement décrites peuvent être transformées à travers des mécanismes de conversion pour correspondre à des modèles de données variés, afin de permettre leur utilisation par de multiples communautés. Le premier article de cette thèse décrit la création de cette approche par le Clinical Data Semantics Interoperability Working Group du SPHN et les premiers résultats de son implémentation en Suisse qui confirment cette première hypothèse.

La seconde hypothèse est basée sur l'observation que la puissance combinatoire de l'interlingua, définie précédemment, dépasse largement les besoins liés à l'activité clinique. Elle postule qu'elle peut être réduite à une taille à la fois utile et appréhendable par des humains.

La construction et implémentation d'une liste commune pour représenter les problèmes des patients au sein des Hôpitaux Universitaires de Genève (HUG) ont permis de tester cette hypothèse. Cette liste a été construite en réunissant et traitant manuellement une liste d'expressions rencontrées dans des documents écrits par des cliniciens. Elle cible spécifiquement leur langage et représente un ensemble de concepts utiles et cliniquement pertinents plutôt qu'une liste exhaustive de tous les possibles. Dans un second temps, chaque expression de la liste a été encodée dans une série de dimensions sémantiques, incluant SNOMED CT, pour permettre de multiples utilisations secondaires de ces données. Après quatre ans d'utilisation et de nombreuses mises à jour et ajustements, l'utilisation de la liste a fait l'objet d'une évaluation qui constitue le deuxième article de cette thèse. Les résultats ont confirmé qu'en quatre ans la liste s'est imposée comme un axe central du dossier patient. Elle s'est montrée utilisable et interopérable tout en conservant une taille limitée qui permet son maintien par une équipe relativement restreinte, démontrant de ce fait la validité de cette seconde hypothèse.

Finalement, la dernière hypothèse de ce travail postule que l'interlingua définie précédemment peut être utilisée pour représenter l'information exprimée dans les documents cliniques narratifs, en concevant ce défi comme une tâche de traduction automatique du langage.

Cette idée est née de deux observations principales : Premièrement, la plupart des concepts utilisés dans un environnement clinique ne peuvent pas être exprimés par un simple élément dans une classification. Comme dans tout langage utilisé par des humains, il est nécessaire de pouvoir exprimer des idées par l'association de concepts. Deuxièmement, SNOMED CT présente des similarités importantes avec un langage naturel. En effet, avec une grammaire compositionnelle, plus de 350,000 concepts et 1,000,000 relations, cette terminologie peut être utilisée pour combiner des concepts en des structures complexes de la même manière que les mots, dans un texte écrit dans une langue naturelle comme le français, se combinent en phrases. De ce fait, le défi de représenter l'information contenue dans des documents en langage naturel peut être vu comme une tâche de traduction du français (langue source) dans une nouvelle langue (langue cible).

Dans un premier temps, une scoping review a permis d'établir qu'il était possible de représenter l'information contenue dans les documents narratifs en utilisant SNOMED CT comme cadre conceptuel. Cependant, cette revue a également mis en évidence que peu avait été entrepris pour exploiter la puissance compositionnelle de SNOMED CT afin de représenter des phrases en langage naturel. C'est pourquoi la dernière partie de ce travail a été consacrée à (a) investiguer la possibilité

de traduction de concepts SNOMED CT en français, allemand et anglais en contribuant à la traduction du starter kit de SNOMED CT en Suisse et en extrayant des concepts SNOMED CT de différents types de langage tels que l'anglais utilisé sur les réseaux sociaux, (b) représenter des expressions médicales complexes contenues dans des documents narratifs en SNOMED CT d'abord manuellement puis (c) en implémentant une tentative de traduction automatique français (langue source) – SNOMED CT (langue cible).

L'expérience accumulée grâce à la traduction de concepts dans de multiples langages a permis de mettre en évidence des spécificités qui ont été injectées directement dans la tâche de traduction du français en phrases SNOMED CT post-coordonnées. Ce travail a donné lieu à l'adaptation d'un outil de traduction automatique multilingue avec des résultats encourageants.

En résumé, dans cette thèse, une approche multi-dimensionnelle est proposée pour rendre sémantiquement interopérables des types de données variés pour des cas d'utilisation multiples. Cette approche souligne les défis posés par un domaine qui regroupe de multiples communautés, types de données, standards et systèmes d'information. Elle confirme qu'il n'existe pas de solution unique et que des approches ciblées et centrées sur la sémantique sont nécessaires. Des grands projets nationaux aux documents en langage naturel ultra spécifiques, l'interopérabilité doit entrer dans chaque secteur de la santé. L'approche proposée est basée premièrement sur une sémantique forte, en utilisant des vocabulaires contrôlés compositionnels pour créer une interlingua formelle lisible par un ordinateur sans forcer l'utilisation d'un modèle de données ; elle restreint ensuite cette interlingua à un ensemble utile et nécessaire de concepts et finalement exploite sa compositionalité pour représenter l'information complexe contenue dans les documents narratifs.

Cette approche dessine un nouveau paysage de l'interopérabilité des données cliniques qui tirera parti des nouvelles possibilités proposées par la santé personnalisée.

5. Outline

1. Acknowledgements 1
2. Academic output during the thesis period 2
   2.1. Journal publications related to this thesis 2
   2.2. Conference papers related to this thesis 2
   2.3. Posters and oral presentations related to this thesis 2
   2.4. Other publications made during the thesis period 3
3. Abstract (English) 4
4. Abstract (French) 7
5. Outline 10
6. General Introduction 12
   6.1. Interoperability 12
   6.2. Electronic Health Records (EHRs) 15
      6.2.1. Types of data in EHRs 16
      6.2.2. Secondary use of data 17
   6.3. Standardization organizations 17
      6.3.1. International Organization for Standardization (ISO) 17
      6.3.2. World Health Organization (WHO) 18
      6.3.3. Integrating the Healthcare Enterprise (IHE) 18
      6.3.4. Health Level 7 (HL7) International 19
      6.3.5. The Clinical Data Interchange Standards Consortium (CDISC) 19
      6.3.6. The Observational Health Data Sciences and Informatics (OHDSI) 19
   6.4. Technical standards 19
      6.4.1. HL7 standards 20
      6.4.2. Digital Imaging and Communications in Medicine (DICOM) 22
      6.4.3. OMOP Common Data Model (CDM) 22
      6.4.4. OpenEHR 23
      6.4.5. ISO 13606-1 24
      6.4.6. CDISC Operational Data Model (CDISC ODM) 25
   6.5. Controlled vocabularies 25
      6.5.1. The uncontrolled vocabulary of controlled vocabularies 26
      6.5.2. Usage of controlled vocabularies in healthcare 27
      6.5.3. Properties of controlled vocabularies 28
      6.5.4. International Statistical Classification of Diseases and Related Health Problems (ICD) 28

6.5.5. Logical Observation Identifiers, Names, and Codes (LOINC) 29
6.5.6. Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) 30
6.5.7. Unified Medical Language System (UMLS) 32
6.6. Projects related to clinical data sharing for research 33
6.6.1. I2b2 tranSMART foundation 33
6.6.2. Strategic Health IT Advanced Research Projects consortium (SHARPn) project 34
6.7. Natural Language Processing (NLP) 34
6.8. Research questions 35
6.8.1. Hypothesis 1 36
6.8.2. Hypothesis 2 39
6.8.3. Hypothesis 3 40
7. Publications 41
7.1. Methodological contributions 41
7.1.1. Contributions to hypothesis 1 41
7.1.2. Contribution to Hypothesis 2 42
7.1.3. Contribution to Hypothesis 3 43
7.1.4. Other publications made during the thesis period 47
7.2. Publication manuscripts 49
7.2.1. Article 1 49
7.2.2. Article 2 59
7.2.3. Article 3 80
7.2.4. Article 4 98
7.2.5. Conference Article 1 116
7.2.6. Conference Article 2 121
8. Conclusions and perspectives 126
8.1. An interlingua for clinical data 126
8.2. A restricted set of useful concepts 127
8.3. An automatic interlingua translation for clinical narratives 127
8.4. General conclusions 128
9. Abbreviations 130
10. Bibliography 132

6. General Introduction

6.1. Interoperability

Interoperability is a concept used in every domain in which there is the need for two systems, people or institutions to work together. It boils down to the need for communication. In 2010, eight Global Health agencies called for action on health data and mentioned the improvement of interoperability between systems in the required actions for meeting the Millennium Development Goals [1]. In 2017, during the first United Nations World Data Forum, a collaborative on Sustainable Development Goals' (SDGs) data was founded [2]. A year later, a practitioner's guide was published by the collaborative to inform about the best practice around data interoperability, recognizing the subject as a key factor for the achievement and monitoring of the SDGs [3].

In healthcare, interoperability is the conceptualization of what happens when a team of health professionals takes care of a patient. A patient entering the emergency ward of a hospital will usually first be seen by the triage nurse assessing priority and overall situation. Based on the initial assessment, the emergency clinician will proceed to a more detailed assessment. In parallel, according to the needs, investigations and early care will be carried on. If required, the patient will be transferred to an inpatients ward for further care. At discharge, a summary of the encounter will be made available to the usual general practitioner of this patient. Such history is characterized by different processes and transitions between people and structures. In a non-digitalized world, this transfer of information is accomplished through paper patient records, telephone calls and face to face meetings. These transitions are important and subject of many studies, especially in handoffs [4,5]. The complexity of natural communication channels is progressively replaced by digitalization and demonstrates the challenges of "analogic" interoperability emphasizing the need to improve interoperability of electronic systems (Figure 1).

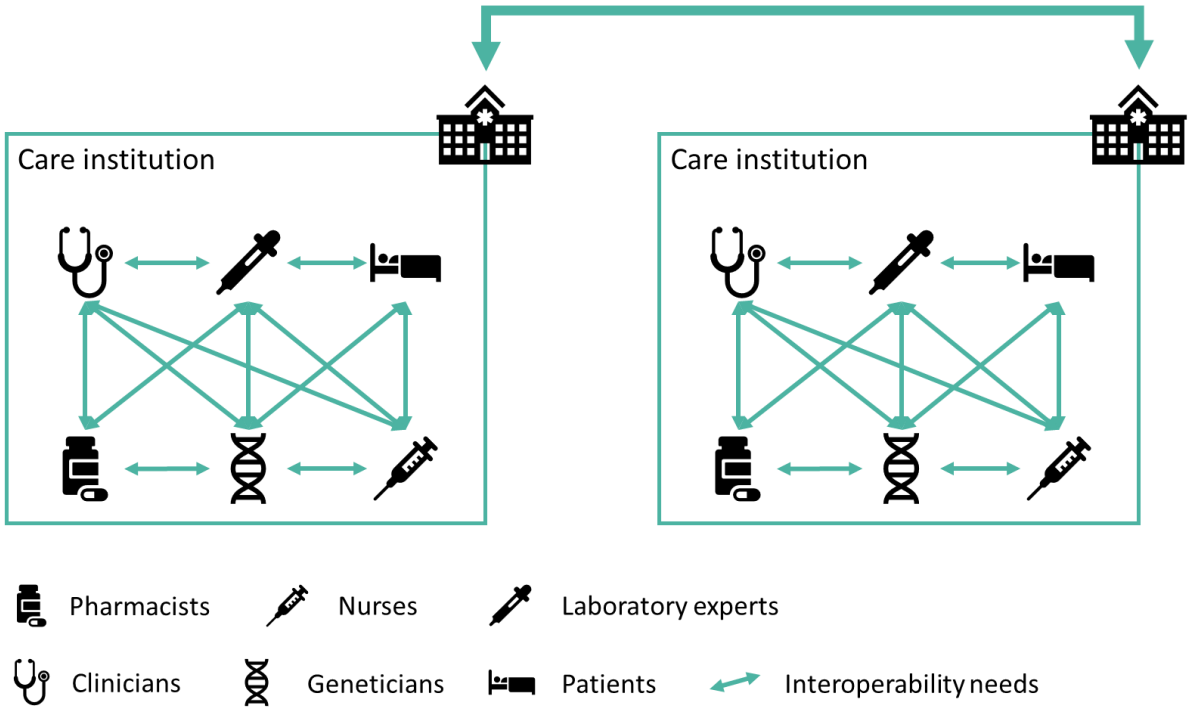


Figure 1 Interoperability needs for clinical data.

The definitions of interoperability are various. According to the Institute of Electrical and Electronics Engineers (IEEE) Standard Computer Dictionary, it is "the ability of two or more systems or components

to exchange information and to use the information that has been exchanged" [6]. In this definition, information exchange is the central goal with two different aspects of it, the ability to exchange the information followed by the ability to use it. According to the Healthcare Information and Management Systems Society's (HIMSS) Dictionary of Health Information and Technology Terms, Acronyms and Organizations, interoperability is "the ability of different information systems, devices or applications to connect, in a coordinated manner, within and across organizational boundaries" [7]. This definition reduces the scope to the technical challenge of connecting systems or otherwise computerized elements. In this work, we will adopt a broader definition that goes from the exchange of information to its usage.

Interoperability is commonly split into multiples layers or different types, but the number and nature of the split can vary. According to Braunstein et al. [8] three layers can be defined, the first being "transport interoperability", meaning the ability to exchange information, which for example would be the technical possibility to make a phone call or to send signals over telegraph infrastructures. The second layer is "structured interoperability" which is defined as the ability to structure the sent message so it can be parsed into defined blocks on the other end and therefore allowing to share data with any level of structure from forms with fields and value set to simple blocks of text. The last layer is "semantic interoperability" which is about sharing data that will be used and understood in the same way by the sender and the recipient. This last layer can be linked to the second part of the IEEE's definition. In healthcare, this could be exemplified by the ability to use the data to take clinical decisions about a patient.

For Benson et al. [9], interoperability in healthcare can be split into four different parts. "technical interoperability" which is broader than in the previous definition since it covers technical and structured interoperability, "semantic interoperability" and two new parts that are "process interoperability" and "clinical interoperability" that are both instantiations of the semantic interoperability. They relate to a level of interoperability that allows the workflow, the people and the institutions to function in a coordinated manner and clinical work to be carried out seamlessly.

Finally, for Blobel [10], there are five layers, first technical and structural which are similar to Braustein's definition of technical and structured, with the specificity of adding a syntactic layer which is a refinement of the structural interoperability specifying that information is well structured, for example following an existing standard such as the Clinical Document Architecture (CDA) [11]. The fourth layer is semantic interoperability followed by organizational interoperability which is similar to the process and clinical interoperability for Benson and represents the ability of an organization to perform common business processes in a coordinated manner. Those different definitions are summarized in table 1:

<table>
<tr><th>Global categories</th><th>Braunstein</th><th>Benson</th><th>Blobel</th></tr>
<tr><td rowspan="3">Technical</td><td>Transport</td><td rowspan="3">Technical</td><td>Technical</td></tr>
<tr><td rowspan="2">Structured</td><td>Structural</td></tr>
<tr><td>Syntactic</td></tr>
<tr><td>Semantic</td><td colspan="3">Semantic</td></tr>
<tr><td rowspan="2">Process</td><td rowspan="2"></td><td>Process</td><td rowspan="2">Organizational</td></tr>
<tr><td>Clinical</td></tr>
</table>

Table 1 Definitions of interoperability layers.

Those layers can be categorized in three coarse categories. The first one being “technical interoperability” in its broader sense, meaning the technical ability to transfer information, structured or not. It is important to note that the information transmitted does not need to be meaningful providing that the information is transmitted without alteration. Since the democratization of the internet and even more in the era of the Internet of Things (IoT), connecting systems do not constitute a big challenge anymore [12]. Any device can access the network and once on it, connect to any other connected device. However, to successfully exchange data, a common communication protocol must be chosen and adopted by the systems. Only then a message can be efficiently emitted and received. One of the most common protocols used online is the hypertext transfer protocol (HTTP) [13]. While most of the internet is accessed and browsed using it, few are the users that can list its specifications. The same is true about communication protocols in healthcare. If the system works properly, users can communicate without any knowledge of how the information is transmitted. In healthcare, this layer has been at the center of numerous projects, research studies or regulations. Technical interoperability is only possible if a specific standard is adopted [9] and multiple standards exist in healthcare to tackle this layer.

The second layer is “semantic interoperability”. Despite varying on other steps, the three definitions share this denomination. The ability to understand the shared information so that its meaning is the same for the recipient and the sender is crucial. A common medical term such as “shock” can have a different meaning depending on who receives the information (Figure 2). Any benefit from data sharing will come from the use of the shared data. For the semantics to be stable and usable by every recipient, the language used must be the same. It is easy to make a phone call to someone in China. The technical part is invisible to the user as he just needs to dial the number, and someone can answer at the other end of the world. But when it comes to understand and use the information transmitted, if both parties do not speak the same language (English or Mandarin in this case), it is not likely that the call will be fruitful. This analogy underscores the importance of being able to represent information correctly so that it can be used by everyone. It is usually achieved in healthcare using controlled vocabularies such as classifications, terminologies or ontologies. Those systems are designed to hold the semantics of the information in a non-ambiguous manner. They will be the focus of section 6.5.

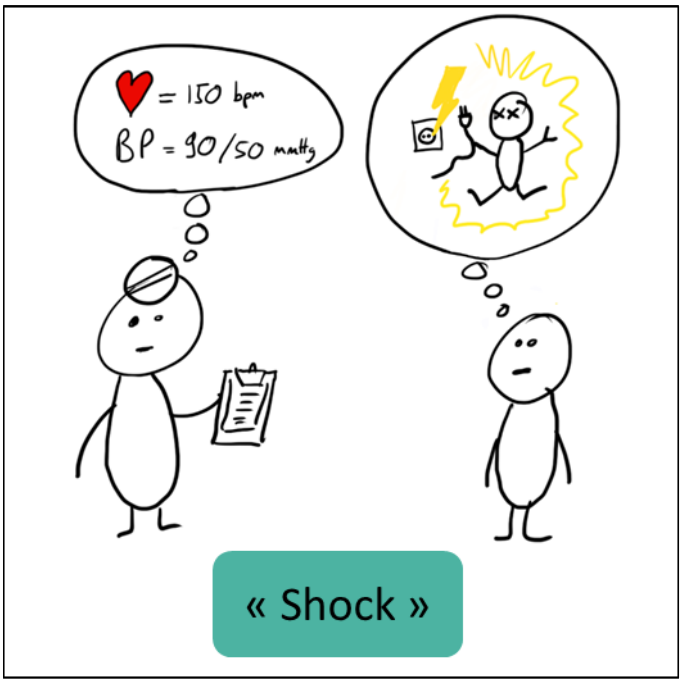


Figure 2 Semantic interoperability requires a communication language.

The last layer of this categorization is “process interoperability” in its broader sense: the ability to convert the shared data into actions or processes in an institution. If the two first layers are adequately implemented, it is then possible to build a workflow based on the shared information. It is worth noting that similarly to a controlled vocabulary for representing knowledge, languages exist to represent processes and reasoning over the data. For example, the Arden syntax [14] created in 1989 and currently maintained by Health Level 7 (HL7) is designed to allow computerized clinical reasoning using medical information.

It is important to emphasize two key aspects of the technical and semantic layers of interoperability. The first is that two sources of data can exchange information while having internal representations that are different. The second is that the data exchanged can be used in a meaningful way. So, similarly to human communication, interoperability does not require everybody to speak the same language or to have the same internal representation, but to be able to adopt an additional communication language that is characterized by two properties: words that convey a clear meaning and a grammar to associate these words into complex sentences.

6.2. Electronic Health Records (EHRs)

HIMSS dictionary defines EHR as “a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting” [7]. EHRs started as early as the ‘60s [15]. The first computerized record systems were not the complete systems that exist today allowing processes such as prescription or complementary exam requests and centralizing patient’s data [16,17]. However, one of the first advocated roles of EHRs has been to give quick access to critical information needed for patient care [17]. This need remained at the center of the EHRs’ role and became broader with the need to share data across institutions and countries. With each digitalization progress, new possibilities to improve the quality of care, efficiency and costs were unveiled [18].

In 2004 in the United States, a report written by the President’s Information Technology Advisory Committee (PITAC) highlighted key challenges in the digitalization of healthcare [19]. This report was the basis for the writing of the Health Information Technology for Economic and Clinical Health (HITECH) Act that was ratified in 2009 [20]. It contained what was later called “the meaningful use initiative” and proposed incentives for hospitals and practitioners that adopted an EHR system. This program that was set to promote the transition to a digitalized healthcare succeeded with an important increase in the percentage of hospitals using at least a basic electronic record. However, in 2016, a report of the Office of the National Coordinator for Health Information Technology (ONC) on the state of adoption of health technology in the United States stated that while 97% of the hospitals and 75% of the physicians adopted an EHR, interoperability was still a challenge and that insufficient specificity regarding standards adoption was a key barrier to its development [21]. Moreover, it was shown that the meaningful use initiative had unintended consequences such as market saturation, innovation vacuum, physician burnout or data obfuscation [22].

Nowadays, in industrialized countries, EHRs are broadly deployed and used in various types of institutions, such as hospitals, ambulatory clinics, insurance companies, pharma companies or research agencies [23]. They have become the norm when working with patient data. However, challenges remain to fully harvest to possibilities offered by a fully integrated EHR. As it has recently been shown, information blocking is still a barrier to interoperability of EHRs [24]. Distributed architectures that allow simultaneous queries and computation on multiple sites are a growing field of research with solutions such as i2b2, i2b2-transmart or Medco [25–27], but they often remain limited to research data and struggle to enter healthcare institutions.

6.2.1. Types of data in EHRs

EHRs are designed to store and allow access to patient information. The data contained are often divided into structured versus unstructured [28–33]. Structured data are usually understood as data that can be easily manipulated by computers for analytics, such as numbers, Booleans, categorical variables. Data such as gender, birthdate, date of death, laboratory results are examples of structured data. Unstructured data usually refers to signal data requiring complex analytics, such as texts, images, video and signals like audio, electrocardiographs or electroencephalographs. Documents and narratives (clinical notes, discharge letters, etc.) are thus considered as unstructured while containing most of the pertinent information. This work focuses on structured data and text data.

It is important to note that while the terms “structured” and “unstructured” are used thoroughly in literature, they are not accurate nor stable in their meaning. Natural language is almost always categorized as unstructured, as software cannot interpret it for its lack of computer readability. However, natural languages are highly structured in that they comply with a specific grammar and use a finite value set of words arranged in sentences with punctuation marks to convey meaning. If natural languages were truly unstructured, it would be a lot harder for humans to communicate. But they are seen as totally unstructured when processed by a computer, even if recent advances have been made in language understanding, the field specifically aiming at tackling this issue [34,35]. Therefore, when labeling data as structured or unstructured, it should be only limited to computer processing. In this view, free text is, indeed, unstructured. Throughout this work, for simplicity, the terms “structured” and “unstructured” will be used in the same way as in the literature.

When confronted to real data, the strict differentiation between structured and unstructured is impossible. The reality is a continuum that spans from totally structured data, such as lab values encoded in a classification with a result value that is represented by a specific data type and representing the result of the test, to a totally unstructured progress note written by a tired resident at 3 a.m. using in-house acronyms with multiple typographic errors. Narratives can be structured [36] and a structured variable can contain unstructured free text. Figure 3 displays a non-exhaustive list of data commonly found in EHRs classified according to their degree of structuring.

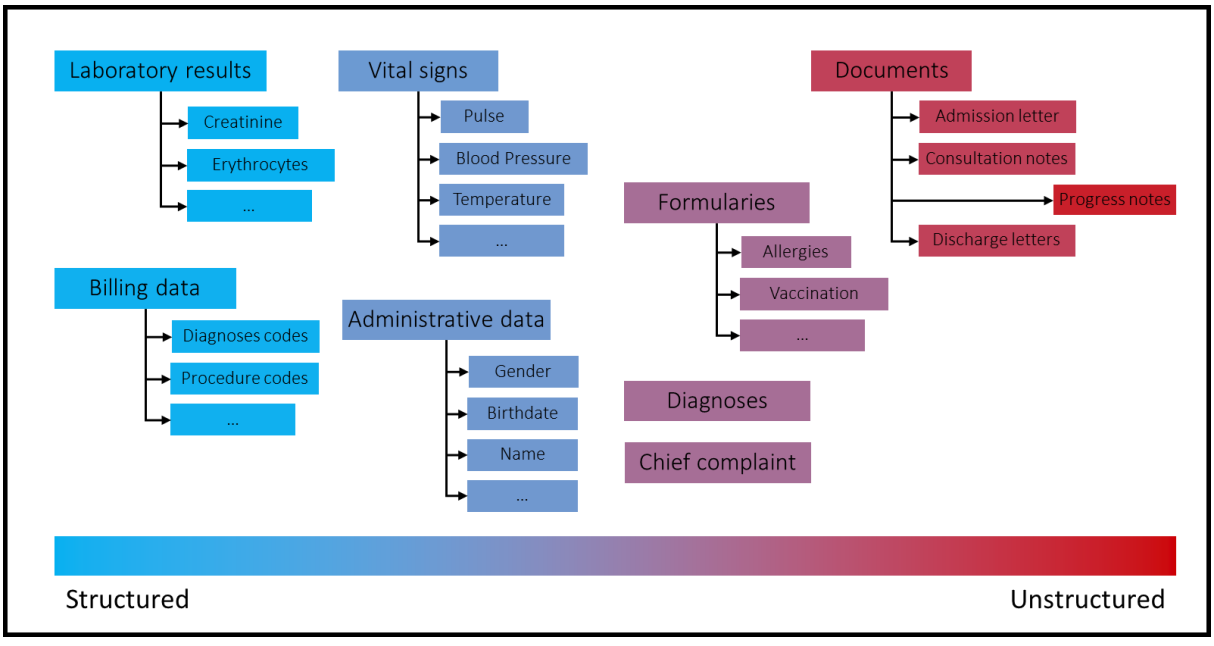


Figure 3 Examples of structured and unstructured data in EHR classified according to their degree of structuring.

While structured data is preferable for allowing efficient computer processing, in EHRs, the largest part of the information stored is in the form of free text [32,37,38]. Free text is notably rich in clinical information [32] and is the main manner in which clinicians express their reasoning about the patient's condition. Crucial data for surveillance such as adverse drug effects are often only documented in the post-event progress note [39] and therefore difficult to use in decision support or automated surveillance systems [40].

On the other hand, diagnoses, commonly stored as codes from the International Statistical Classification of Diseases and Related Health Problems (ICD), ninth or tenth revision [41], are broadly used for billing and therefore easily available and well structured. However, multiple studies reported challenges and issues reusing these data [42–44]. It has been shown that they are not always fit for purpose as the assignation of codes for billing is not performed with a clinical intention and therefore does not necessarily represent the clinical truth which prevents their usage in research [43,45]. The PITAC report already highlighted in 2004 that the classification systems historically used to code medical diagnoses and procedures for reimbursement and population statistics are not adequate for research purposes [19]. Moreover, even when the coding is made for clinical purpose and uses a standard terminology, such as the Logical Observation Identifiers, Names, and Codes (LOINC), it has been shown that there is often some disagreement and that research on coded information compared to research on textual sources can yield different results [38,46].

6.2.2. Secondary use of data

Secondary use of data is the use of data for additional purposes than the primary reason for their creation [7]. In an EHR, the data is generated by the clinician, the patient or by devices, but its main goal is to provide care. Therefore, any use of the generated data for another purpose is considered as secondary use. Common goals of secondary data use are surveillance, quality assurance, pharmacovigilance, public health and research [47].

6.3. Standardization organizations

The field of interoperability has been largely shaped by standardization organizations and the multiple standards they produce. Those organizations can be global as the International Organization for Standardization (ISO), or domain specific such as the World Health Organization (WHO), HL7 and the Clinical Data Interchange Consortium. As interoperability is a broad domain and this work focuses on the specific field of clinical data interoperability, only organizations dedicated to the standardization of clinical data will be described. However, it must be acknowledged that numerous other organizations exist. Interoperability in the biomedical domain is rich and rapidly evolving thanks to organizations such as the Global Alliance for Genomics and Health which aims at providing guidelines for the use of standards in genomic data sharing [48], the Open Biological and Biomedical Ontologies Foundry which develops ontologies for the biomedical domain [49], or the Gene Ontology Consortium [50] whose goal is to build computational models of biological systems and centralize information on the function of genes. The following non-exhaustive inventory aims at describing key stakeholders in clinical data standardization. This review focuses on describing organizations related to clinical data from three different communities: healthcare, research and industry.

6.3.1. International Organization for Standardization (ISO)

ISO is a global standard-setting organization with a membership of 165 national standards bodies. Created in 1947, it acts as a network of the world's leading standardizers and develops and publishes standards, technical specifications or guides [51]. As a global organization, ISO is not focused on a specific field and is involved in domains as various as quality management, energy management, environmental management, food safety or online security. One of these domains being health and

safety standards, it is one of the actors of healthcare interoperability. Since 1998, ISO includes a technical committee on health informatics that was involved in the creation of standards such as the ISO 13606-1 on electronic health record communication [52].

6.3.2. World Health Organization (WHO)

The WHO was created in 1948 as a specialized agency of the United Nations with the mandate to coordinate authority on international health issues. It includes employees from 150 countries and operates on a wide range of issues. One of the many products released by WHO is the WHO Family of International Classifications, a set of integrated classifications used to represent and classify health information across the world [53]. The most renowned member of this family is the International Statistical Classification of Diseases and Related Health Problems (ICD) which is described in the next section. Other classifications released by WHO include the International Classifications of Functioning or the Disability and Health, and the International Classification of Health Interventions.

6.3.3. Integrating the Healthcare Enterprise (IHE)

The IHE initiative was launched in 1998 by the Radiological Society of North America and the HIMSS. Coming from the healthcare and the industry, it was based on the hypothesis, later confirmed by the 2016 ONC report [21], that even if standards, such as HL7 and Digital Imaging and Communications in Medicine (DICOM), exist, interoperability was still not easily achieved as each vendor and healthcare facility implemented them in a specific manner that often prevented data sharing [54]. To solve this issue, IHE defined a technical framework that could be used by vendors and users to implement existing standards in a way that ensures specific functionalities. This was done by creating a common vocabulary to describe implementation needs and IHE Profiles that state how to use existing standards to meet those needs [55]. The process to produce and implement those profiles is described in Figure 4 and includes connectathons to make vendors, users and IHE members meet and connect real systems.

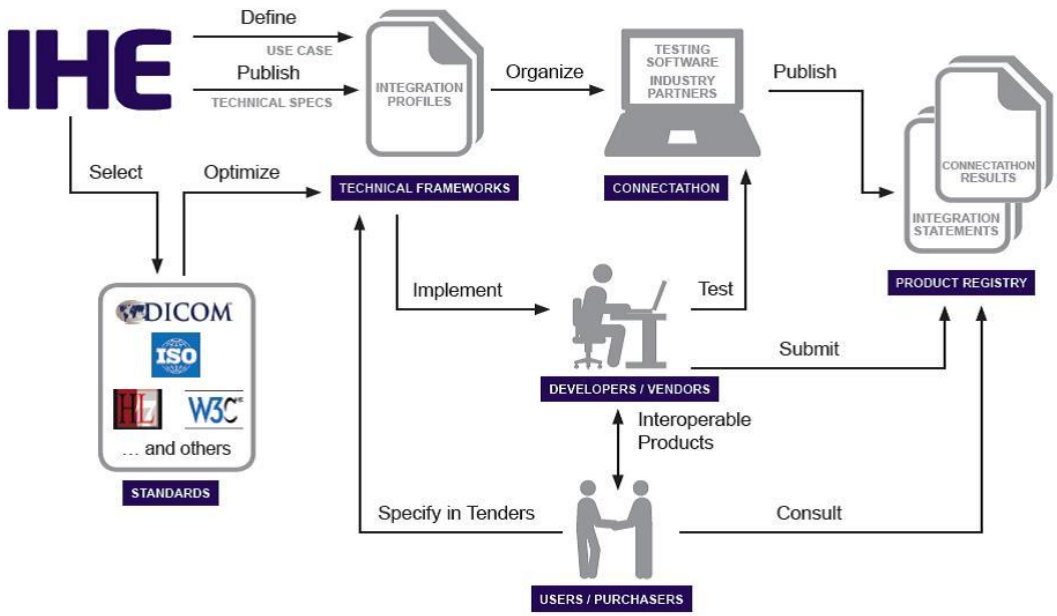


Figure 4 The IHE process [56].

6.3.4. Health Level 7 (HL7) International

Created in 1987, HL7 International is a nonprofit organization dedicated to interoperability. It aims at creating standards that allow the implementation of the seventh layer of the Open Systems Interconnection model (OSI model) [57]. The OSI model, initiated by ISO in 1977, describes the seven layers needed to build network communication, from the physical devices and cables that compose the system to the high-level application that will be accessed by the end user. This last layer is the focus of HL7 since its creation. Standards produced by HL7 include transactional messaging standards, such as HL7 Version 1, 2 and 3 and its last version called the Fast Healthcare Interoperability Resource (FHIR) [9,58]. HL7 also created the Reference Information Model (RIM), a data model on which is based HL7 version 3 and the Clinical Document Architecture (CDA), and exchange model for clinical documents [11,59]. It also maintains and develops the Arden Syntax, a markup language used to represent and share medical knowledge in an actionable format for decision support systems [14].

6.3.5. The Clinical Data Interchange Standards Consortium (CDISC)

CDISC, formed as a volunteer group in 1997 and a non-profit organization since 2000, is creating standards to support clinical and non-clinical research. These standards cover all steps of research from Protocol Representation Model [60] to Study Data Tabulation Model [61] and Analysis Data Model (ADaM) [62]. Some of the CDISC standards, such as the Biomedical Research Integrated Domain Group (BRIDG) Model, are recognized by ISO as international standards [63]. Since 2010, CDISC's standards are required to submit to the Food and Drug Association (FDA) in the United States and by the Pharmaceutical and Medical Devices Agency in Japan [64].

CDISC also provides a Controlled Terminology that can be used in its standards to encode variables [65]. This terminology is maintained and distributed by the National Cancer Institute (NCI) as part of the NCI thesaurus. On the possibility of using an existing terminology, such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), CDISC state on their website that it requires a paying license and that their controlled terminology should not depend on a non-free terminology. Moreover, SNOMED CT is lacking definitions which their controlled terminology has. On the other hand, other terminologies are mandatory in CDISC such as the Medical Dictionary for Regulatory Activities (MedDRA) for encoding adverse events [66].

6.3.6. The Observational Health Data Sciences and Informatics (OHDSI)

The OHDSI (pronounced odyssey) is a project born from the Observational Medical Outcomes Partnership (OMOP), a 5-year initiative that ran between 2008 and 2013 [67]. Its primary goal was to unify partners from the public and private sector (mainly pharmaceutical companies) to develop secondary use of observational medical data for pharmaceutical research by proposing a shared analytic platform based on a common data model. This partnership launched by the FDA was transferred in 2013 to the Reagan-Udall foundation and gave birth to OHDSI. The main contribution of OMOP was the creation of the OMOP Common Data Model (OMOP CDM) [68,69] which aims to propose a model in which a center could extract, transform and load (ETL) its data. In 2019, the OHDSI project spanned over 20 countries and more than 100 different healthcare facilities [70].

6.4. Technical standards

While this work focuses mainly on semantic interoperability, it is important to mention some of the standards focusing on technical interoperability, since they usually include a mention or a pathway to allow semantic interoperability and because any approach dedicated to solving semantic interoperability needs to interact with the technical layer. The described systems have been chosen empirically based on the relevance of the standards in the literature and with the aim of giving a picture of the available standards.

6.4.1. HL7 standards

6.4.1.1. HL7 version 1 (HL7 v1)

HL7 version 1 was the first standard released by the organization. While it was more of a proof of concept to define what would be the structure and content of such standard, it was allowing exchange of data related to patient admission, discharge and transfers in structured messages.

6.4.1.2. HL7 version 2 (HL7 v2)

This first version was soon replaced by HL7 v2 which improves considerably the range of information covered by adding messages to exchange orders, laboratory tests and treatments [71]. This version progressively became the most widely used healthcare standard in the world with more than 95% of the United States healthcare organizations using some variation of the version 2 and more than 35 countries having implementations of it [58]. HL7 v2 is being constantly enriched since its creation with its 2.9 version released in 2019 [72]. Part of its wide adoption is due to the backward compatibility of each of these versions. Adopting a standard is an important investment and the perspective to be able to keep a version for a long time without suffering compatibility issues is a strong incentive. Another advantage of HL7 v2 is that while its complete specification is enormous, the messages' syntax is simple to grasp. An HL7 message is composed of segments that are separated in fields. The fields are composed of components. A message starts with the message header component that specifies metadata like the sender, the type of separator used or the datetime of the message. Then, other components can be used to define the event that triggered the message, the patient identification, the encounter, the requested observation or the result of an observation. The complete specification of HL7 v2 messages is extraordinarily rich and cannot be summarized in a few words, but the general structure remains the same. Importantly, it is possible to exchange data encoded in a classification which is of utmost importance for semantic interoperability.

6.4.1.3. HL7 version 3 (HL7 v3)

In 1992, HL7 started to work on the successor of HL7 v2. The third version of the well-known standard called HL7 v3 was focused on solving the main pitfalls of the version 2 [73]. By design, this third version is strongly linked to the RIM [59], an object model designed to be the backbone of HL7 version 3 and a unique reference for the healthcare domain (Figure 5). Its classes and their relations are the building blocks of HL7 v3 messages and define what can be exchanged. It was designed with an object-oriented paradigm around three main classes: Act, Roles and Entities. "Act" represents something that happened or will happen, "Entity" represents any living or nonliving thing and "Role" can be understood as a competency that is expressed by an Entity. Those classes can have specializations, attributes and relationships with other classes. For example, the class "Patient" is a specialization of the Role class and linked to a spatialization of an Entity which is a living human and can be the subject of a Procedure which is a specialization of Act. Some attributes are present in multiple classes such as classCode which indicate the name of an Act, Role or entity. While classCode is supposed to define what the class is by giving it a name and is purely internal to HL7 v3, the code attribute is there to hold an external code from a classification such as ICD-10. Other attributes such as negationId or statusCode allow to respectfully negate the class or specify its state as active, inactive, cancelled, etc. As the RIM is designed to be the unique model for all information in healthcare, it is not usually used in its entirety. Constrained information models, also called profiles, can be defined to specify what classes and attributes can be used, what value-set, etc. HL7 v3 messages derived directly from the constrained models used and are rendered in Extensible Markup Language (XML) structures. Importantly, to enable interoperability between two systems the profiles used must be shared. Figure 5 displays the normative content of the RIM:

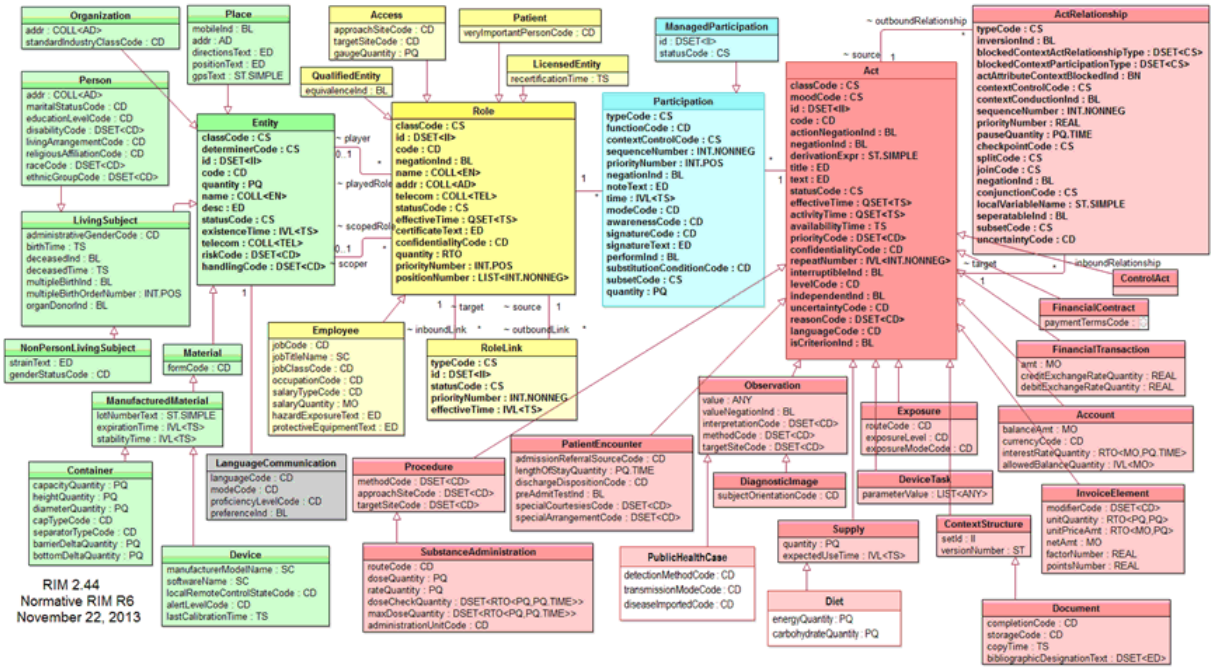


Figure 5 Normative content of the RIM [74].

6.4.1.4. Clinical Document Architecture (CDA)

In 2000, HL7 released the first version of a new implementation of HL7 v3, called the Clinical Document Architecture. This new specification aimed at proposing an interoperable solution for exchanging clinical documents. On many aspects, documents are different from rows of data in a database. According to HL7, documents in this architecture must express six characteristics: Persistence, Stewardship, Potential for authentication, Context, Wholeness and Human readability [75]. As a document can contain large parts of free-text data, the CDA is designed to be able to encapsulate the content of a document in an HL7 compliant structure with three different structure levels of increasing complexity [11]. The level 1 is only an XML header that contains the document's metadata and a body which contains the document itself. CDA level 2 contains the header and a body composed of sections. Each section covers a part of the content of the document and contains specific metadata. The third level, the most complex, has each section linked to a set of entries, called Clinical Statements, which represents the content of the section in a fully structured manner. Clinical Statements can be Observations, Procedures, etc. They can be linked to other nested entries and coded in any external encoding system required [76]. In the years following its release, CDA was widely adopted as a standard to store and exchange documents, with national implementations in the United Kingdom [77], Australia [78] and the United States through the Continuity of Care Document specification [79,80]. This wide adoption is partly due to the fact that the only mandatory part of a CDA document is the narrative part, allowing for early adoption even if structured data are unavailable in the system [81].

6.4.1.5. Fast Healthcare Interoperability Resource (FHIR)

In 2009, after more than 10 years since the beginning of the development of the third version of HL7, it appeared that it did not meet its goal of becoming the successor of HL7 v2. The broad RIM specification, while being successful in reducing the inconsistencies of v2, proved to be too complex for easy implementation. The investment required to successfully deploy HL7 v3 was only possible in national programs and the required knowledge of the specification was only present in people participating to the specification design [82]. Moreover, the specification was criticized for the design of the RIM that could be interpreted as incoherent from a semantic point of view [83].

The same year, a working group was created to explore new paths for the future of HL7. In August 2011, a proposal was published on the blog of Grahame Grieve, head of the working group and future founder of HL7 FHIR [84]. FHIR was first released as a Draft Standard for Trial Use in February 2014 and is now in its fourth release. FHIR proposes a new approach for interoperability. Since the beginning, it aims at simplifying implementations while preserving data integrity [85]. It is based on a given set of resources that define all exchangeable content. Every resource contains the same metadata and a hierarchical set of elements. The resources can be represented in multiple formats such as JSON, XML, UML or RDF languages like Turtle. Like in HL7 v3, it is possible to define FHIR profiles to constrain the usage of a resources.

On the implementation side, FHIR draws its inspiration on Application Programming Interfaces (APIs) commonly found on the web. More specifically, it is based on a Representational State Transfer API, also called RESTful API, meaning that every resource in FHIR has a predictable address. They can be accessed and managed through a set of HTTP services. The fact that FHIR is designed as a RESTful API means that a FHIR server is an external interface accessible by clients. Therefore, the internal design behind the API, the type of architecture, the databases or the data model are irrelevant. This results in the possibility to implement it on legacy systems with limited cost. Since its first release, FHIR has been built in contact with the community through regular connectathons [86]. Those events usually last two days and aim at connecting systems together.

Since its release, it has been widely adopted. Large healthcare entities in the United States, such as Medicare and the Veterans Associations, adopted it as their interoperability layer to give patients access to their data [8].

6.4.1.6. Substitutable Medical Applications and Reusable Technologies (SMART) on FHIR

In 2010, an interoperability project began in the United States with the goal of developing a platform that would allow medical application to be developed once and run in multiple facilities, providing they implemented the platform. In 2013, this initiative, called Substitutable Medical Applications and Reusable Technologies (SMART), decided to focus on the FHIR standard to achieve its goal. Among other features, SMART on FHIR allows authentication and authorization of users to access resources [87].

6.4.2. Digital Imaging and Communications in Medicine (DICOM)

DICOM was created in 1983 by the American College of Radiology and the National Electrical Manufacturers Association. Although not called DICOM at that time, it was an early effort to bring technical interoperability to digital imaging. The overall goal was to make medical imaging (Magnetic Resonance Imaging, Computed tomography, X-rays, etc.) independent of the manufacturer of the device that produced the image. This standard was an impressive effort at a time where no standards existed in the rest of the healthcare domain. The last iteration of the standard DICOM 3.0 release in 1993 is still the basis of most -if not all- Picture Archiving and Communication Systems (PACS) [88]. It allows production, storage and sharing of digital images along with their metadata. All current medical imaging devices produce DICOM formatted images and thanks to the backward compatibility of the standards with previous versions, it is possible to connect old devices in modern PACS with the necessary tweaking [89]. DICOM is one of the most striking examples of a successful standard introduction and adoption.

6.4.3. OMOP Common Data Model (CDM)

The OMOP CDM is patient centric and focused on drugs and their effects. The proposition of the OMOP CDM is different from HL7 standards. The overall goal of the OHDSI project is to give access to the common medical knowledge of multiple centers from a single-entry point and propose a galaxy of

software services to analyze those data. While its main focus is not interoperability, to provide such multicentric analytics, it had to include a specific approach for it. OHDSI proposes that each participating center maps its information into the OMOP CDM and performs an ETL procedure to load the data. Once in the common data model, it is possible to benefit from the multiple tools developed by OHDSI [90]. ATLAS is defined as a unified interface to patient level data and analytics. It allows cohort definition and analytics. The Health Analytics Data-to-Evidence Suite (HADES) is a set of open source R packages for large scale analytics [91]. Each of those software is open source and can be installed and used on an OMOP CDM database.

6.4.4. OpenEHR

The openEHR project was born in 1998 from the Good European Health Record Project. It is currently managed by openEHR international, a nonprofit organization created in 2003 by the openEHR Foundation. OpenEHR defines itself as a "technology for e-health, consisting of open specifications, clinical models and software that can be used to create standards and build information and interoperability solutions for healthcare." [92] All of its releases are open source. It proposes a method for designing and implementing health information systems. It is based on the idea that single level approaches creating entities directly in the system's data model have too much shortcomings and that a two-level approach (also known as "dual model approach") is both more robust and flexible [93]. OpenEHR proposes a stable reference model which will be the first level of the architecture. All data are stored in the reference model. Then, small domain models aimed at a specific piece of information and called archetypes can be defined and will constrain the reference model. For example, the heartbeat archetype contains information about the observation itself (rate, regularity, clinical description, etc.), the protocol (method, body site, device, etc.) and other elements. Figure 6 displays the Blood Pressure archetype:

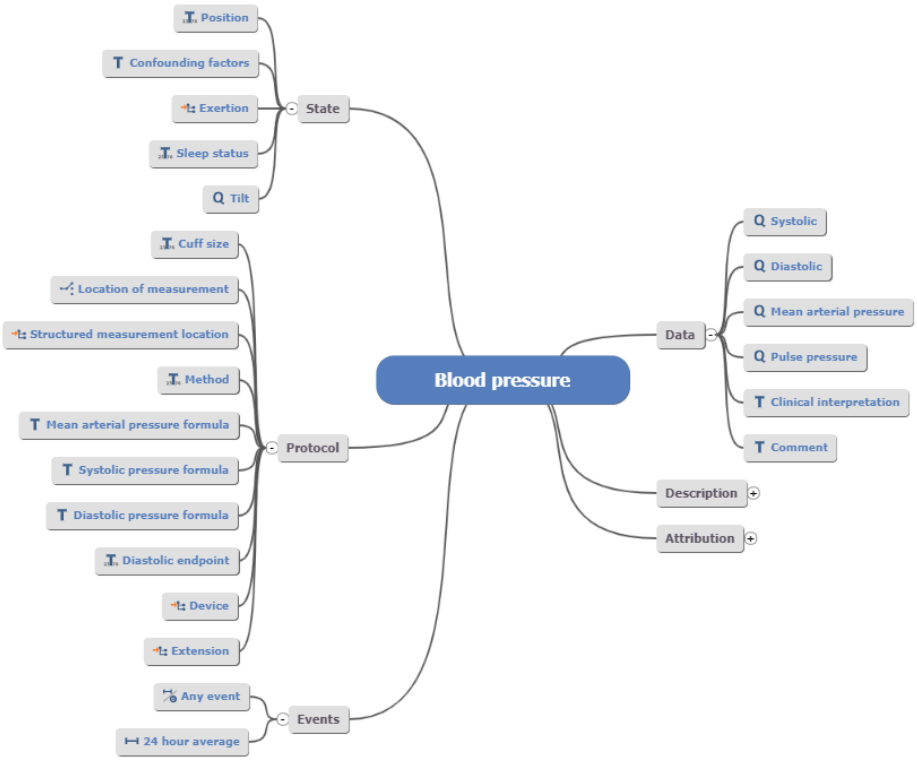


Figure 6 Blood pressure archetype as available in the openEHR Clinical Knowledge Manager [94].

Elements in archetypes can be linked to terminologies or classification systems. The archetypes are defined with an abstract syntax called Archetype Definition Language (ADL) and can be managed using different open source software published by openEHR International. They can be combined and constraint in templates that will be used to represent more complex sets of clinical information, such as health risk assessment, triage assessment, etc. The openEHR architecture does not enforce any controlled vocabularies. By providing mechanisms to design new archetypes and templates for everyone, there is a separation of the conceptual model which is represented by the archetype layer and can change continuously and the reference model that will stay the same and should contain a "relatively small number of non-volatile concepts" [93]. Figure 7 expands on the blood pressure archetype by displaying the link between elements of the archetype and SNOMED CT concepts.

Systolic Q Quantity Optional [**SNOMED-CT(2003)**::271649006 | Systolic blood pressure (observable entity)]	Peak systemic arterial blood pressure - measured in systolic or contraction phase of the heart cycle.	Property: Pressure Units: 0.0..<1000.0 mm[Hg] Limit decimal places: 0
Diastolic Q Quantity Optional [**SNOMED-CT(2003)**::271650006 | Diastolic arterial pressure]	Minimum systemic arterial blood pressure - measured in the diastolic or relaxation phase of the heart cycle.	Property: Pressure Units: 0.0..<1000.0 mm[Hg] Limit decimal places: 0
Mean arterial pressure Q Quantity Optional	The average arterial pressure that occurs over the entire course of the heart contraction and relaxation cycle.	Property: Pressure Units: 0.0..<1000.0 mm[Hg] Limit decimal places: 0
Pulse pressure Q Quantity Optional	The difference between the systolic and diastolic pressure.	Property: Pressure Units: 0.0..<1000.0 mm[Hg] Limit decimal places: 0
Clinical interpretation T Text Optional	Single word, phrase or brief description that represents the clinical meaning and significance of the blood pressure measurement.	
Comment T Text Optional	Additional narrative about the measurement, not captured in other fields.	

Figure 7 Data definition of the Blood pressure archetype with terminology mappings as available in the openEHR Clinical Knowledge Manager [94].

Archetypes and templates can be modeled using the Clinical Knowledge Manager (CKM) [94], an online and local tool that unifies more than 2600 users around more than 80 projects and incubators. Archetypes and templates defined by the users can be made available for everyone to reuse and modify. At the time of writing, 983 active archetypes and 166 active templates are available on CKM.

6.4.5. ISO 13606-1

The ISO 13606 standard created by the European Committee for Standardization and first published in 2008 has a similar approach to openEHR. Aimed at proposing a stable architecture for electronic health records, it is based on a dual model architecture by proposing a reference model as the building blocks of archetypes that will be meaningful combinations of those blocks [52]. In 2019, ISO released a new version of its standard, ISO 13606:2019 [95], based on the same principles. Archetypes can be described using the same ADL as openEHR. The reference model defined by ISO 13606-1 contains six clinical classes: folder, composition, section, entry, cluster and element (Figure 8). Those classes can be instantiated and linked to archetypes.

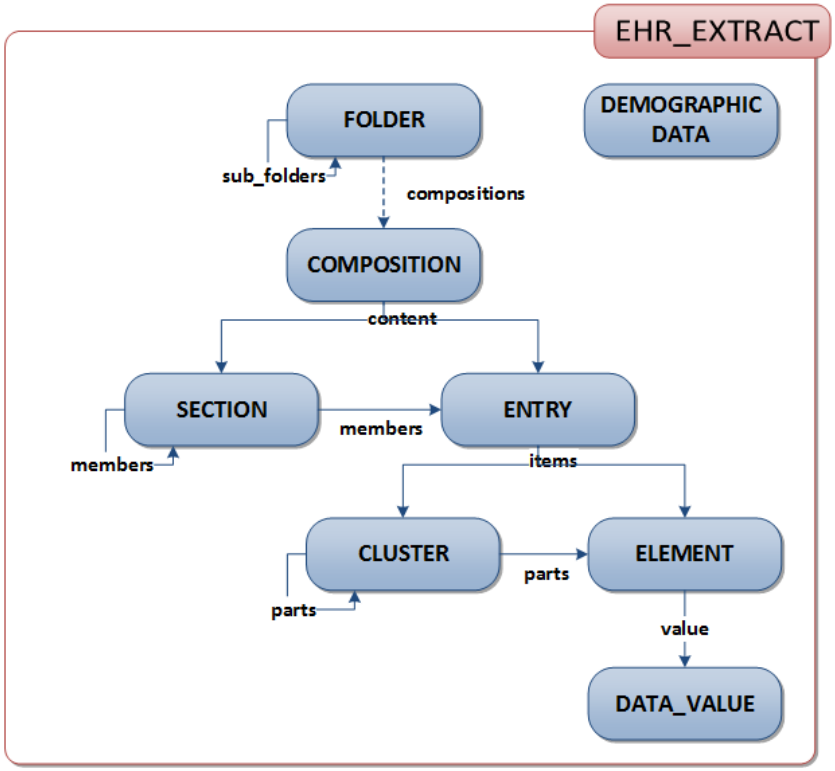


Figure 8: Classes of ISO 12606-1 reference model [52]

6.4.6. CDISC Operational Data Model (CDISC ODM)

Although not specifically aimed at interoperability for healthcare data, CDISC ODM needs to be presented as part of this landscape. First released for review in 1999, CDISC ODM is an XML-based data exchange standard designed to exchange information in Case Report Forms and their metadata [96]. Data in CDISC ODM is broken down into three blocks: study, administrative and clinical. Study data describe global variables, definitions as well as every case report form and every field. Administrative cover data about users of the system, investigators, centers included, etc. Finally, clinical data will store the actual instantiation of the data.

6.5. Controlled vocabularies

Semantic interoperability relates to the ability to use exchanged data meaningfully. As in every other field in which more than one person need to work together, people are sharing data, orally, or in written language. We usually do not refer to semantic interoperability when a clinician is presenting a case to its team, but it is a perfect display of what it must be accomplished: the transfer of information that will influence actions of people who did not gather it themselves. The difference in the medical informatics domain is the addition of computers to the equation (the word “computer” designates any computerized device or server that is processing information in a computerized manner). Medical informatics relies on the belief that there is a gain to obtain using computers for tasks that usually required people, pen and paper. Therefore, with this new actor generating, processing and sharing information, it must be ensured that the process keeps the meaning of the information unmodified and that it can be used by a machine that does not comprehend it. Grahame Grieve, the project lead of HL7 FHIR, wrote that while developing a standard such as FHIR, what is pursued is not semantic interoperability but “un-semantic” interoperability [97]. Indeed, technical standards are built in the aim to process information automatically without any interest for its meaning but keeping it intact. Therefore, while it is currently illusory to build computers that understand the meaning of a health

record, it is crucial to ensure that the semantics of the data is sufficiently consistent and unambiguous so that it can sustain processing, comparison and transfer without being altered.

The barriers to semantic interoperability are various. Some relate to person-to-person interactions and some to person-to-computer or computer-to-computer interactions. When free-text data is transmitted from a clinician to another through a computerized process, the major barrier preventing data from being usable is the language. Clinicians need to share the same language (e.g. English, French, etc.) to understand each other. They also need to share some knowledge about the topic or medical specialty to which the data relates. In medical practice, words and abbreviations can have multiple meanings -as it has been shown by the study of polysemy in large classifications such as the Unified Medical Language System (UMLS)- and their usage can hinder semantic interoperability [98]. Those barriers are not specific to computers and existed before them.

Moreover, when data must be processed by the computer itself, for example to trigger an alert, new challenges arise. First, data should be structured in order to be processed (as stated in the structured interoperability layer), but structure does not convey meaning. To bind data to a stable meaning, a common approach is to use a controlled vocabulary. By binding data to specific labels or codes in a classification, semantics is transferred from the data itself to an external system specifically targeted at organizing meaning (Figure 9).

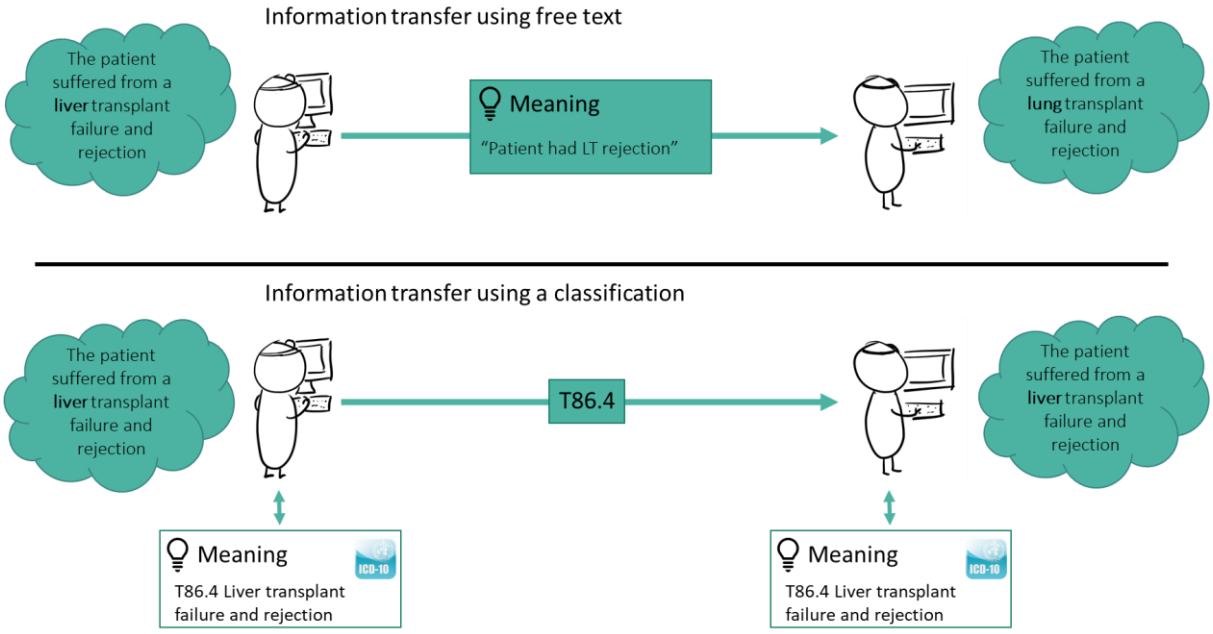


Figure 9 Effect of encoding the meaning of data into a controlled vocabulary.

6.5.1. The uncontrolled vocabulary of controlled vocabularies

It is interesting to note that the vocabulary used to describe what can be generally addressed as controlled vocabularies can be unclear and require definitions. This issue is best exemplified by SNOMED CT being named a nomenclature in its acronym, referred as a terminology in its documentation and often called an ontology in articles about it [99,100].

This problem can be approached by classifying controlled vocabularies according to their general goal or by their coverage. Controlled vocabularies can be built to gather every existing expression and to group them according to their meaning. This approach produces what is commonly called a thesaurus, or a dictionary [101], the most known thesaurus in the biomedical domain being the UMLS Metathesaurus which regroups more than 200 controlled vocabularies in a unified list [102].

Classifications, as defined by the Cambridge dictionary are “the division of things into groups by type” [103]. This division is materialized as a finite set of boxes in which every covered element can be placed. Classifications define the exhaustive list of what can be used when speaking about a domain. When using ICD, every disease and mortality cause should be expressed using a code contained in it and no meaning can be expressed if it is not contained in a code. Finally, combinatorial ontologies (or concept systems) provide a list of concepts, and rules describing how to combine them to express meaning. The concepts composing them include descriptive relationships with other concepts that define their essence [104–106]. What such a system contains is not a closed list of expressible meanings, but a list of the possible words available in a language to build sentences that will convey a much wider set of meanings. Although referred frequently as a terminology, SNOMED CT is currently the most prominent member of medical combinatorial ontologies.

Taking another angle, each of these types of controlled vocabularies can have different coverage. They can target specific activities such as ICD for diseases and mortality causes, the NANDA International Nursing Diagnoses [107] for nursing diagnosis or the International Primary Care Classification [108] for consultations motives. Controlled vocabularies can also target a domain, like the International Classification of Diseases for Oncology, third edition (ICD-O-3) which aims at representing the site and the histology of neoplasms [109], the Anatomical Therapeutic Chemical Classification System which is used to classify drugs or the Diagnostic and Statistical Manual of Mental Disorders published by the American Psychiatric association and aimed at mental disorders [110]. Finally, large transdomain controlled vocabularies such as SNOMED CT, LOINC or the UMLS Metathesaurus cover even broader parts of the biomedical domain.

It is worth noting that hyper specific clinical scales and scores, such as the Glasgow Coma Scale (GCS) [111] or the Apgar score [112], are also numerous in medicine. Scores and scales are an integral part of medical practice and clinical data. They are used every day by clinicians and are so pervasive that applications exist to provide easy access to them [113]. However, those small classification systems are not the focus of this work as they can be considered as small sets of specialized concepts and are sometimes already included in large controlled vocabularies (for example, Apgar score is included in SNOMED CT and LOINC).

Other words can be encountered in the literature about controlled vocabularies. They will be briefly defined for exhaustiveness but do not enter in the scope of this work. A taxonomy can be considered as a synonym of classification but use of this term in healthcare is less frequent and usually bond to the field of biology [114–116]. A nomenclature represents a list of words commonly recognized and validated in a specific domain to name elements of this domain. Finally, the most frequent term used when reading about controlled vocabularies in healthcare is “terminology”. The meaning of terminology is close to a nomenclature. It is a collection of special words or expressions used in relation to a particular subject or activity [117]. This profusion of terms and their sometimes fuzzy use underscore the challenges of semantic interoperability. Throughout this work, the term “controlled vocabulary” is used to cover all types described before and, for coherence with other works, the word “terminology” is used for SNOMED CT.

6.5.2. Usage of controlled vocabularies in healthcare

While there exist probably several hundreds of controlled vocabularies in the biomedical domain, it is difficult to find an exhaustive list. However, some insights can be drawn from organizations or projects aiming at regrouping them. The UMLS Metathesaurus (described in 6.5.7) contains more than 200 controlled vocabularies. The Health Terminology/Ontology Portal [118,119], a French platform that allows search of concepts in multiple languages, covers 70 terminologies and ontologies. Finally, the

HIMSS terminology standards list 11 major common terminology standards on its website, including SNOMED CT, LOINC and ICD-10 which will be described in section 6.5 [120].

Integration of controlled vocabularies vary in technical standard specifications. Standards published by HL7, CDISC, OMOP or openEHR include different solutions to encode or represent meaning. FHIR resources can contain elements with the data type "CodeableConcept" which specifies that this element will be represented by a code in a value set of codes, internal or from an external classification [121]. CDISC's Controlled Terminology is mandatory to encode CDISC compliant datasets [65]. OpenEHR allows to integrate external controlled vocabularies to archetypes or templates to define value sets or encode a concept without many restrictions. Finally, the Standardized Vocabularies released by OMOP are mandatory to encode data in the OMOP CDM. Those vocabularies can be created internally or adopted from existing standards such as SNOMED CT [122].

However, each of those solutions have limitations: either inconsistencies highlighted in the design of FHIR resources [123,124], the mandatory data model at the center of openEHR, or the way OHDSI integrates external vocabularies by assigning new unique identifiers to concepts and therefore imposing a controlled terminology instead of proposing the liberty of using existing ones.

6.5.3. Properties of controlled vocabularies

The properties that a controlled vocabulary should express to be widely used and future proof have been summarized by Cimino et al. in 1998 [125]. Such a vocabulary should be concept oriented; it should provide concept permanence to avoid that data encoded in the past becomes unusable in the future. This forbids the use of residual aggregation since a concept that is defined by exclusion such as "pneumonia not elsewhere classified" is bound to a semantic drift as new knowledge reduces its scope by adding new types of pneumonia. It should be built around a polyhierarchy to allow efficient navigation and avoid choices such as classifying the concept "pneumonia" in the lung disease category over the infectious disease category. In the same aim, the identifier of the concept should carry no meaning to avoid modification when new knowledge appears or limitation in the addition of new codes. Indeed, many classification systems, such as ICD or Clinical Terms Version 3, use position-dependent codes where the structure of the code specifies part of its meaning and can limit evolution possibilities [126]. It should include formal definitions materialized as typed relationships between concepts and all granularities should coexist in it. Finally, it should be updated regularly and in a consistent manner without creating any break in compatibility. The SNOMED CT documentation includes a page where desiderata of Cimino are listed with an explanation of how SNOMED CT meets each one of those requirements [127].

6.5.4. International Statistical Classification of Diseases and Related Health Problems (ICD)

While the first attempt to classify diseases is traced back in the 18th century, the precursor of ICD was born in 1853 with the International List of Causes of Death. This classification created after the first International Statistical Congress aimed at a statistical study of causes of death around the world [128]. This list first released in 1893 was adopted by multiple countries including the United States and Canada. It was updated multiple times during the first half of the 20th century. Its sixth revision in 1949 was marked by the transfer of the custody of the list to the World Health Organization (WHO) and its modification into the International Statistical Classification of Diseases which included new elements for the coding of morbidity data. In 1989, the tenth revision of the classification, ICD-10, was released and is still to this day the most used version of ICD with 117 countries using it for reporting mortality data [129]. ICD-11, for which the work started more than a decade ago, was released for the members of WHO in May 2019 and should start to be used for health reporting in January 2022 [130–132].

The ICD structure changed across its revisions and will change again with its eleventh revision. ICD-10 is composed of twenty-one chapters (from A00 to Z99). Every ICD-10 code is composed of a letter followed by up to three digits. The U chapter is reserved for the provisional assignment of new codes as new diseases are discovered. Starting with the tenth revision, ICD is updated annually by WHO to include new diseases or refine existing codes.

Multiple countries have developed their own adaption of the ICD classification. Examples of such local adaptations are the German modification, ICD-10 GM [133], the Australian Modification ICD-10 AM [134] or the United States clinical modification ICD-10 CM [135]. The latter succeeded in 2015 to ICD-9 CM as the mandatory classification system for mortality and morbidity [136,137].

As stated in its name, ICD is a statistical classification. Its role is to provide a common language for reporting and monitoring diseases. Its primary use is summarizing data about mortality and morbidity. ICD is best used when aggregating data from a large set of patients in order to gain knowledge about the evolution of a disease or the evolutions of the causes of death in a population. Due to its ability to classify and express what disease or pathology the patient suffered from with common codes across a country, it has been largely used for billing and reimbursement. For example, in the United States, ICD codes are required in health care claims. In Switzerland, a translated version of ICD-10 GM is the basis for billing of inpatient stays. Each stay is coded using ICD codes that must be assigned according to a rulebook edited by the public health administration [138]. While coding was first made by clinicians in Switzerland, it is currently performed by coding experts that follow a specific training and must hold a certification. The ICD codes assigned to an inpatient stay are combined with procedure codes from the "Classification Suisse des interventions chirurgicales" (CHOP) and additional data. That information is then grouped and mapped to Diagnoses Related Groups (DRGs), a system created in the United States to classify inpatient stays according to their cost in resources.

ICD codes are broadly used and available healthcare data. Therefore, the incentive to use them for research is high. However, it has been shown that they are not a reliable source of information for research or clinical care as they are biased by the rules applied during coding.

6.5.5. Logical Observation Identifiers, Names, and Codes (LOINC)

LOINC has been created in 1994 by the Regenstrief Institute, a United States non-profit organization [9,139]. Its goal was to solve the problem of internal, idiosyncratic coding of observations and laboratory values [140]. It is designed to provide a unique identifier for the observation and not its result. LOINC describes the question that is asked; for example "glycemia level in plasma", and not the answer that would be "4.2mmol/L". This distinction is crucial as terminologies such as SNOMED CT can express both the question (Glucose measurement, blood (procedure)) and its answer (Random blood sugar raised (finding)).

LOINC is mainly composed of two large categories of codes, clinical terms and laboratory terms. The first includes all types of observations that can be made about a patient such as "Left ventricular Cardiac index by US" or "History of Kidney disorders". It also includes a document ontology which describes types of documents in compliance with HL7 CDA and codes for radiologic observations, patient-reported outcomes measures or nursing assessments [139]. The laboratory terms category covers laboratory tests such as "Sodium [Moles/volume] in Blood". Panels are elements in LOINC that regroup multiple observations. "Sodium and Potassium panel [Moles/volume] – Blood" regroups two laboratory exams. The documents in the document ontology are also panels and are linked to a set of codes that are recommended or optional to be part of the document.

The concept model of LOINC defines six mandatory properties for a code: Component or Analyte describes what is the focus of the observation (e.g. glucose), Property defines what is observed about the component (e.g. substance concentration), Time-Aspect describes if the observation is over time or punctual, System is related to the sample or system in which the observation is made (e.g. blood), Scale differentiates if the observation is qualitative or quantitative and, if relevant, the Method used for the observation is filled (e.g. Glucometer).

LOINC is updated twice a year and is distributed freely. The current 2.69 version contains 94,895 codes [140]. Since its creation, it has been widely adopted for encoding laboratory measures around the world with 88,647 users representing 176 countries in 2019 with 20 translations [141]. However, issues have been raised on its usage, specifically about the comparability of aggregated results due to different devices and references ranges encoded with the same LOINC codes. The task of mapping internal coding to LOINC codes is time consuming, requires interdisciplinary teams and can result in different codes for the same analysis depending on the person performing the mapping. Finally, the unit of the laboratory test can vary and the absence of a systematic unit conversion system inside LOINC can hinder interoperability [142–144].

6.5.6. Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)

SNOMED CT was first released in 2002 from the merger of two existing healthcare classifications, SNOMED Reference Terminology (SNOMED RT) and Clinical Terms Version 3 (CTV3) [145].

SNOMED RT was the latest version of a classification created in 1965 by the College of American Pathologists and called Systematized Nomenclature of Pathology (SNOP). In 1975, its scope was broadened to include concepts relative to medicine and therefore was renamed “Systematized Nomenclature of Medicine (SNOMED)”. Three different versions of SNOMED were released until the end of the nineties when the building of SNOMED RT began. While the third version of SNOMED already included 10 different taxonomies such as diagnoses, procedures, topography or living organisms, and a mechanism to combine concepts to create a composed statement, it lacked a proper syntax to do so and typed relationships between concepts [146]. SNOMED RT aimed to fill this gap by adding formal description logic-based definitions to the system. It was released in early 2001 when the development of SNOMED CT had already begun. Therefore, it was never updated and represented more as a transition step to SNOMED CT.

CTV3, on the other hand, was created in 1985 in the United Kingdom by Dr James Read as a set of clinical terms for use in EHRs in primary care settings. The so-called Read Codes were updated and renamed Clinical Terms when their license was purchased by the United Kingdom government and the updates were put under the responsibility of the National Health Service [147,148]. The third version of CTV3 also included features to post-coordinate codes and further specify a concept.

In 1999, the College of American Pathologist and the National Health Service agreed to merge CTV3 with SNOMED RT. After mapping common concepts and working on concept modelling, this merger gave birth to SNOMED CT [147].

SNOMED CT is currently considered as the most comprehensive, multilingual clinical healthcare terminology in the world [99]. It contains more than 350,000 concepts and a million relationships. It is maintained and published by SNOMED International [149], a non-profit organization composed of 39 member countries [150]. Each member country contributes a license fee to SNOMED International related to its gross domestic product. The license gives access to SNOMED CT for the entire country as well as the possibility to contribute, suggest modifications, define specific reference sets for local usage and benefit from training provided by SNOMED International.

SNOMED CT is organized in three main components: Concepts, Descriptions and Relationships [99]. Concepts are the only component in which the meaning resides. Descriptions are a natural language representation of the concept. Relationships are specific concepts used to link concepts together. Each concept has a unique formal logic-based definition that is materialized as relationships to other concepts and complies to rules defined in the SNOMED CT Concept model (Figure 10) [151]. As stated before, while named a "systematized nomenclature", SNOMED International designates SNOMED CT as a terminology. However, from its concept-based structure including typed relationships, synonyms and preferred term, it is closer to a formal ontology or to a group of ontologies (Figure 11). However, it cannot be defined completely as an ontology as issues with its concept model have been raised, such as the possibility to violate axioms of the SNOMED CT concept model based on the intuitive meaning of synonyms or the usage of relationship groups [152–154].

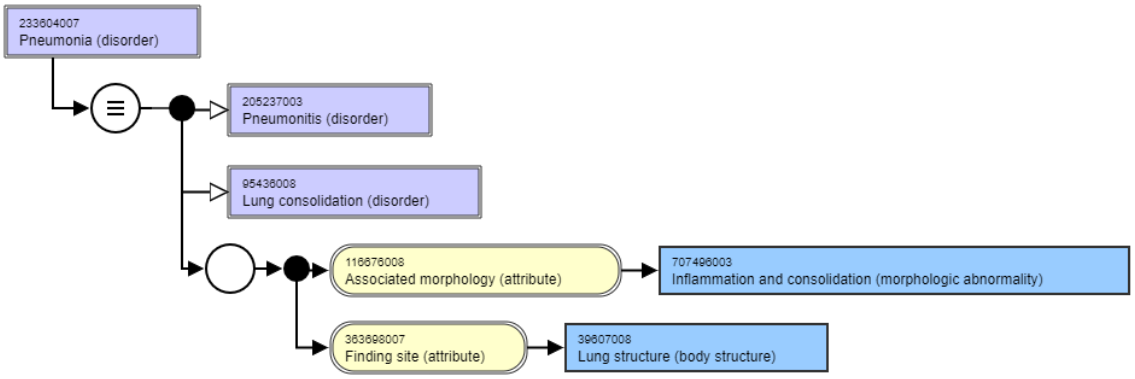


Figure 10 The diagram of the concept Pneumonia in SNOMED CT [155]..

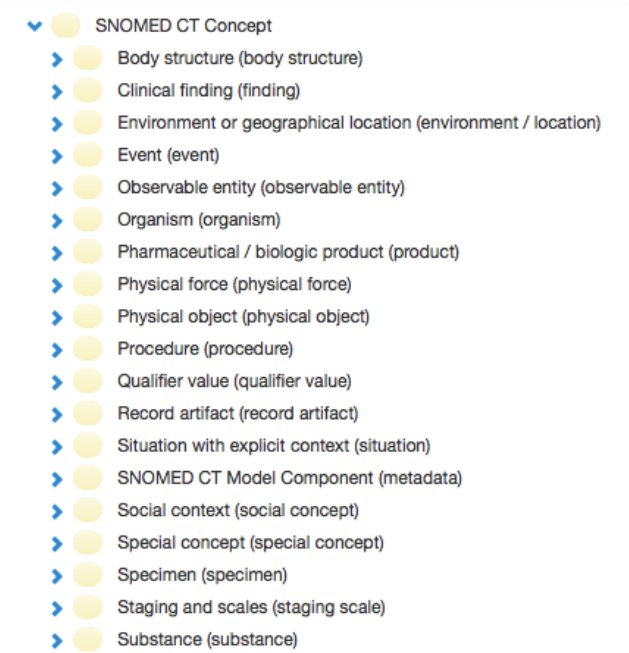


Figure 11 The 19 top hierarchies of SNOMED CT [156].

The highly connected structure contained in SNOMED CT allows complex queries under the form of expression constraint queries. This formalism can be used to retrieve concepts following a set of constraints that can be simple like "being a child of a specific concept" or more complex like "having a specific relationship with a set of other concepts". Information retrieval through expression constraint

language cannot be accomplished with other health terminologies due to their lack of formal logic-based definition [157].

SNOMED CT allows post-coordination. When a concept is not present as a single code in the terminology, it can be created by composing existing codes and relations. By following the compositional grammar edited by SNOMED International, anyone can create new post-coordinated concepts [158]. This has multiple advantages. It avoids the combinatorial explosion that comes with every terminology aiming at exhaustiveness and it allows users to accurately encode the information with the needed granularity. This is the reason why SNOMED CT is described as the universal language of healthcare [159].

With 39 member countries and more than 5'000 affiliate licenses distributed [150], SNOMED CT is widely used in both healthcare and research settings. The use cases are various and can range from purely theoretical analysis of the properties of SNOMED CT [154,160] to pragmatic usage of a subset of concepts to encode a specific information [161,162]. Among those use cases, the ability of SNOMED CT to be used as a language to represent complex concepts expressed in unstructured data is very promising.

6.5.7. Unified Medical Language System (UMLS)

The building of the UMLS began in 1986 by the National Library of Medicine with the goal of improving the capability of computer programs to behave as if they understood the biomedical languages [102,163]. This national and international long-term effort resulted in the creation of three knowledge sources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon & Lexical Tools (Figure 12).

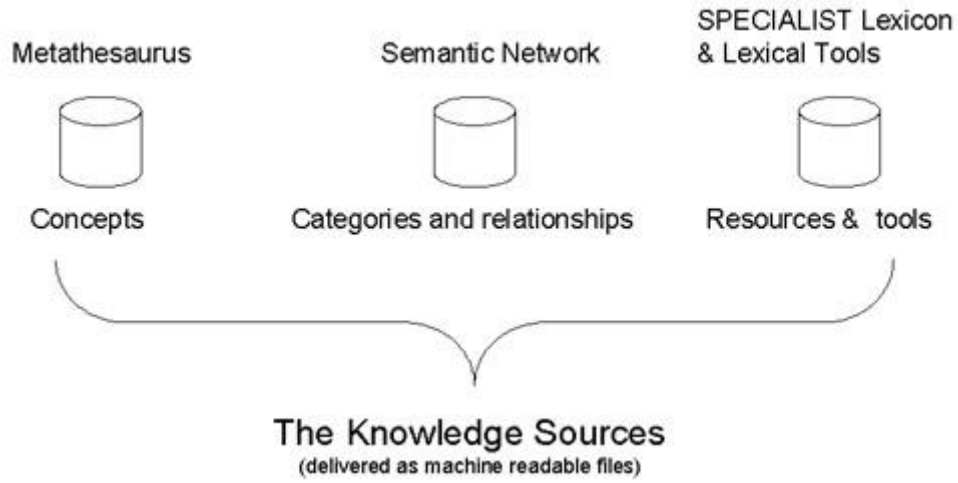


Figure 12 Knowledge sources of the UMLS [102].

The Metathesaurus is a unification of more than two-hundred controlled vocabularies relevant to the biomedical domains into more than one million biomedical concepts and five million terms. Each term is linked to a unique identifier specific to the UMLS. Although it includes many classifications and terminologies already widely used in the clinical domain and research, it does not aim to replace them. However, it can be used to organize concepts from multiple sources and facilitate the creation of mappings between them. Examples of included vocabularies are SNOMED CT, LOINC, MeSH or ICD-10-CM. The Semantic Network is composed of 133 semantic types and 64 relationships. They are used to categorize each concept in the Metathesaurus, therefore reducing the complexity of navigating through many concepts. Finally, the SPECIALIST Lexicon is an English lexicon of more than 200,000 terms with syntactic, morphological and orthographic information aimed at Natural Language

Processing (NLP) applications. Lexical tools such as a lexical variant generator or a normalized string generator are also available.

6.6. Projects related to clinical data sharing for research

6.6.1. I2b2 tranSMART foundation

I2b2 is a clinical research platform aiming at combining data emerging from research (biology) with clinical data coming from patient records (bedside). It has been created in the framework of the NIH Roadmap National Centers for Biomedical Computing initiative (46, 47). Its first version was released in 2007. This platform includes a database model, an application layer and core APIs. Each module of the application is called a cell and is integrated in a hive. The cells communicate with each other via web services.

At the center of the system is the data repository cell, also called Clinical Research Chart, which contains phenotypic and genomic data. It is accessed by most of the other cells to produce analytics. Data in i2b2 is stored using a star schema. This data model is built around a central table called observation facts and a finite set of dimensions linked to it (Figure 13). Each piece of data is a fact linked to a patient, an encounter, a provider and a concept. Even though the model has changed along the years, it remains based on the same axioms. The concept dimension is especially important because it holds the link between the data and its semantics. It is designed to contain any classification or coding system needed to encode the data. Therefore, each observation fact can be linked to a set of concepts. It is important to note that no such link is available for the other dimensions and therefore their semantic needs to be documented elsewhere. The modular design allows i2b2 to be improved by new cells when needed. Cells have been developed both by the i2b2 foundation and by the research community for various tasks such as managing ontologies and patient identities [166,167] or providing FHIR compatibility [87,168]. I2b2 also includes a cell dedicated to NLP. The Health Information Text Extraction (HITEx) tool combines a set of language processing modules that can be selected to build an NLP pipeline and perform extraction from narrative documents. Its components are derived from the open source General Architecture for Text Engineering (GATE) [169] and are integrated in the NLP cell of the i2b2 Hive [170]. This cell can perform various information extraction tasks and contains a UMLS concept mapper that can find and link concepts to text. The NLP cell has been used and evaluated to extract principal diagnosis, co-morbidities and smoking status [170–172].

In 2017, the i2b2 foundation completed its merger with the tranSMART foundation. TranSMART is described as knowledge management and high content analytics platform [26]. It is aimed at providing reusability to research data. It uses parts of the i2b2 platform and provides search capabilities based on Apache Lucene Solr [173], a well-known search engine. TranSMART was first released in 2010 and was put in open source in 2012. The i2b2TranSMART foundation currently manages multiple software with the two core products being i2b2 and tranSMART. Since the merge, the integration and compatibility of the two platforms have been improved and different use cases using the two platforms are described and additional software proposed by the foundation [174].

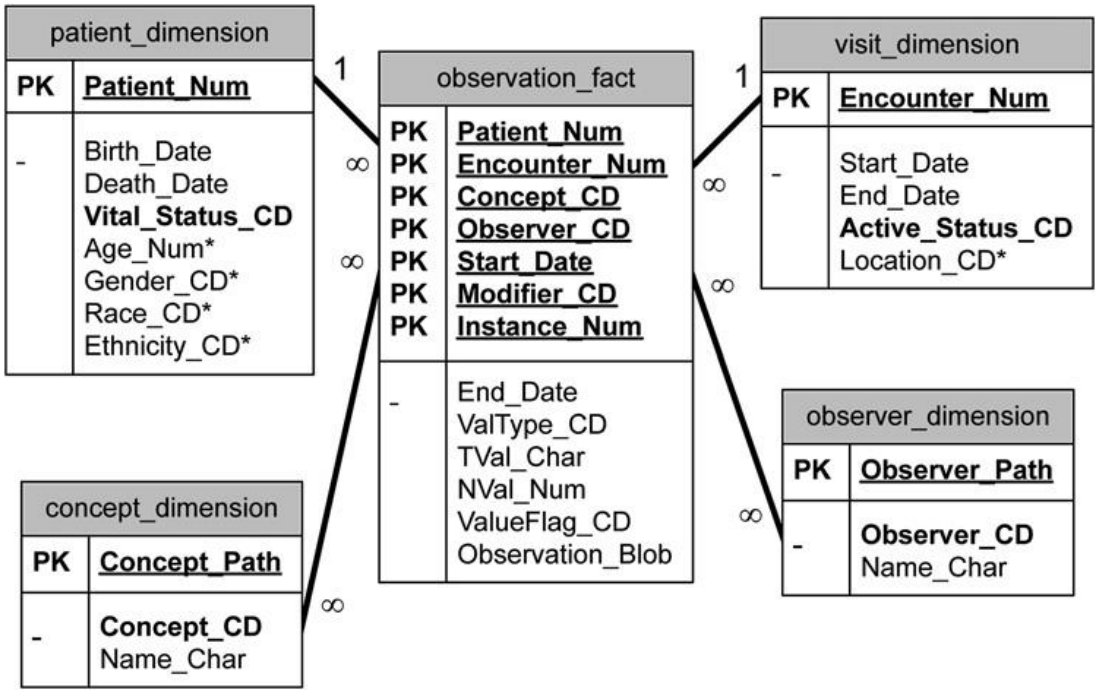


Figure 13 i2b2 star schema [165]

6.6.2. Strategic Health IT Advanced Research Projects consortium (SHARPn) project

The SHARPn project emerged from a set of projects called SHARP and piloted by the ONC in 2012. It is a framework aimed at secondary use of EHR data and is designed to be scalable and standard driven [175]. It can take as an input data in HL7 messages and normalize them using an Unstructured Information Management Architecture (UIMA) pipeline which is an architecture designed by IBM and maintained by the Apache Software Foundation. The pipeline includes different normalization steps depending on the type of data (HL7 messages, tabular data or free text). The data is then persisted in an SQL database according to a set of data models named Clinical Element Models (CEMs) [176]. CEMs are small data models designed according to a common abstract structure. They can be expanded or modified according to a specific need. Specifically, for the SHARPn project, 28 CEMs were used. They are like openEHR archetypes as they are small domain models and can be constraints and linked to terminology codes for semantic purposes. The goal of SHARPn is to propose a framework for high throughput phenotyping by rapid harmonization of various data from standardized HL7 messages to free text using NLP.

6.7. Natural Language Processing (NLP)

NLP can be defined as the field of computer sciences that takes natural language as an input. It is an active research branch in the biomedical field and has been broadly applied on scientific literature and clinical text [177] for automatic information extraction, automatic document classification and sentiment analysis [178–180]. NLP tools used in clinical text mining are often seen as a way to improve healthcare [181]. Scientific literature represents convenient datasets for NLP tools as these documents are freely available and usually well-structured [182–184]. However, NLP applications on clinical documents are less frequent. Among the reasons explaining this disparity, the limited access to corpora of clinical documents and the lack of publicly available annotated corpora are mentioned [185]. Limited access to data is partly due to the difficulty of defining the appropriate method to enhance the privacy of data as well as setting a threshold above which the data is shareable [186]. These barriers slow down the development of large-scale solutions for NLP and information extraction on clinical documents.

Moreover, while there exist freely available corpora of clinical documents (among which, MIMIC III [187] and CLEF [188]) written in English, such datasets in other languages are scarce.

One interesting application of NLP in connection with clinical data interoperability are the challenges organized in the framework of the i2b2 project. Those challenges were designed as shared tasks, meaning that each participating team was given the same annotated data set and their results were evaluated against each other. Those tasks aimed at tackling challenges met by the i2b2 projects. From 2006 to 2014, seven challenges were organized by the i2b2 foundation and produced numerous results for tasks such as de-identification [189], medication extraction [190], temporal relation extraction [191] etc. Since 2018, those challenges are organized by the National NLP Clinical Challenges (n2c2) under the stewardship of the Department of Biomedical Informatics at the Harvard Medical School [192]. The last challenge organized by n2c2 was focusing on medical concept normalization [193]. Those challenges are an important source of annotated datasets since all datasets used in challenges are available for research under registration. Beside those tasks, NLP can be used to provide automation of mandatory steps for interoperability such as structuring documents or mapping of text to controlled vocabularies [194].

6.8. Research questions

Despite a profusion of technical standards and controlled vocabularies and important investments, interoperability of clinical data remains a challenge. Several observations can be made on these barriers and limitations.

A standard can only bring interoperability if it is adopted. The wide implementation of DICOM in PACS or the broad usage of ICD for the billing of inpatient stays and for public health statistics are good examples of successful standards. Unfortunately, such unity in the adoption is not common in healthcare. The 2016 ONC report [21] mentioned that the reluctancy of the stakeholders to adopt standards that would support collaborative work and meaningful engagement from the patients resulted in poor adoption of standards or even information blocking, an issue still present in 2020 [24]. Elements such as calls for action by large EHR companies to hospitals, urging them to oppose to regulations promoting interoperability are a strong sign that the industry has a responsibility in the slow progress of semantic interoperability in healthcare [195,196].

Regardless of their type, standards are not neutral. The intention that drove their creation influences their structure and their adoption depends on this intention. The three main classes of the HL7 v3 model, Act, Role and Entity, are designed to be able to represent actions taken and their associated metadata describing what, where and to whom it happened. OMOP CDM is constructed to identify and evaluate associations between interventions and outcomes. While both approaches are person centered, harmonization between the two models have proven to be non-trivial, underscoring the fundamental differences of the two models [197]. Similarly, the resources proposed by FHIR were primarily targeted at EHR interoperability, OMOP CDM specifically targets clinical research and the CDISC ODM is the model used for submissions made to the FDA. Each of these standards is different and answers different needs but standards overlap to some extent. As for controlled vocabularies, ICD-10 is designed as a statistical classification of diseases, LOINC aims at representing observations and laboratory tests and SNOMED CT targets the complete medical domain. SNOMED CT contains a hierarchy named “disorders” that covers most of the elements of ICD-10, and LOINC includes concepts related to observations and laboratory values that also appear in SNOMED CT.

While there are overlaps in the coverage of major controlled vocabularies, it is not reasonable to enforce a single standard, whether technical or semantic, in every community and for all purposes. Clinical care is composed of many actors, roles, cultural habits and needs, and enforcing a unique

standard is neither possible nor desirable. As an example, the identification of a drug by its active substances may be meaningful for a clinician but classifying it by its therapeutic indications is preferable for a patient and logistic identification is needed to handle supplies and orders in the hospital's pharmacy. Each of these views on the drug concept are equally correct, needed and different. Moreover, if an international standard can be used to represent a drug dose, the analytical method of a laboratory test or the specific type of cells observed in a pathological specimen, concepts such as the expression of the complaint of a patient are strongly linked to his/her language and cultural background. Finally, controlled vocabularies have limitations. They can be too comprehensive for specific usages (the 71,000 LOINC codes can be useful for a large laboratory provider but are difficult to use for a researcher) or lack specific features (SNOMED CT is able to represent close to any medical concept, but lacks solution when expressing concrete values [198]). Concerning statistical classifications, such as ICD-10, that are widely used and available, they have shown their limits in terms of semantic representation outside of the billing process [42,43,45]. Therefore, it does not seem desirable or even applicable to enforce a single standard for every use case.

Finally, in a connected digitalized world, data must be shareable and understandable across domains and standards. Since it is not realistic to enforce a single technical standard and a single controlled vocabulary to represent, store and share clinical data, conversion mechanisms are needed. However, they would require mapping from each standard to every other one which is, to date, not done and would represent unreasonable maintenance costs.

> - Standards do not create interoperability; their adoption creates it.
> - Standards are not neutral; their specificities depend on the community that created them and their goal.
> - It is not possible to enforce a single standard in every community for every purpose because of the differences in needs, purposes and cultural habits of the communities that use them.
> - The multiplicity of standards requires many-to-many mappings from each standard to every other for the data to be shareable and understandable by everyone.

Textbox 1 Summary of the observations made about interoperability in healthcare.

To resolve this issue, three hypotheses are made.

6.8.1. Hypothesis 1

> SNOMED CT in conjunction with a limited number of other knowledge representations can be used as a formal interlingua to represent clinical information properly.

Most of the existing controlled vocabularies in healthcare, such as ICD-10 or LOINC, do not include mechanisms to combine their elements to create concepts absent from the classification. But SNOMED CT is published with a compositional grammar allowing the post-coordination of concepts. This highlights an important difference in the philosophy behind those systems. The list of ICD-10 codes can be considered as the list of "what is expressible in ICD-10". Therefore, any concept that is not present in the classification cannot be represented with it. This partly explains why residual aggregation is needed in ICD-10: to provide a way of expressing diseases not yet encoded in a proper ICD-10 code. In a compositional system, the list of its concepts is no more "what is expressible" but "the set of concepts that can be used and combined to express something in this classification". This simple perspective

shift differentiates SNOMED CT from the classification realm and brings it closer to a natural language. This is exemplified by the fact that it is possible to create a syntactically correct but totally nonsensical post-coordinated SNOMED CT sentence. As the famous sentence from Noam Chomsky, “Colorless green ideas sleep furiously”, created to highlight the difference between syntax and semantics, the syntactic correctness of a SNOMED CT sentence does not guarantee anything regarding its meaning [199]. Textbox 2 displays an attempt at representing Noam Chomsky’s sentence using the SNOMED CT compositional grammar.

> *247623004 |Exciting ideas (finding)|:*
>
> *{*
>
> *103366001 |With color (attribute)|:371246006 |Green color (qualifier value)|:*
>
> *{*
>
> *103366001 |With color (attribute)|:263716002 |Colorless (qualifier value)|*
>
> *},*
>
> *246090004 |Associated finding (attribute)|:248220008 |Asleep (finding)|,*
>
> *246090004 |Associated finding (attribute)|:75408008 |Feeling angry (finding)|*
>
> *}*

Textbox 2 SNOMED CT representation of the sentence: Colorless green ideas sleep furiously.

The coverage of SNOMED CT has been the focus of a growing number of publications [200,201], including a large European project named ASSESS CT and aimed at evaluating the terminology as a standard for semantic interoperability for European eHealth deployments [202]. This project stated in its final recommendations that the content coverage of SNOMED CT was superior to any other single terminology [203].

The complexity of semantic interoperability in healthcare can be summarized as the interface between two realms. On one side, the humans have a common understanding of health-related concepts. Those conceptual frames hold the meaning of things as understood by humans. When communicating with each other, humans create representations of those concepts using natural language. Natural languages are numerous and overlapping. General languages such as English or French are large distinct representational systems. But smaller, domain related languages can be defined such as the “medical language”, or even smaller jargons used in specific professions such as nurses, surgeons, etc. The language used by a hematologist and a laboratory worker will overlap but still be distinct in representing some concepts.

On the other side, computers do not strictly understand medical concepts but still manipulate and store them. Those machine-readable concepts are the pendant of the conceptual frames for humans. As for humans, multiple machine-readable languages are used to store and share concepts between machines. ICD-10, LOINC for large systems, and APGAR or GCS for smaller ones.

In this setting, it is hypothesized that it is possible to bridge human conceptual frames with machine readable concepts by using a restricted set of controlled vocabularies to represent health concepts. Figure 14 summarizes this approach.

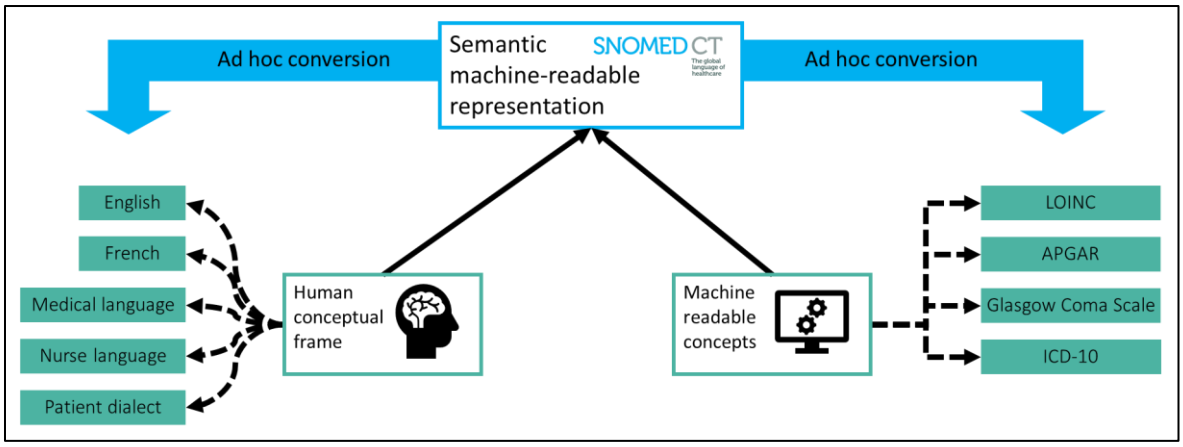


Figure 14 Schema of the semantic convergence proposed in hypothesis 1.

The follow-up of this first hypothesis is rooted in the observation explained before that no single standard should be enforced for all purposes. Therefore, once concepts are adequately represented in selected controlled vocabularies, the storage of the data should not be made by constraining the data into a data model. Instead, a descriptive formalism should be used to describe the data without an a priori definition of a model.

Descriptive formalisms, such as the Resource Description Framework (RDF) [204], are based on the building of statements composed of a subject linked by a predicate to an object. The repetition of this pattern, called a triple, becomes a graph with nodes and edges (Figure 15). By using such a formalism, it is possible to describe and store data into databases named Triplestore in the case of RDF [205]. From this common representation, ad hoc conversions to any data model are possible, reducing the number of mappings needed (Figure 16).

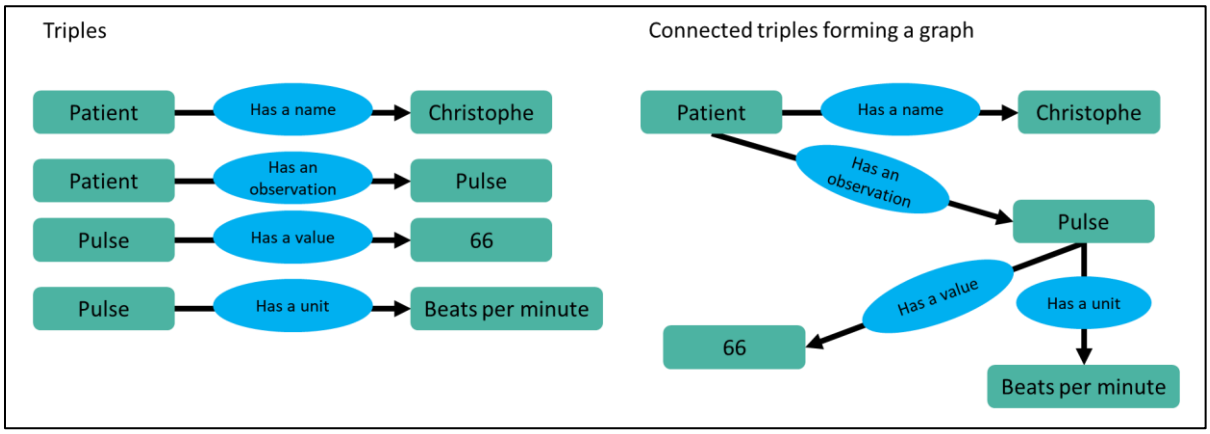


Figure 15 Triple representation as describe in descriptive formalisms such as RDF.

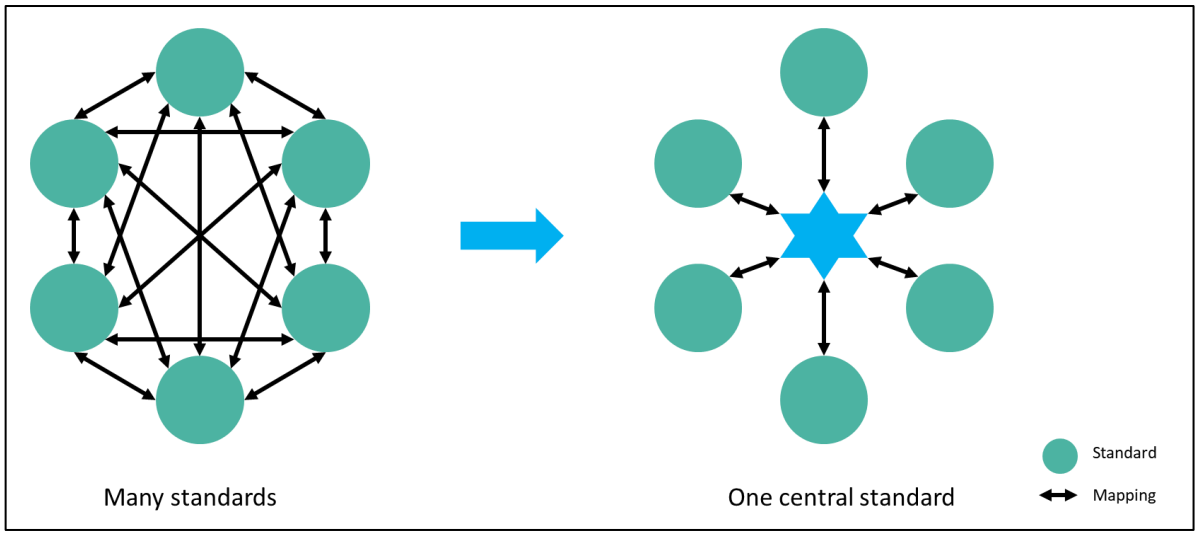


Figure 16 The benefit of a single central standard on the number of mappings needed (adapted from [206]).

6.8.2. Hypothesis 2

> The combinatorial power of such an interlingua exceeds the effective needs for clinical activities and can be reduced to a meaningful and manageable size.

The expressivity of natural languages or combinatorial controlled vocabularies is enormous. While ICD-10 has a finite set of codes of around 14'000, the post-coordination possibilities of SNOMED CT concepts or the number of possible sentences in French are almost limitless. This extreme expressivity brings new challenges when managing semantic representations of information. However, when focusing on the use of a language in a particular setting, only a small set of what is expressible in this language is used regularly. As a travel guide will gather a set of simple words or sentences that are specifically useful when in a foreign country, it should be possible to define a set of useful expressions extensively used in clinical settings. This set would be arguably smaller than the expressivity domain of a language such as medical French.

The situation is the same with the controlled vocabularies relevant to a certain use of the data. As explained previously, multiple views of the same concepts are relevant to different stakeholders in the hospital and beyond. Those different views can be called semantic dimensions as they represent a dimension in which a part of the semantic of the expressions can be represented. For example, the term "appendicectomy" means something totally different whether it is read by an internal medicine clinician, a surgeon, a nurse or the billing department. Therefore, from all the possible semantic dimensions in which an expression can be represented, we believe that it is possible to define a closed subset of relevant dimensions.

Based on those two assumptions, it should be possible to define a set of relevant expressions and to represent them in a set of relevant semantic dimensions. This would transform the impossible task of representing every expression in every dimension into a human-sized list of interoperable expressions and would bridge the gap between the expressions used by clinicians and the various representations of information needed in a hospital (Figure 17).

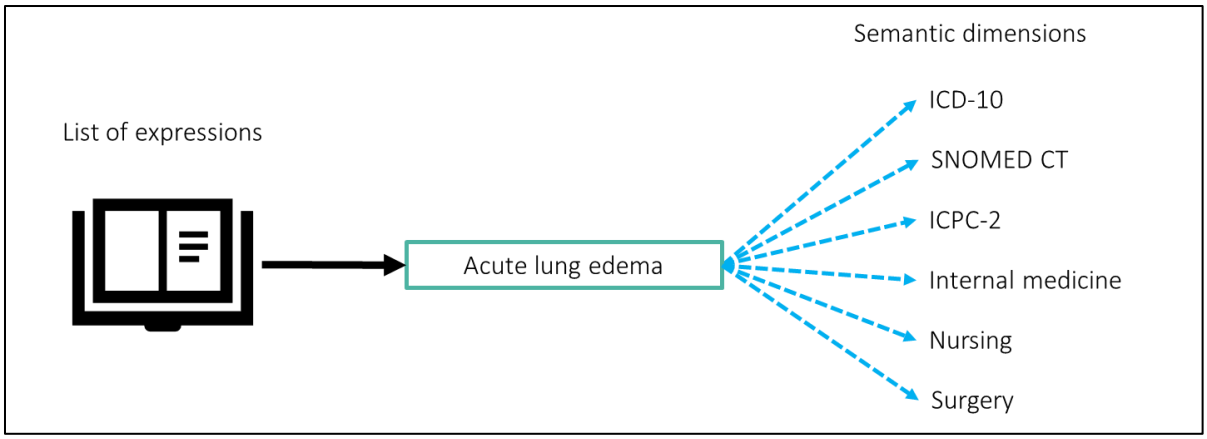


Figure 17 Structure of a possible common list of expressions enriched with semantic dimensions.

6.8.3. Hypothesis 3

> The interlingua can be used to represent the information expressed in clinical narratives, framing the challenge as an automatic translation task.

In hypothesis 1, it is assumed that a formalization of clinical data into a healthcare interlingua is a solution to the semantic challenge of interoperability. However, the method to attain this convergence is not defined. The problem can be split in two different approaches for two different types of data.

For what is commonly described as structured data, the conversion to the chosen interlingua can be made by binding the structured variables to concepts. For example, SNOMED CT has been mapped to multiple other controlled vocabularies and implemented for multiple types of structured data such as problem lists, chief complaints for emergency department, wound assessments or procedure codes [207–209]. For representing such data, it is possible to link a SNOMED CT representation to the variable as well as to the set of its possible values.

On the other hand, a major part of the EHR information is still hidden in free text or narratives [47,210]. Representing these “unstructured” data using SNOMED CT is not trivial. While it is possible to link a set of SNOMED CT concepts to a free-text document or, to be more specific, to annotate a document linking concepts to expressions [211], we propose a new approach based on two observations.

Most concepts expressed in clinical settings cannot be expressed as single elements in a classification. As any language used by humans, there is a need to be able to express a meaning with an association of concepts. For example, “acute myocardial infarction” can be found in single concept-based classifications such as ICD-10. However, clinical activity requires to express many additional information that will specialize this concept, such as uncertainty, severity, probability, extent, precise location or timing. “Acute myocardial infarction” can then become “images compatible with sequalae of probably repeated small antero-lateral myocardial infarction in the past”.

SNOMED CT presents similarities with a natural language. Indeed, with a compositional grammar, more than 350,000 concepts and 1,000,000 relations, SNOMED CT concepts can be used and combined into complex post-coordinated sentences in a similar way words are combined into sentences in French or English.

Considering SNOMED CT as a language with words (concepts) that can be combined in sentences (post-coordinations), it is possible to consider the challenge of representing narrative data in SNOMED CT as a translation task. To confirm the innovative potential of this approach and to review previous work in

this domain, a literature review was needed. Based on this knowledge, it should then be possible to develop a method to translate narrative data into SNOMED CT concepts, first manually, then automatically using NLP.

7. Publications

7.1. Methodological contributions

7.1.1. Contributions to hypothesis 1

Targeting the first hypothesis, the first article focuses on a large framework for national interoperability of structured variables. The crucial axiom of this work is that a framework aiming at creating interoperability for multiple communities, such as research healthcare and regulatory agencies, needs to be strongly semantically driven but agnostic of any data model.

In Switzerland, the Swiss Personalized Health Network (SPHN), an initiative started in 2017 and funded up to the amount of 137 million CHF until 2024, aims at leveraging research in the field of personalized health by building a nationally coordinated infrastructure that supports exchange and reuse of health data produced by the healthcare system [212,213]. The SPHN implemented a new approach to solve the interoperability challenge based on three pillars: a semantically robust framework, an agnostic descriptive formalism and ad hoc conversions to data-models. This approach was defined by the Clinical Semantic Interoperability working group (CSI) of the SPHN's Data Coordination Center (DCC) and has been implemented in every University Hospital and every Polytechnical School in Switzerland for sharing of clinical data.

During the first phase of the SPHN, the author of this thesis, Christophe Gaudet-Blavignac (CGB), was involved in multiple projects funded by the initiative. As a member of the Division of Medical Information Sciences (SIMED) of the Geneva University Hospitals (HUG), he participated to the LOINC for Swiss Laboratories infrastructure project in which he unified and aligned the LOINC coding of the five Swiss university hospitals [214]. He performed data extraction and annotation for the De-Identification of clinical narrative data in French, German and Italian infrastructure project [214]. He participated to the Swiss Frailty Network Repository driver project in which he encoded and mapped the project's codebook to the HUG's data warehouse [215] and he participated in the mapping and extraction of the data for the Swiss Personalized Oncology driver project [215].

The SPHN's three pillar approach was largely defined and driven by the CSI through monthly meetings and publication of datasets and strategic papers [216,217]. As an active member of this group, CGB participated to the creation and dissemination of the strategy. He was involved in the building of the core and extended datasets that were released by the SPHN. To scientifically formalize the strategy and to report on its implementation, CGB, Jean-Louis Raisaro (JLR) and Christian Lovis (CL) wrote a scientific article that was submitted to the Journal of Medical Internet Research (JMIR), (cf. 7.1.1.1).

This article confirms the first hypothesis by reporting on the successful implementation of the SPHN three-pillar strategy. The first pillar consists in the definition of a set of concepts encoded in SNOMED CT and a set of controlled vocabularies. They can later be combined to create new concepts fitting the needs of the driver projects. This confirms that it is possible to only use a restricted set of relevant information representations as an interlingua to represent clinical data.

The main contribution of this work is the description of a strategy for clinical data interoperability at the national level and the report on the successful implementation of this approach in Switzerland.

7.1.1.1. Article 1: A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study.

The outline of this article presenting the SPHN three pillar strategy, was created conjointly by CGB, JLR and CL. Then, CGB did the literature research and wrote the first draft of the article with the help of JLR for implementation parts, then CL reviewed it and the feedbacks were integrated by CGB. The article was then sent to Katrin Crameri, head of the Personalized Health Informatics Group responsible for the Data Coordination Center of SPHN for official endorsement and review. Finally, Vasundra Touré and Sabine Österle, also members of the SPHN DCC reviewed and provided feedbacks on the article. Once CGB integrated final feedbacks, the article was submitted to JMIR for peer review and publication in January 2021. In February 2021 the article came back with a request for revision. CGB JLR and CL made the revisions and the article was resubmitted and accepted for publication in May 2021.

7.1.2. Contribution to Hypothesis 2

The common problems list implemented in HUG since 2017 constituted a use case to confirm the second hypothesis of this work. When clinicians list the problems of their patients, they use natural language and create an expression. The complete set of possible expressions is not closed. Therefore, creating a list of all possible French expressions representing a problem is impossible. On the other hand, existing controlled vocabularies, such as the ICD-10, are not fit for a problem list due to various reasons such as residual aggregation or their lack of synonyms. Finally, a problems list contains information that could be useful beyond the medical domain. However, each profession or division of a hospital has its own vocabulary in which they are used to communicate. There is therefore a need for a problems list that fits the expressions the clinicians desire to enter and can represent them in multiple semantic dimensions.

This could only be achieved by reducing the list of possible expressions to a subset used in practice small enough to be manually represented into multiple semantic dimensions chosen based on specific use cases. The problems list presented in this article represents an attempt at creating a central source of useful information representations articulated around a restricted set of useful expressions, wide enough to represent the richness of clinicians' language.

Since 2016, CGB has overseen the development, deployment and maintenance of this manually created problems list. This list of more than 40,000 labels manually extracted from clinical documents was chosen to become the HUG's common problems list under the responsibility of CGB. The list required curation, additions and encoding into SNOMED CT and multiple other controlled vocabularies called semantic dimensions. Those tasks were accomplished by CGB directly or by the training and supervision of a multidisciplinary team of clinicians, nurses, medical students and semantic experts shifting the burden of encoding the information from the responsibility of the working clinicians. Once this first phase of curation ended, in January 2017, the list was deployed into the production environment of the HUG to be used as the common problem list of the hospital. After the first deployment, the list needed to be completed, improved and additional semantic dimensions needed to be mapped to the expressions. For four years and to this day, CGB pilots the development of this list, he supervises a team of two to five people working on different aspects of the list, has been creating verification and security checks and delivering regular production releases to the HUG, in collaboration with the Information Systems Directorate and the Medical and Quality Directorate of the HUG. After four years of implementation, CGB extracted usage data of the list from the data warehouse and described this work in a scientific article to report on the creation and deployment of this list as well as on its use and adoption by the users.

In four years, the list has become a central axis of the EHR. Usage data showed that it was the most used source of expressions for entering problems, that the number of created problems was rising and that the proportion of problem entered as free text, not using the list, was decreasing. Those results are indicators of the list's success in representing the language of the clinicians. After four years, the 20,120 expressions of the list seem to adequately cover more than 80% of the clinicians needs in term of expressions. Moreover, the various semantic dimensions added to the list allowed its usage for numerous other goals such as surgical theater planning, dietetic and nutrition diagnosis, decision support or clinical research.

These results confirmed the second hypothesis by showing the successful convergence to a restricted list of manually curated expressions mapped into SNOMED CT and useful semantic dimensions to answer multiple interoperability needs in the hospital.

7.1.2.1. Article 2: One list to rule them all and many semantics to bind them: Building a shared, scalable and sustainable source for the problem oriented medical record.

This article reporting on the building and evaluation of the common problem list was written by CGB, then reviewed and corrected by CL and Andrea Rudaz, the clinician in charge of the development and deployment of the problem module in HUG (the software used by clinicians to access the list). CGB then included the feedbacks and the article was submitted to JMIR in March 2021.

7.1.3. Contribution to Hypothesis 3

The last goal of this thesis is to develop a method to bring semantic interoperability to the notoriously complex narrative data. The third hypothesis of this work suggests that representing narrative data in the proposed interlingua can be framed as an automatic translation task. To validate this approach, a review of the actual uses of SNOMED CT for narratives was needed. This article was framed as a scoping review on the use of SNOMED CT to process or represent clinical free text. The results showed that if SNOMED CT had been widely used to process this type of data, it was often with rule-based systems, in English and without taking advantage of its compositional capabilities. Specifically, no article mentioned an approach framing the problem as a translation task. This review grounded the possibility to represent narratives using SNOMED CT as a conceptual framework and confirmed that the translation approach had not been attempted before.

When shifting the approach from an encoding task to a translation task, new insights about automatic translation had to be considered. Automatic translation approaches can be classified in three categories: rule-based, corpus-based and hybrid [218]. Rule-based approaches can be summarized using the Vauquois' triangle of automatic translation [219] (Figure 18). This triangle considers that automatic translation should be made by abstracting the meaning of the input text into a theoretical pivotal language or "interlingua" that represents the semantic content of the text. Then, the generation of the target language representation is the reverse task of generating language from the abstracted interlingua form. The transfer could occur at a different level with different expected quality. The direct transfer from the source to the target, meaning a direct replacement of the words, requires no syntactic processing but an important morphological analysis. The syntactic transfer occurs after the syntax has been abstracted to the interlingua and requires syntactic processing. Finally, the semantic transfer requires the complete abstraction of the source into the interlingua and generation into the targeted language.

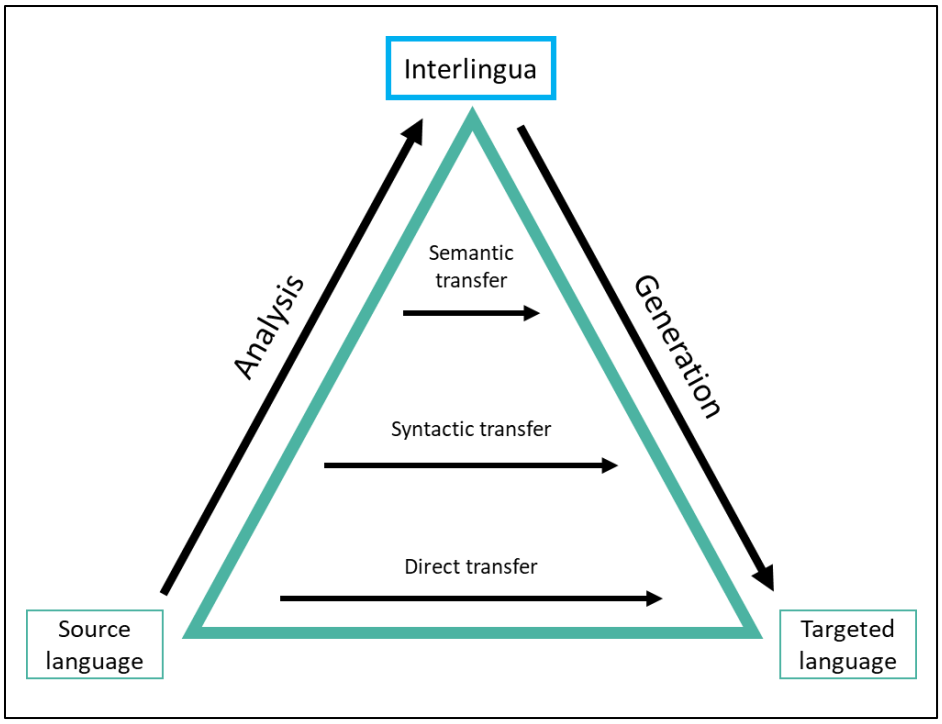


Figure 18 Summarization of the Vauquois' triangle [219].

It is assumed in this thesis that SNOMED CT can be the interlingua acting as a bridge between human conceptual frames and machine representation. The parallel with the first half of the Vauquois' triangle seems obvious. Therefore, the translation task should be only the abstraction process that will lead from French medical text to the SNOMED CT representation. Unfortunately, the reality was not as simple, due to the creation process of SNOMED CT and its primitive concepts.

The influence of a language on the thoughts of its speaker, a concept known as linguistic relativity, has been thoroughly discussed in the linguistic domain [220–222]. If the extent of this influence is subject to debate and beyond the scope of this work, it is commonly accepted that the language of a speaker can facilitate some conceptualizations or distinctions over others [221]. In other words and as a reference to the language philosopher Ludwig Josef Johann Wittgenstein, it is the words we choose that give meaning to the world [223]. While SNOMED CT is supposedly language independent [224], meaning that the information contained in a concept is not linked to its natural language representation, it was however designed by English speakers, in the United Kingdom for the CTV3 terminology and in the United States for SNOMED RT. Therefore, since SNOMED CT has been designed and first represented in English, its concepts are more likely to represent an English-speaking way of representing the reality.

Moreover, SNOMED CT contains primitive concepts [225]. A concept is named "primitive" if its formal definition, the set of relationships linking it to other concepts, is not sufficient to distinguish it from every other concept. In such concepts, the description in natural language is the only way of capturing its complete meaning. Since this description is in English for the International version of SNOMED CT, it is arguable that SNOMED CT is not language independent, but strongly linked to English.

Therefore, the process of translating French medical text into SNOMED CT cannot be reduced to the abstraction side of the Vauquois' triangle but should be seen as a translation between French and an entity that is an English-influenced interlingua.

To confirm this insight concerning the link between SNOMED CT and natural languages, several tasks were started. First, to investigate the translation of SNOMED CT concepts into French and German, the author of this thesis participated in the translation of the starter kit of SNOMED CT concepts for Switzerland. This starter kit was composed of a set of 6,393 concepts translated in French and German and was supervised by eHealth Suisse [226]. CGB piloted the work to validate this translation, managing a team of two experts and acting as a negotiator for difficult items. This mandate helped gather experience on the translation of SNOMED CT concepts in two languages and was a first step toward the final goal of translating French medical language into SNOMED CT.

Secondly, building on a collaboration with a group from the Yonsei University in Seoul, a dictionary-based NLP tool named PKDE4J was adapted to extract biomedical entities and relations from user-generated text on a social media platform. The objectives of this work were first to evaluate the adaptability of the tool to a new type of text, differing from scientific publications, to evaluate if a dictionary created using expressions directly extracted from SNOMED CT was able to capture biomedical entities represented in social media content and finally to analyze how people express themselves about chronic diseases on social media.

The dictionaries used for the entity extraction covered the following categories: drugs, anatomy, procedures, symptoms, findings and side effects. Among those, the Finding and Procedure dictionaries were built by extracting specific sets of SNOMED CT concepts based on semantic tags. Semantic tags are notation at the end of a SNOMED CT description indicating its relevant domain. All concepts with a "finding" or "procedure" semantic tag were extracted from the complete list of SNOMED CT concepts. Concepts of more than two words were filtered out to simplify the task on the ground that the probability of a long concept to appear in user generated content was low. The final dictionaries contained 7,800 procedures and 7,620 findings.

Focusing on the SNOMED CT dictionaries, on a corpus of 17,580 user messages and 2,137,115 tokens, 11,549 entities representing 296 different procedures and 8,741 entities covering 483 different findings were extracted. An evaluation of the entity extraction was made over 1,000 random messages and showed a performance of 78.48% (3,682/5,151 entities correctly extracted).

This work showed that, for the English language, the English description of SNOMED CT concepts could be used to automatically extract concepts from a natural language outside of the realm of scientific publications and medical writing. However, the number of different concepts was arguably low with 679 different entities over dictionaries of 15,420 concepts.

Then, CGB started to work on the representation of complex medical French sentences in post-coordinated SNOMED CT sentences. This started with the manual translation of sentences found in discharge letters of the HUG. CGB used an annotating tool named Brat [227] to annotate portions of text with SNOMED CT concepts, then adding attributes as relations between those annotated entities. After a first set of letters manually translated, CGB wrote guidelines to perform the translation and trained four annotators for this task. A set of 60 discharge letter discussions was selected and fully translated into SNOMED CT sentences. Each letter was translated conjointly by a group of two annotators and a set of 20 letters was annotated by both groups. This work has not yet been the subject of a publication but allowed to gather significant insight on the common difficulties encountered when performing this translation.

Then, CGB started a collaboration with Vasiliki Foufi (VF) and Professor Eric Wehrli (EW) to adapt Fips, a syntactico-semantic parser developed at the University of Geneva (UNIGE) by EW [228,229] to automatize the translation of medical French into SNOMED CT post-coordinated sentences. This work required that the compositional grammar of SNOMED CT be integrated in the tool, that the description

model rules of SNOMED CT be converted in rules applicable by the tool, which was made by EW and CGB, and an important work on lexico-semantic resources. The French translation of 170,000 SNOMED CT concepts were provided by the Professor Stefan Darmoni. 80,000 of those translations were then curated manually by CGB, VF and trained students. The curation was accomplished according to translating rules published by SNOMED International and adapted by CGB to French.

A selection of those verified translations was then made by lexical analysis of a large corpus of discharge summaries, and around 15,000 terms were manually selected and inserted by CGB and VF in the electronic dictionary of Fips, along with lexical and grammatical information. Then, the tool processed multiple corpora of clinical text, allowing fine tuning of the parameters.

The post-coordination capability of the software was implemented for three different SNOMED CT relationships. To provide preliminary results on automatic post-coordination, the method was applied to a new corpus of discharge letters. The results were then manually validated by CGB for semantic correctness. The validation was split in two parts: the correctness of the annotation of simple concepts and the correctness of post-coordinated triplets composed of a concept linked to another concept by a relation. Only when the three elements were correct, the triplet was considered correct.

7.1.3.1. Article 3: Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review.

The protocol for this scoping review was created by CGB, CL, VF and Mina Bjelogrlic (MB) based on similar work in the field [186,230]. The selection and reading of the papers were made by CGB and VF and disagreements were discussed and resolved. The results of the review were documented and analyzed in an excel sheet. CGB wrote a first version of the article that was reviewed by MB, VF and ultimately CL. Comments and modifications were gathered and included by CGB. When a stable version of the article was finalized, it was submitted to JMIR. The peer review requested some modifications. CGB included them and the article was resubmitted and accepted for publication.

7.1.3.2. Article 4: Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases with Extracted Entities and their Relations

The research and writing of this article was performed by three researchers collaborating equally. Tatsawan Timakum (TT) from the Yonsei University helped in the tuning of the NLP tool, VF focused on the processing of the data and CGB on the dictionary creation and the evaluation. A draft of the article was written conjointly, TT writing the description of the tool, VF the method and results and CGB the discussion. Then feedback was gathered from the two Professors CL and Min Song from the Yonsei University. The feedbacks were included by the three authors and the article was submitted to JMIR. After a round of revisions, the article was accepted for publication.

7.1.3.3. Conference article 1: Automatic Annotation of French Medical Narratives with SNOMED CT Concepts

An evaluation of the annotation was performed by CGB by manually annotating free text with SNOMED CT concepts and comparing with the output of the tool on the same documents. Precision, recall and F-score were then computed. CGB wrote the first version of the article that was then reviewed by CL, EW and VF. The feedbacks were added to the manuscript and it was submitted, revised and presented by CGB at the Medical Informatics Europe Conference 2018 in Gothenburg.

7.1.3.4. Conference article 2: Reconnaissance et représentation automatiques de concepts médicaux français en SNOMED CT

This work on the automatic recognition and representation of SNOMED CT concepts and the evaluation of the automatic post-coordination was accomplished by the author of this thesis. The article written by the author of this thesis was then reviewed by CL, EW and VF. Feedbacks were included by CGB and

the article was submitted and presented by CGB during "TALMED 2019: Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical" [231]. This symposium took place during MEDINFO 2019 in Lyon.

7.1.4. Other publications made during the thesis period

The results presented in this dissertation are a relevant subset of multiple years of work. However, multiple other publications have been made by the author during this time, around the same thematic, without strictly entering the scope of this thesis. They are listed and briefly summarized in the following section.

7.1.4.1. Vishnyakova D, Gaudet-Blavignac C, Baumann P, Lovis C. Clinical Data Models at University Hospitals of Geneva. Stud Health Technol Inform. 2016;221:97-101.

This work is a brief report evaluating the advantages and disadvantages of different data models commonly used in healthcare. It was based on existing projects: the Swiss Transplant Cohort Study for which CGB was in charge of the IT during five years [232], the European FP7 DebugIT project [233] and the EHR4CR project [234]. The comparison shows that the described data models are strongly dependent on the objectives of the projects which underscores again the observation made about the non-neutrality of data models.

7.1.4.2. Walpoth BH, Meyer M, Gaudet-Blavignac C, Baumann P, Gilquin P, Lovis C. The International Hypothermia Registry (IHR): Dieter's ESAO Winter Schools and Beat's International Hypothermia Registry. Int J Artif Organs. 2017 Jan 1;40(1):40-42

In the frame of his work in the SIMED, CGB piloted the IT maintenance of five web-applications used for national and international cohort projects. The International Hypothermia Registry was one of them. This application allowed the gathering of clinical data about hypothermia patients by multiple centers around the globe through specifically designed forms. This publication reports on the creation of this registry.

7.1.4.3. Foufi V, Gaudet-Blavignac C, Chevrier R, Lovis C. De-Identification of Medical Narrative Data. Stud Health Technol Inform. 2017;244:23-27.

This publication reports on the development of a rule-based method for the de-identification of French medical narratives. The work on de-identification of natural language is an ongoing project in the SIMED. When starting to work on the automatic French to SNOMED CT translation tool, the need for a solution to de-identify documents became obvious. A set of finite state automata was developed by VF and CGB to process documents and remove personal health information. An evaluation on 20 random discharge summaries showed a 0.98 total recall and 1.0 precision. This project is still ongoing combining three approaches, knowledge based, rule-based and machine learning based.

7.1.4.4. Lovis C, Gaudet-Blavignac C, Chevrier R, Robert A, Issom D, Foufi V. BigData, intelligence artificielle, blockchain : guide pratique [Bigdata, artificial intelligence and blockchain for dummies]. Rev Med Suisse. 2018 Sep 5;14(617):1559-1563. French. PMID: 30226672.

This publication was a collective effort in the SIMED to propose a simple introduction to common subjects in the field of medical informatics such as NLP, artificial intelligence, and graph databases.

7.1.4.5. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res. 2019 May 31;21(5):e13484.

Along with the work on de-identification of French medical narratives, a scoping review was conducted to explore the understanding of the concepts of anonymization and de-identification in healthcare. This review was mainly conducted by Raphaël Chevrier, in close collaboration with the rest of the team.

This review concluded that there was a need for clearer definitions as well as for better education and dissemination of information on the subject.

7.1.4.6. Rochat J, Gaudet-Blavignac C, Del Zotto M, Ruiz Garretas V, Foufi V, Issom D, Samer C, Hurst S, Lovis C. Citizens' Participation in Health and Scientific Research in Switzerland. Stud Health Technol Inform. 2020 Jun 16;270:1098-1102.

This work reports on a survey conducted by the SIMED team named Evalab. This team is dedicated to human factors, man-machine interactions and user-centered design. In this frame, a survey about the factors motivating Swiss citizens to participate to research was made highlighting the lack of opportunity as the main factor blocking participation.

7.1.4.7. Foufi V, Ing Lorenzini K, Goldman JP, Gaudet-Blavignac C, Lovis C, Samer C. Automatic Classification of Discharge Letters to Detect Adverse Drug Reactions. Stud Health Technol Inform. 2020 Jun 16;270:48-52.

One of the applications of the multiple tools developed in the SIMED for processing clinical documents is the detection of adverse drug events. Using manual annotations of discharge letters and three different machine learning based classification methods, we managed to obtain an F1 score of 0.95 on classifying letters containing an adverse drug event and 0.91 on letters without adverse drug event.

7.1.4.8. Turbé H, Bjelogrlic M, Robert A, Gaudet-Blavignac C, Goldman JP, Lovis C. Adaptive Time-Dependent Priors and Bayesian Inference to Evaluate SARS-CoV-2 Public Health Measures Validated on 31 Countries. Front Public Health. 2021 Jan 21;8:583401.

This work is a direct product of the effort made by the SIMED to support the HUG during the first wave of the pandemic. With a new rapidly evolving situation, it became crucial to be able to foresee the impact of health measures on the reproduction number of the epidemic. This method estimates the reproduction number using Bayesian inference with time-dependent priors enhancing previous approaches by considering a dynamic prior continuously updated as restrictive measures and comportments within the society evolve. The main work was accomplished by Hugues Turbé, helped by the rest of the team in gathering, annotating and analyzing the data.

7.2. Publication manuscripts

7.2.1. Article 1

Original Paper

A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study

Christophe Gaudet-Blavignac[1,2], BSc, MSc; Jean Louis Raisaro[3,4], PhD; Vasundra Touré[5], PhD; Sabine Österle[5], PhD; Katrin Crameri[5], MPH, PhD; Christian Lovis[1,2], MD, MPH, FACMI

[1]Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland
[2]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
[3]Data Science Group, Division of Information Systems, Lausanne University Hospital, Lausanne, Switzerland
[4]Precision Medicine Unit, Department of Laboratories, Lausanne University Hospital, Lausanne, Switzerland
[5]Personalized Health Informatics Group, SIB Swiss Institute of Bioinformatics, Basel, 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 223726201
Email: christophe.gaudet-blavignac@hcuge.ch

Abstract

Background: Interoperability is a well-known challenge in medical informatics. Current trends in interoperability have moved from a data model technocentric approach to sustainable semantics, formal descriptive languages, and processes. Despite many initiatives and investments for decades, the interoperability challenge remains crucial. The need for data sharing for most purposes ranging from patient care to secondary uses, such as public health, research, and quality assessment, faces unmet problems.

Objective: This work was performed in the context of a large Swiss Federal initiative aiming at building a national infrastructure for reusing consented data acquired in the health care and research system to enable research in the field of personalized medicine in Switzerland. The initiative is the Swiss Personalized Health Network (SPHN). This initiative is providing funding to foster use and exchange of health-related data for research. As part of the initiative, a national strategy to enable a semantically interoperable clinical data landscape was developed and implemented.

Methods: A deep analysis of various approaches to address interoperability was performed at the start, including large frameworks in health care, such as Health Level Seven (HL7) and Integrating Healthcare Enterprise (IHE), and in several domains, such as regulatory agencies (eg, Clinical Data Interchange Standards Consortium [CDISC]) and research communities (eg, Observational Medical Outcome Partnership [OMOP]), to identify bottlenecks and assess sustainability. Based on this research, a strategy composed of three pillars was designed. It has strong multidimensional semantics, descriptive formal language for exchanges, and as many data models as needed to comply with the needs of various communities.

Results: This strategy has been implemented stepwise in Switzerland since the middle of 2019 and has been adopted by all university hospitals and high research organizations. The initiative is coordinated by a central organization, the SPHN Data Coordination Center of the SIB Swiss Institute of Bioinformatics. The semantics is mapped by domain experts on various existing standards, such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), and International Classification of Diseases (ICD). The resource description framework (RDF) is used for storing and transporting data, and to integrate information from different sources and standards. Data transformers based on SPARQL query language are implemented to convert RDF representations to the numerous data models required by the research community or bridge with other systems, such as electronic case report forms.

Conclusions: The SPHN strategy successfully implemented existing standards in a pragmatic and applicable way. It did not try to build any new standards but used existing ones in a nondogmatic way. It has now been funded for another 4 years, bringing

XSL•FO
RenderX

the Swiss landscape into a new dimension to support research in the field of personalized medicine and large interoperable clinical data.

(JMIR Med Inform 2021;9(6):e27591) doi: 10.2196/27591

KEYWORDS

interoperability; clinical data reuse; personalized medicine

Introduction

Background

Interoperability is a well-known challenge in medical informatics and is one of the main obstacles preventing data-driven medicine from realizing its full potential. Efforts to classify and express meaning in health care are as old as the International Classification of Diseases (ICD) [1]. Organizations, such as Health Level Seven International (established in 1987) [2] and SNOMED International, which maintains and releases the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [3], are dedicated to promoting interoperability in health care. Moreover, multiple national and international programs are seeking to promote interoperability. Examples of major initiatives designed to tackle the interoperability challenge in health care include the Meaningful Use program under the Health Information Technology for Economic and Clinical Health Act [4] in the United States and the Integrating the Healthcare Enterprise initiative [5], with more than 175 member organizations worldwide.

Semantic Interoperability

Semantic interoperability usually involves controlled vocabularies. In the medical field, the equivalent is part of the culture, but is named differently, such as scales, scores, and classifications. These involve organization of medical knowledge into a finite set of classes. They are used in daily practice to evaluate, describe, and prognose many situations or conditions. For example, medical scales or scores are narrow-scope classifications used in everyday medical practice. The Glasgow Coma Scale [6] and the Apgar score [7] are examples to describe the level of consciousness of patients and the health of newborns, respectively. In some cases, there are several of them for a specific condition with different perspectives, such as for heart failure [8]. Clinicians commonly use dozens of scores and scales in their daily practice, and there are numerous applications that combine and facilitate their use [9,10].

More extensive medical classifications, such as the 10th revision of the International Classification of Diseases (ICD-10) [11] and the Logical Observation Identifiers Names and Codes (LOINC) [12], are large systems that attempt to organize broader areas of medical knowledge, such as diseases and causes of death (ICD-10), or health measurements, observations, and documents (LOINC).

They can be articulated into larger representations (meta-organizations) that consolidate several classifications, ontologies, terminologies, etc. The Unified Medical Language System (UMLS) Metathesaurus [13,14], for example, combines several classifications having different purposes, such as diagnosis encoding and literature indexing. SNOMED CT is another example, which combines 19 top-level hierarchies into one representation.

Specific classifications are characterized by a partitioning of the knowledge represented according to a specific purpose, usually the intention for which the classification has been designed. Thus, SNOMED CT is historically dedicated to pathology and was extended later with clinical codes. ICD-9 and 10 are well adapted to represent diagnosis and morbidity causes, while LOINC is mostly used to represent laboratory analytical and preanalytical characteristics. Drugs are often handled using Global Trade Item Number (GTIN) for logistical needs and Anatomical Therapeutic Chemical (ATC) classifications for order entry decision support [15,16], while adverse drug reactions are reported in MedDRA [17].

Challenges for Semantic Interoperability

As a result of having specific classifications well designed for specific purposes, they are usually not well adapted to express other types of knowledge or different organizations (partitioning) of that knowledge. SNOMED CT is able to represent almost any pathological test result and has been used to represent free text, but it fails to express some types of concrete values [18,19]. ICD-10 can be used to assign a code to any disease, but its mono-hierarchical structure prevents meaningful information reuse (eg, it is not possible to easily extract all codes representing infectious diseases). Finally, GTIN identifies commercial drug products, but it does not efficiently represent active substances, while ATC expresses only substances, but not the products. Classifications are tools used to represent the meaning of the data, but they always carry an intent, and none can be used for every purpose.

Data Organization

Data are usually organized with data models, and the first and most simple is the text or tabular file that is still widely used, notably in clinical research settings. The serialization of data in comma-separated value (CSV) files can be expanded into more complex representations. Data models structure the data into entities and relationships that fit a given purpose. These have existed in health care for a long time, and some of them are widely used. Health Level 7 (HL7) version 2, which is the most widely implemented standard for health care in the world [20], is linked to the Reference Information Model (RIM), a data model designed to be the backbone of HL7, with the following three main classes: Act (representing something that has happened or will happen), Entity (any living or nonliving thing), and Role (a competency expressed by an Entity). These three classes will then be used to build an event using a "connector" named "Participation" that allows building of complex nested structures [21]. Finally, as for controlled

XSL•FO
RenderX

vocabularies, data models can be articulated in meta-models, such as the bridge recently created between the Observational Medical Outcomes Partnership (OMOP) and the Informatics for Integrating Biology and the Bedside (i2b2) [22] data models.

Challenges of Data Interoperability

The structure of each data model depends on the goal of the standard and on the community that will use the data. For example, the RIM was primarily targeted at electronic health record (EHR) interoperability, while the Common Data Model of OMOP specifically targeted clinical research [23]. The data model of i2b2 [24] is designed to integrate genetic and phenotypic data, while the Clinical Data Interchange Standards Consortium (CDISC) operational data model [25] is required for drug regulatory constraints by the United States Food and Drug Administration. The openEHR project is built around another paradigm and is composed of archetypes that are small domain models aimed at providing a specific piece of information. The definition of archetypes and templates of archetypes are very flexible and can solve numerous interoperability challenges; however, it still requires adopting the reference model for the storage of data [26,27]. The design of these models is based on specific goals, and there is no one-size-fits-all data model that can serve every purpose.

The Swiss Personalized Health Network

The Swiss Personalized Health Network (SPHN) aims to leverage research in the field of personalized health in Switzerland by building a nationally coordinated infrastructure network that supports exchange and reuse of health-related data produced by the health care system and in biomedical and clinical research settings [28]. This national initiative was launched in 2017, with funding of up to CHF 137 million (US $153 million) assured until 2024 [29,30]. In essence, the goal of the SPHN is to connect the Swiss health care system, the research community, regulatory agencies, and eventually industrial partners involved in personalized medicine. Consequently, the SPHN is at the interface between three communities and must overcome the multiple challenges of exchanging data in a secure, interoperable, and meaningful manner.

Objectives

The challenges of interoperability described above have been the focus of active research in recent decades. Every year, new standards appear with the goal of addressing the remaining challenges. Interestingly, each of these new standards solves some problems but also generates new ones.

As opposed to conventional approaches, which are aimed at mapping data to one common standard and are in practice only effective for specific use cases, our interoperability strategy uses existing standards in a purpose-specific and complementary manner without depending on any particular one, thus providing great flexibility and sustainability. As such, it enables data interoperability between various communities, each of which has different needs or follows different requirements with regard to the type of data model to be used.

Vocabulary

Interoperability is by essence an interdisciplinary process. Therefore, the vocabulary used to describe its components can vary. This section aims to define the words used in this work and their meaning. *Data model* is an abstract model that organizes elements of data in structures. *Data model-independent* is used to describe a system that does not depend on a predefined data model. *Encoding* is the action of expressing something with a specific coding system. For example, encoding a concept into a terminology means linking this concept to the elements of the targeted terminology that adequately represent it. *Interoperability* is the ability of two different entities to connect, share, understand, and use data in their processes. *Semantics* is the encoding of meaning into one or more knowledge representations (KRs). *Knowledge representation* is organization of knowledge into a list of elements, such as controlled vocabularies, terminologies, classifications, taxonomies, ontologies, thesauri, and coding systems.

Methods

Overview

Based on the lessons learned from previous attempts, this work addresses the interoperability challenge adopting a semantic-driven data model–independent framework based on the following three pillars (Figure 1):

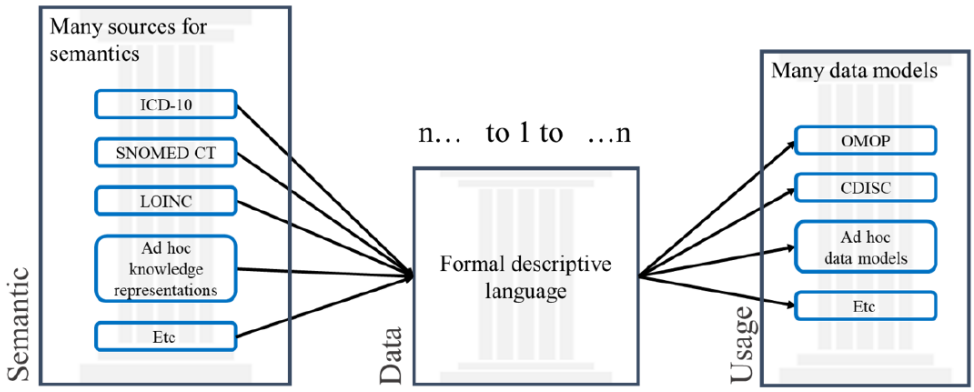
1. A multidimensional encoding of the concepts. Only the required concepts (variables) are encoded in any KR system. This decision is completely agnostic, so that several international standards can be used at any time, according to the needs.
2. Resource description framework (RDF)-based storage and transport of the instances of these concepts when used to express clinical data. The RDF is well suited for a federated national exchange format. As it is a formal descriptive language, it is very scalable to any future needs not yet known.
3. Conversion of the RDF to any target data model that is needed for a specific research community or usage, according to the needs of the users.

This ends up with the first two pillars being completely data model independent. Only at the third pillar will the data be available in any required model, such as CDISC and OMOP, according to needs. We thus considered this strategy "semantic agnostic" and "model independent."

This strategy is being implemented stepwise since January 2019. This paper focuses on the strategy. The deployment and societal challenges will be discussed in a further publication.

XSL•FO
RenderX

Figure 1. The three pillars of the proposed data interoperability strategy. CDISC: Clinical Data Interchange Standards Consortium; Etc: et cetera; ICD-10: International Classification of Diseases; LOINC: Logical Observation Identifiers Names and Codes; OMOP: Observational Medical Outcomes Partnership; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



Integrative and Usability-Focused Semantic Approach

As stated in the Introduction, it is illusory to believe that different communities will adopt a single standard for the sake of mutual compatibility. Therefore, this strategy does not enforce a specific KR to express meaning. The goal is to enable the use of an adequate KR, based on the purpose and context of use, without imposing any specific one. However, the presence of a semantic definition of the data is crucial and must be the central axis of the strategy.

The first pillar of our approach consists, therefore, of developing a semantic framework comprising a set of concept definitions relying on existing KRs or new ones if needed. The concept definition must be adapted to the granularity required by the use case. Each concept can be encoded into as many KRs as required. For instance, the concept "Heart Rate" can include encoding into SNOMED CT and LOINC. The power of representation and usability is prioritized over conceptualization. It is thus possible to express the meaning of the data without enforcing a specific KR. Finally, instantiations of the concepts can use an adequate KR, depending on the context. Axioms of the first pillar are summarized in Textbox 1.

Textbox 1. Axioms of the first pillar of the strategy.

Axioms

- Framework composed of a set of concept definitions.
- Semantic encoding using a knowledge representation.
- Multiencoding of a concept in several knowledge representations allowed.
- Selection of concepts defined by use cases.
- Combination and extension of concepts allowed.

Descriptive Formalism for Transfer and Storage

Transport and storage of information are essentially the same. Since transport is a "moving storage" and storage is a "nonmoving transport," they can be regarded as a single challenge. The data and concept landscapes in health care are constantly evolving with new elements to be exchanged. To best answer this need for sustainability, scalability, and plasticity, the strategy is based on the use of a descriptive formalism (eg, the RDF, the Arden syntax, and the Web Ontology Language [OWL]) [31-33]. These languages offer flexible storage and transport of information (be it data, semantics, processes, or rules). This differs from a data model–based approach, as it does not constrain data to fit a specific format but only describes the data and its semantics in an intuitive and unconstrained way as it is collected at the source. Our approach allows the use of different formalisms when needed. For example, RDF can be used to store and transport the data, and the Arden syntax can be used to describe rules, such as automatic alert and clinical decision support systems [32]. Similarly, other formalisms can be used for other types of information and purposes (eg, Guidelines Interchange Format for guidelines [34] and Java Business Process Model for workflows [35]). Textbox 2 summarizes the approach for the second pillar.

XSL•FO
RenderX

Textbox 2. Axioms of the second pillar of the strategy.

Axioms

- Common approach for storage and transport.
- No a priori definition of a data model.
- Use of descriptive formalisms to describe the data encoded in the first pillar.
- Choice of the formalism depending on the use case.

Purpose-Specific Transformation to Data Models

The final building block of our strategy is the transformation of data from a flexible representation, based on formal descriptive languages, to a more rigid but application-oriented representation, such as relational data models. The goal is to provide a way of efficiently sharing data between different communities used to working with their own data models. As mentioned above, no common data model can be adopted by all communities, and mappings across data models are often partial because of incompatible information representations.

As a result of the first and second pillars, it is possible to create ad hoc conversions based on users' needs. In particular, the use of a data model–independent formalism to store data enables the implementation of one-to-many mappings to any target data model. For example, existing work has already proposed the transformation of RDF resources into customized relational data models [36] or standard common data models, such as i2b2 [37,38] and OMOP [39]. This approach addresses the complexity of the current many-to-many mappings and will enable the sharing of data with any community, provided that the mapping is done while keeping the data unchanged. Textbox 3 summarizes the approach.

Textbox 3. Axioms of the third pillar of the strategy.

Axioms

- Ad hoc conversions from the descriptive formalism of the second pillar to data models.
- Building of a reusable one-to-many mapping catalog.
- Selection of the targeted data models based on use cases.

Results

The SPHN

The proposed interoperability strategy was implemented to serve the data-sharing needs of the SPHN. The projects supported are all large multicentric projects, multihospitals, multiresearch centers, and data-driven research related to personalized medicine [28]. They vary in terms of not only methodology and research questions, but also the clinical data concepts requested from the data providers involved. The projects are designed to generalize the use of the Swiss General Consent, improve clinical data management systems on care providers, build a national data interoperability landscape for research, and leverage research organizations.

The defined approach was implemented by every university hospital and high research organization of Switzerland as the national standard for sharing clinical data. Twelve driver projects were funded and used the approach for their data needs.

Organization

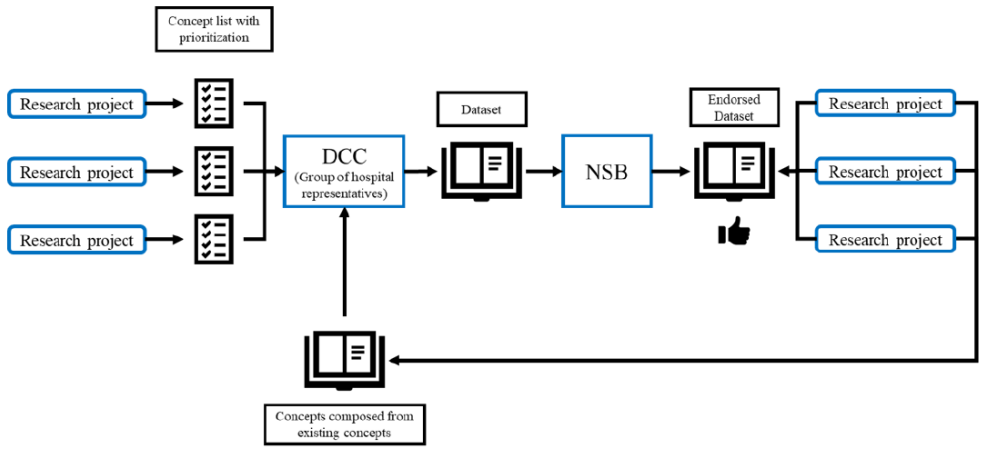
In the implementation of the first pillar, a semantic framework has been built and maintained by the SPHN Data Coordination Center (DCC). The DCC is the central hub for data interoperability in the SPHN and part of the SIB Swiss Institute for Bioinformatics. Its mandate is to coordinate the development of the specification of the structure and semantics of the SPHN data set, which describes the type of data that is available and potentially shareable within the network (hereafter referred to as the SPHN semantic data set). A full description of the DCC is available on the SPHN website [40].

First Pillar

The content of the SPHN semantic data set is defined by leveraging domain knowledge from the Swiss clinical research community. Every research project provides the list of variables they need to the group in charge of aligning the semantics. This group includes domain experts and clinical semantics specialists. This SPHN semantic data set is periodically reviewed and extended according to experience obtained in projects by extracting and using the data and, of course, the new needs of research projects. There is a validation process that ends up in the publication of a new release of the core list of concepts endorsed by the SPHN National Steering Board (NSB). After official release, the new concepts are used by university hospitals for interoperable data exchange. The steps involved in this process are shown schematically in Figure 2. The complete structure of the SPHN is beyond the scope of this article and openly available in published reports [30].

XSL•FO
RenderX

Figure 2. Flowchart of the validation process. DCC: Data Coordination Center; NSB: National Steering Board.



The concept list is evolving, such that each element contains, in addition to semantics, management metadata, such as unique ID, a name, a description, and several fields for versioning. All data transfer for SPHN projects should comply with these concepts once enforced by the NSB. Examples of the encoding of these concepts with SNOMED CT and LOINC are shown in Table 1, in which the code is linked to the row where relevant and applicable. As more use cases arise, new encodings can be created.

The DCC has the task of exploring common international KR when validating new concepts, so as to select the most appropriate one. A KR for a concept is chosen taking into consideration not only its capacity to represent the concept correctly and unambiguously but also the ability of hospitals to comply with it and the research project to use it. Currently, more than 300 concepts are being used, which can describe demographics, laboratory analysis and results, drugs and prescriptions, clinical and physiological variables, etc [41].

Table 1. Examples of encoded concepts used to describe a temperature measurement.

Concept name	SNOMED CT[a]	LOINC[b]
Temperature	386725007 |Body temperature (observable entity)|	8310-5 Body temperature
Unit	767524001 |Unit of measure (qualifier value)|	N/A[c]
Body site	123037004 |Body structure (body structure)|	39111-0 Body site

[a]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.
[b]LOINC: Logical Observation Identifiers Names and Codes.
[c]N/A: not applicable.

Second Pillar

The data storage and transport step of the SPHN was implemented using RDF as proposed by the World Wide Web Consortium [42]. RDF allows to map instances of real data originating from a clinical database with the conceptual framework defined in the first pillar. The RDF allows to build a labeled directed multigraph, where nodes and edges are identified by uniform resource identifiers. The basic entity in the RDF graph is known as a "triple" and is composed of a subject, a predicate, and an object. Several triples compose a graph. Since the RDF does not depend on a specific semantic standard, it allows for the use of different ontologies and value sets, as required by the strategy. The reasons for choosing RDF technologies are summarized in Textbox 4 [43-47].

Textbox 4. Reasons for choosing the resource description framework.

Reasons

- Flexibility to represent complex knowledge with simple statements (ie, triples of information).
- Scalability to other fields (eg, the resource description framework [RDF] has been adopted by systems biology and molecular biology for specific data representation [43,44]).
- Advanced query system (ie, with the SPARQL language).
- Existing tools in a rich community to create, maintain, validate, explore, and visualize RDF representation (eg, Protégé and WebVOWL [45-47]).

XSL•FO
RenderX

A set of rules and conventions has been defined to guide the creation of an SPHN RDF schema, that is, how RDF classes and properties required to generate instances (RDF resources) for storing hospitals' data should be created [48,49]. Particularly, such rules stipulate (1) how concepts defined in the SPHN semantic data set should be converted into RDF classes or RDF properties and (2) how concepts that are not semantically linked to each other by composition should be linked to encapsulate contextual information provided at the time of data capture.

Swiss hospitals' clinical research data warehouses are primarily based on relational database management systems. To transform data from a relational model representation into a graph representation based on RDF, extract, transform, and load (ETL) pipelines have been implemented by data providers' informatics teams. They typically include an RDF transformer step where raw data from the EHR is converted and loaded into a triple store. Then, data can be extracted and serialized into RDF files for each specific project.

Third Pillar

Converters are used to transform the RDF data into purpose-specific data models, serializing the RDF data into other common formats such as XML, JSON, JSON-LD, and TSV/CSV. For example, SPARQL queries have been implemented to convert data into flat files that can be processed by research-enabling software or machine learning pipelines [49].

Discussion

Overview

While the proposed data interoperability strategy offers a number of advantages in terms of flexibility and extensibility over more conventional approaches based on common data models, several challenges had to be addressed to allow effective implementation.

Granularity Challenges

Finding the right representation for a concept is not trivial. Data can be represented in many ways (eg, "arm circumference" defined as a concept or a "circumference" concept connected with a "body site" concept taking the value "arm"), and agreeing on a common way to represent data is a challenging process. While both of these representations may be correct, interoperability is not always ensured if both are used, even though an international KR is used. This difference in the level of granularity also influences the way the user can query the data. When only one level of granularity is used in a specific data set, querying for relevant information is trivial. The user simply queries for the data of interest using the relevant defined concepts. However, if the data set comes from two different sources with different levels of granularity for the same type of information, either the querying needs to be adapted so that it can recognize both patterns or mapping must be performed beforehand to ensure that the results obtained are complete. Within the SPHN community, the granularity challenge has been addressed in the following two complementary ways: (1) when possible, a specific level is agreed by consensus and (2) in all other situations, all levels are encoded using a KR (for example SNOMED CT), allowing to query at different levels of granularity.

Different Needs

Defining a common concept for different use cases proved to be complex when creating the semantic framework. Depending on the project, needs may vary widely. For example, one project may require the temperature of a patient, without any information on the site or the device used to measure it, while another project may require the exact device and site for the temperature. This problem is addressed by representing the meaning strongly, therefore allowing the different concepts to be represented. Thus, it is possible to express temperature and many additional (present and future) concepts, and associate them freely. This is a major advantage when compared to any formal data model. When a concept requires further specification, it can be combined with other existing concepts (eg, body site and device) or extended by new project-specific properties.

Implementation Challenges

The process of clinical data acquisition passes through numerous filters before it ends up in a data warehouse for further usage. From acquisition of the data through questionnaires, formularies, texts, devices, etc in many different systems to the warehouse, several ETL processes usually will be required, resulting in loss of information. Therefore, the granularity and precision of the back-office semantic linkage can only represent the information richness known at that time. For example, the status "covid positive" cannot be coded in LOINC as this would require knowing the analytical method used by the laboratory. During that process, similar data in the data warehouse might originate from different contexts, which are not represented in the data warehouse. This is true within a care provider organization and is amplified when aggregating data originating from different care facilities and sources. These challenges were addressed in the strategy in several manners. The semantic framework with clear definitions of the concepts and their encoding in KR limited the ambiguity when creating the ETL procedures in the hospital. Second, the task of mapping the raw data to SPHN concepts was performed in each hospital by people knowing the internal data acquisition processes. Finally, the possibility to include relevant KR depending on the use case allowed the inclusion of relevant classifications used directly in care facilities, such as clinical, logistic, and billing classifications.

Resource Challenges

The creation, evolution, and management of these semantic descriptions raise several challenges, notably scalability and coherence. Since the data sets rely on multiple external standards, there is versioning required, especially because the data considered can cover decades. The same is true for the maintenance of KR created in the project and for the infrastructure and human resources that will handle the storage and transport layers in hospitals. Most hospitals did not know RDF before the SPHN strategy. Competencies had to be built internally to ensure the sustainability of local solutions. Adoption by care facilities has thus been a critical factor to

XSL•FO
RenderX

improve successful and sustainable implementation, with development of strategies for internal added value.

Competencies and Educational Challenges

The introduction of several new approaches in care facilities (semantic-centered data handling, formal descriptive language for storage and transport, and relegating data models to the end of the data pipeline) has been a huge challenge and still encounters resistances in the information technology (IT) community. Dedicated efforts in building several working groups for semantics, RDF, and data model bridging involving numerous hospital representatives have been important to handle this challenge. This was managed by the DCC, which gathered representatives from all stakeholders. The task of identifying the list of variables to be exchanged and their prioritization was given to the research projects.

The semantic framework is bound to evolve as the user base grows, and this evolution must follow the needs of projects without compromising the strategy. This will only be possible if the strategy is well understood both centrally and at the hospital level by specialists in medical informatics within IT departments. A strong effort is therefore currently underway within the SPHN to disseminate the strategy via the publication of strategic papers, webinars, and courses given to members of the SPHN community [50-52].

Conclusions

The main contribution of this work involves a new strategy for enabling nationwide intercommunity health data interoperability. The proposed strategy relies on the development of a semantic-based framework, which is designed to not replace existing standards but use them in a synergistic, pragmatic, and purpose-specific way. As the framework is built on the compositionality principle, it offers high flexibility and sustainability. The use of formal descriptive languages, such as RDF, as a data storage and transport layer ensures strong scalability to new needs. At the final stage, building specific bridges to fulfill the many data models used in research or required to comply with regulatory frameworks has proven successful and has been an important asset to ensure continuity of existing processes.

The wide adoption of the proposed strategy by every university hospital and high research organization in Switzerland as the national standard for sharing clinical data marks an important transition to an interoperable landscape for personalized health in Switzerland.

Acknowledgments

We would like to acknowledge the Swiss Personalized Health Network (SPHN) Clinical Data Semantic Interoperability, and the Hospital IT Working Groups and the resource description framework (RDF) Task Force, as well as all five university hospitals for their contribution to the implementation of the strategy. This work was funded by the Swiss Government through the SPHN Initiative.

Conflicts of Interest

CL is the Editor-in-Chief of this journal (JMIR Medical Informatics).

References

1. International Classification of Diseases, 11th Revision (ICD-11). World Health Organization. URL: http://www.who.int/classifications/icd/en/ [accessed 2021-06-09]
2. Health Level Seven International-Homepage. HL7 International. URL: http://www.hl7.org/ [accessed 2021-06-09]
3. Bhattacharyya SB. SNOMED CT History and IHTSDO. In: Introduction to SNOMED CT. Singapore: Springer; 2016.
4. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 05;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
5. Integrating the Healthcare Enterprise (IHE). IHE International. URL: https://www.ihe.net/ [accessed 2021-06-09]
6. Sternbach GL. The Glasgow Coma Scale. The Journal of Emergency Medicine 2000 Jul;19(1):67-71. [doi: 10.1016/s0736-4679(00)00182-7] [Medline: 10863122]
7. American Academy of Pediatrics, Committee on Fetus and Newborn, American College of Obstetricians and Gynecologists and Committee on Obstetric Practice. The Apgar score. Pediatrics 2006 Apr;117(4):1444-1447. [doi: 10.1542/peds.2006-0325] [Medline: 16585348]
8. What Are the Classifications of Heart Failure? Heart Failure. 2019. URL: https://heart-failure.net/classification [accessed 2021-06-09]
9. Schoonjans F, Zalata A, Depuydt CE, Comhaire FH. MedCalc: a new computer program for medical statistics. Comput Methods Programs Biomed 1995 Dec;48(3):257-262. [doi: 10.1016/0169-2607(95)01703-8] [Medline: 8925653]
10. Elovic A, Pourmand A. MDCalc Medical Calculator App Review. J Digit Imaging 2019 Oct;32(5):682-684 [FREE Full text] [doi: 10.1007/s10278-019-00218-y] [Medline: 31025219]
11. ICD-10 Version:2019. World Health Organization. URL: https://icd.who.int/browse10/2019/en [accessed 2021-06-09]
12. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633. [doi: 10.1373/49.4.624] [Medline: 12651816]
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]

XSL•FO
RenderX

14. The Unified Medical Language System (UMLS). National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html [accessed 2021-06-09]
15. GTIN Definition: Information. GTIN. URL: https://www.gtin.info/ [accessed 2021-06-09]
16. WHO Collaborating Centre for Drug Statistics Methodology-Home. WHOCC. URL: https://www.whocc.no/ [accessed 2021-06-09]
17. MedDRA. URL: https://www.meddra.org/ [accessed 2021-06-09]
18. Planned transition to concrete domains. SNOMED. 2020. URL: https://www.snomed.org/news-and-events/articles/planned-transition-concrete-domains [accessed 2021-06-09]
19. 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 Jan 26;23(1):e24594 [FREE Full text] [doi: 10.2196/24594] [Medline: 33496673]
20. Benson T, Grieve G. Implementing Terminologies. In: Principles of Health Interoperability. Health Information Technology Standards. Cham: Springer; 2016:189-219.
21. Benson T, Grieve G. The HL7 v3 RIM. In: Principles of Health Interoperability. Health Information Technology Standards. Cham: Springer; 2016:243-264.
22. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLoS One 2019;14(2):e0212463 [FREE Full text] [doi: 10.1371/journal.pone.0212463] [Medline: 30779778]
23. OMOP Common Data Model. OHDSI. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2021-06-09]
24. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
25. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. J Biomed Inform 2016 Apr;60:352-362 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.016] [Medline: 26944737]
26. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform 2005;115:153-173. [Medline: 16160223]
27. Thomas B. Archetypes: Constraint-based Domain Models for Futureproof Information Systems. CiteSeerX. 2000. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.1158 [accessed 2021-06-09]
28. Swiss Personalized Health Network (SPHN). URL: https://sphn.ch [accessed 2021-06-09]
29. First review report of the International Advisory Board. SPHN. 2019. URL: https://sphn.ch/2019/12/20/iab-report/ [accessed 2021-06-09]
30. Swiss Personalized Health Network. Report from the National Steering Board 2016–2019. Zenodo. URL: https://zenodo.org/record/4044123 [accessed 2021-06-09]
31. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C. URL: https://www.w3.org/TR/rdf-concepts/ [accessed 2021-06-09]
32. Samwald M, Fehre K, de Bruin J, Adlassnig K. The Arden Syntax standard for clinical decision support: experiences and directions. J Biomed Inform 2012 Aug;45(4):711-718 [FREE Full text] [doi: 10.1016/j.jbi.2012.02.001] [Medline: 22342733]
33. OWL Web Ontology Language Semantics and Abstract Syntax. W3C. URL: https://www.w3.org/TR/2004/REC-owl-semantics-20040210/ [accessed 2021-06-09]
34. Peleg M, Boxwala AA, Ogunyemi O, Zeng Q, Tu S, Lacson R, et al. GLIF3: the evolution of a guideline representation format. Proc AMIA Symp 2000:645-649 [FREE Full text] [Medline: 11079963]
35. jBPM. URL: https://www.jbpm.org/ [accessed 2021-06-09]
36. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H, Bürkle T, et al. Ontology-based data integration between clinical and research systems. PLoS One 2015;10(1):e0116656 [FREE Full text] [doi: 10.1371/journal.pone.0116656] [Medline: 25588043]
37. Stöhr MR, Majeed RW, Günther A. Metadata Import from RDF to i2b2. Stud Health Technol Inform 2018;253:40-44. [Medline: 30147037]
38. Solbrig HR, Hong N, Murphy SN, Jiang G. Automated Population of an i2b2 Clinical Data Warehouse using FHIR. AMIA Annu Symp Proc 2018;2018:979-988 [FREE Full text] [Medline: 30815141]
39. Pacaci A, Gonul S, Sinaci AA, Yuksel M, Laleci Erturkmen GB. A Semantic Transformation Methodology for the Secondary Use of Observational Healthcare Data in Postmarketing Safety Studies. Front Pharmacol 2018 Apr 30;9:435 [FREE Full text] [doi: 10.3389/fphar.2018.00435] [Medline: 29760661]
40. Data Coordination Center (DCC). SPHN. URL: https://sphn.ch/network/data-coordination-center/ [accessed 2021-06-09]
41. SPHN Dataset Release. SPHN. URL: https://sphn.ch/document/sphn-dataset/ [accessed 2021-06-09]
42. Decker S, Melnik S, van Harmelen F, Fensel D, Klein M, Broekstra J, et al. The Semantic Web: the roles of XML and RDF. IEEE Internet Comput 2000 Sep;4(5):63-73. [doi: 10.1109/4236.877487]
43. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol 2010 Sep;28(9):935-942 [FREE Full text] [doi: 10.1038/nbt.1666] [Medline: 20829833]

XSL•FO
RenderX

44. Antezana E, Blondé W, Egaña M, Rutherford A, Stevens R, De Baets B, et al. BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics 2009 Oct 01;10 Suppl 10:S11 [FREE Full text] [doi: 10.1186/1471-2105-10-S10-S11] [Medline: 19796395]
45. Lohmann S, Link V, Marbach E, Negru S. WebVOWL: Web-based Visualization of Ontologies. In: Lambrix P, Hyvönen E, Blomqvist E, Presutti V, Qi G, Sattler U, et al, editors. Knowledge Engineering and Knowledge Management. EKAW 2014. Lecture Notes in Computer Science, vol 8982. Cham: Springer; 2015:154-158.
46. Musen MA, Protégé Team. The Protégé Project: A Look Back and a Look Forward. AI Matters 2015 Jun;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
47. Protégé. URL: https://protege.stanford.edu/ [accessed 2021-06-09]
48. SPHN RDF Schema. SPHN. URL: https://sphn-semantic-framework.readthedocs.io/en/latest/sphn_framework/sphnrdfschema.html#technical-specification [accessed 2021-06-09]
49. SPHN RDF quality control. GitLab. URL: https://git.dcc.sib.swiss/sphn-semantic-framework/sphn-rdf-quality-control [accessed 2021-06-09]
50. SPHN Webinar Series. SPHN. URL: https://sphn.ch/services/seminars-training/ [accessed 2021-06-09]
51. Clinical Data Semantics Interoperability Working Group Strategy. SPHN. URL: https://sphn.ch/document/csi_wg_strategy/ [accessed 2021-06-09]
52. Fact sheet Semantic Strategy. SPHN. URL: https://sphn.ch/document/fact-sheet-semantic-strategy/ [accessed 2021-06-09]

Abbreviations

ATC: Anatomical Therapeutic Chemical
CDISC: Clinical Data Interchange Standards Consortium
DCC: Data Coordination Center
EHR: electronic health record
ETL: extract, transform, and load
GTIN: Global Trade Item Number
HL7: Health Level 7
i2b2: Informatics for Integrating Biology and the Bedside
ICD: International Classification of Diseases
IT: information technology
KR: knowledge representation
LOINC: Logical Observation Identifiers Names and Codes
NSB: National Steering Board
OMOP: Observational Medical Outcomes Partnership
RDF: resource description framework
RIM: Reference Information Model
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SPHN: Swiss Personalized Health Network

Edited by CL Parra-Calderón, G Eysenbach; submitted 29.01.21; peer-reviewed by G Jiang, A Rector, S Nelson; comments to author 22.02.21; revised version received 27.04.21; accepted 19.05.21; published 24.06.21

Please cite as:
Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C
A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study
JMIR Med Inform 2021;9(6):e27591
URL: https://medinform.jmir.org/2021/6/e27591/
doi: 10.2196/27591
PMID:

©Christophe Gaudet-Blavignac, Jean Louis Raisaro, Vasundra Touré, Sabine Österle, Katrin Crameri, Christian Lovis. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 24.06.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

7.2.2. Article 2

JMIR Publications
The Leading eHealth Publisher

One list to rule them all and many semantics to bind them: Building a shared, scalable and sustainable source for the problem oriented medical record

Christophe Gaudet-Blavignac, Andrea Rudaz, Christian Lovis

Submitted to: Journal of Medical Internet Research
on: March 29, 2021

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript 5
Supplementary Files 22
Figures 23
Figure 1 24
Figure 2 25
Figure 3 26

Preprint JMIR Publications

One list to rule them all and many semantics to bind them: Building a shared, scalable and sustainable source for the problem oriented medical record

Christophe Gaudet-Blavignac[1, 2] BSc, MSc; Andrea Rudaz[3] MD, MHSA; Christian Lovis[1, 2] MPH, MD, FACMI

[1]Division of Medical Information Sciences Geneva University Hospitals Geneva CH
[2]Department of Radiology and Medical Informatics University of Geneva Geneva CH
[3]Medical and Quality Directorate Geneva University Hospital Geneva CH

Corresponding Author:
Christophe Gaudet-Blavignac BSc, MSc
Division of Medical Information Sciences
Geneva University Hospitals
Rue Gabrielle-Perret-Gentil 4
Geneva
CH

Abstract

Background: Since the creation of the Problem Oriented Medical Record, the building of problem lists has been the focus of many researches. To this day, this issue is not well resolved, and building an appropriate contextualized problem list is still a challenge.

Objective: This paper presents the process of building a shared multi-purpose common problem list at the University Hospitals of Geneva, a consortium of all public hospitals and 30 outpatient clinics of the state of Geneva. This list aims at bridging the gap between clinicians' language expressed in free text and secondary usages requiring structured information.

Methods: The strategy focuses on the needs of clinicians by building a list of uniquely identified expressions to support their daily activities. In a second stage, these expressions are connected to additional information, building a complex graph of information. A list of 45,946 expressions manually extracted from clinical documents has been manually curated and encoded in multiple semantic dimensions, such as ICD-10, ICPC-2, SNOMED-CT or dimensions dictated by specific usages, such as identifying expressions specific to a domain, a gender, or an intervention. The list has been progressively deployed for clinicians with an iterative process of quality control, maintenance and improvements, including addition of new expressions, or dimensions for specific needs. The problem management of the electronic health record allowed to measure and correct the encoding based on real-world usage.

Results: The list was deployed in production in January 2017 and was regularly updated and deployed in new divisions of the hospital. In 4 years, 684,102 problems were created using the list. The proportion of free text entries reduced progressively from 37.47% (8,321/22,206) in December 2017 to 18.38% (4,547/24,738) in December 2020.
In the last version of the list, over 14 dimensions were mapped to expressions, among them 5 international classifications and 8 other classifications for specific usages. The list became a central axis in the EHR, being used for many different purposes linked to care such as surgical planning or emergency wards, or in research, for various predictions using machine learning techniques.

Conclusions: This work breaks with common approaches primarily by focusing on real clinicians' language when expressing patient's problems and secondly by mapping whatever is required, including controlled vocabularies to answer specific needs. This approach improves the quality of the expression of patients' problems, while allowing to build as many structured dimensions as needed to convey semantics according to specific contexts. The method is shown to be scalable, sustainable and efficient at hiding the complexity of semantics or the burden of constraint structured problem list entry for clinicians. Ongoing work is analyzing the impact of this approach at influencing how clinicians express patient's problems.

(JMIR Preprints 29/03/2021:29174)
DOI: https://doi.org/10.2196/preprints.29174

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint JMIR Publications

Original Paper

One list to rule them all and many semantics to bind them: Building a shared, scalable and sustainable source for the problem oriented medical record.

Abstract

Background: Since the creation of the Problem Oriented Medical Record, the building of problem lists has been the focus of many researches. To this day, this issue is not well resolved, and building an appropriate contextualized problem list is still a challenge.

Objective: This paper presents the process of building a shared multi-purpose common problem list at the University Hospitals of Geneva, a consortium of all public hospitals and 30 outpatient clinics of the state of Geneva. This list aims at bridging the gap between clinicians' language expressed in free text and secondary usages requiring structured information.

Methods: The strategy focuses on the needs of clinicians by building a list of uniquely identified expressions to support their daily activities. In a second stage, these expressions are connected to additional information, building a complex graph of information. A list of 45,946 expressions manually extracted from clinical documents has been manually curated and encoded in multiple semantic dimensions, such as ICD-10, ICPC-2, SNOMED-CT or dimensions dictated by specific usages, such as identifying expressions specific to a domain, a gender, or an intervention. The list has been progressively deployed for clinicians with an iterative process of quality control, maintenance and improvements, including addition of new expressions, or dimensions for specific needs. The problem management of the electronic health record allowed to measure and correct the encoding based on real-world usage.

Results: The list was deployed in production in January 2017 and was regularly updated and deployed in new divisions of the hospital. In 4 years, 684,102 problems were created using the list. The proportion of free text entries reduced progressively from 37.47% (8,321/22,206) in December 2017 to 18.38% (4,547/24,738) in December 2020.
In the last version of the list, over 14 dimensions were mapped to expressions, among them 5 international classifications and 8 other classifications for specific usages. The list became a central axis in the EHR, being used for many different purposes linked to care such as surgical planning or emergency wards, or in research, for various predictions using machine learning techniques.

Conclusions: This work breaks with common approaches primarily by focusing on real clinicians' language when expressing patient's problems and secondly by mapping whatever is required, including controlled vocabularies to answer specific needs. This approach improves the quality of the expression of patients' problems, while allowing to build as many structured dimensions as needed to convey semantics according to specific contexts. The method is shown to be scalable, sustainable and efficient at hiding the complexity of semantics or the burden of constraint structured problem list entry for clinicians. Ongoing work is analyzing the impact of this approach at influencing how clinicians express patient's problems.

Keywords: medical records, problem-oriented; electronic health records; semantics

Introduction

Background

The concept of a Problem Oriented Medical Record (POMR) is as old as 1968 [1]. One of the key elements of this approach is a list of relevant problems, current or past, important to understand the

patient's condition. A problem can be anything, complaints, symptoms, existing or previous conditions, diagnosis, procedures, socio-economic issues, etc. This list is the corner stone of clinician's education and of the patient record. It is used from the first encounter, where it is named "chief complaint", to drive clinical reasoning, but increasingly to support electronic decision support and diagnostic or care pathways. With the widespread adoption of Electronic Health Records (EHRs) and since the Meaningful Use Act established the problem list as a requirement for care facilities [2,3], it has been the focus of multiple research and improvements. However, its digitization brought new opportunities and challenges. Problem lists vary in time, are influenced by the conditions of the population a care facility deserves or the specialties covered.

Creation and evolution

The building of a problem list can be driven by free-text entries made by the clinicians or by the creation of a finite list of items they can choose from. Terms included in those premade lists are often taken from an existing terminology. Compared to the use of free text, a premade list allows for more structured data and easier secondary use [4–6]. The usage of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [7,8] for the problem list seems to provide good coverage [9–11]. In 2009, the Clinical Observations Recording and Encoding (CORE) Problem List Subset of SNOMED CT was created using data from 8 institutions [12]. The terms extracted from those institutions were mapped to SNOMED CT concepts to create a subset useable as a problem list. Its current version contains 6,565 SNOMED CT concepts. Other approaches have been explored such as the automatic generation of a patient's problem list using NLP and international terminologies, but with lists of less than 200 problems focused on a clinical specialty [13–17]. When creating a problem list, the equilibrium between a list representing what care professionals need to express and an interoperable controlled vocabulary is hard to find [18].

Using a terminology such as the International Classification of Diseases, Tenth Revision (ICD-10) [19] as a source of expressions for a problem list can bring multiple issues. A classification is a partition of reality in a finite set of categories, resulting in a phenomenon called residual aggregation or residual category. For example: "Other specified immunodeficiencies", "Disorder of pancreatic internal secretion, unspecified", or even "Fracture of unspecified phalanx of other finger " exist in ICD-10 to cover all concepts that don't fall in another category [20]. This type of terms is not suited for a problem list because they do not represent problems that clinicians could reasonably enter.

Another challenge at using a classification as a source is based on its organization, tightly connected to the "intention", or "usage" which supported its development. For example, ICD-10 aims to properly express causes of deaths and morbidity, the International Classification of Primary Care, 2nd edition (ICPC-2) [21] focus on primary care problems and the Logical Observation Identifiers Names & Codes (LOINC) [22] covers observations and laboratories. Each of these have thus a specific structure, and a dedicated organization of their hierarchy to answer the requirements of their usage.

The problem list should be able to represent any of those "intentions", regardless of their future interpretations according to specific classificatory intentions, and without restricting elements to only one classification nor requiring clinicians to know the organization of all of them. A hierarchy such ICD-10 results in choices that will favors some dimensions over others. As an example, there is no infectious diseases chapter in ICD-10, which complicates seriously the identification of infectious diseases. As a consequence, our approach focuses on using real-world clinician's expressions as primary source, and then manually adding as many "semantically meaningful" dimensions as needed.

Maintenance and update of problem lists are also challenging. For example, during the current pandemics, it was suddenly required to add several new entries to express the specific spectrum of COVID-19. Such rapid adaptations of the list must be rapidly implementable and should not depend

on the update cycle of an international classification.
Using efficiently a problem list requires a lot of background information. For example, the same problem can be added several times. This is sometimes appropriate, such as repeated fractures, sometimes inappropriate such as repeating at each encounter that the patient has hypertension. Describing properly the semantics allows to facilitate and speed up the work of clinicians [23,24]. Semantic dimensions should support recognition and reconciliation algorithms, different views of the list, by specialty, organ, severity, to name a few [25,26] or to support graph-based, symbolic, machine learning or clustering algorithms to group concepts along a navigation that answers the needs of clinicians, case managers, researchers, etc. [27].

Implementation and adoption

Although the advantages of a well-maintained problem list are clear, numerous issues have been raised in the way it should be implemented. Engaging the users into documenting a list of problems for their patients in a complete and efficient manner is a challenge. Clinicians in hospitals are under constant pressure and the effort to pivot from a free-text problem list to a dedicated EHR module can be important. Factors such as gap reporting, problem-oriented charting or links to billing codes have shown some positive impact on the completeness of the list documented by clinicians [23,28]. Additionally, training and education seems to be a key factor for adoption [29–31]. In 2016, Simons et al. proposed a list of determinants for the successful implementation of a POMR [32]. It includes, among others, completeness, interoperability, usability and training of staff.
In this paper we aim to address the challenges of building and implementing a shared multi-purpose common problem list at HUG with an approach based on clinician's language and semantic dimensions encoding. The driving concepts of this work is that the content of the list should be created with the care professionals to match their needs, and that the list should be mapped to terminologies to a) improve adoption, with metadata for completers, and b) for secondary use of data [4,33]. After the description of the building, implementation and iterative improvements of the list, an analysis of its usage over 4 years is presented.

Methods

Approach

This approach focuses on two goals. First, allowing clinicians to express themselves as freely as possible with a list representing the language used every day in clinical interactions and working with a free-text completer rather than a constrained closed list. Second, use of the list for multiple purposes in the hospital other than medical care. The latter is done by back-office multi-dimensional extension of metadata of the free-text expressions.

Common list creation

To represent the language of the clinicians, the starting point are sentences expressing problems written by clinicians. The initial list has been created based on 40'000 discharge and admissions summaries. All sentences were extracted using automatic tools, and then manually selected if they represented a potential candidate. These have been curated for typos and grammatical normalization such as plural or uppercase reserved to proper names. Abbreviations have been expanded but kept. Rules applied to build this list were inclusive, covering problems of any type including but not limited to medical, surgical, socio-economic, psychologic, logistic, etc. Synonymy is allowed, so that multiple expressions expressing the same problem are present, such as "generalized pain", "pain everywhere" but connected as synonyms. Every granularity is allowed as long as the expression has been used by clinicians. The only strict rule is that an expression must be syntactically and morphologically unique.

The list of expression is improved based on two axes: vertical (expressions) and horizontal (dimensions). Extensions of the list require to allow the deployment in a specific clinical context, for example neurosurgery. In this case, discussion with clinicians and analysis of their clinical documents allows to build a set of specific expressions for that context and these are added to the common list prior to the deployment. Adjustments of the list are also iteratively made based on usage, helped by the fact that the problem list management module is based on a syntactic completer allowing clinicians to enter free text, and then select an expression if appropriate, or keep the original free text. The modifications of the list, expressions, activity state of expressions are fully historicized based on usage. Deletions are usually forbidden, which happened only once after a one-year evaluation of the impact of deleting entries: ensuring they had never been used and the impact of their absence on tools such as completers, parsers and co-locations, word embedding, etc.
A monthly usage analysis with all expressions chosen, by whom, in which context, and the potential free text added is used to improve the list.

Semantic dimensions

We consider a semantic dimension as any metadata added to the list of expressions to improve its usage for a specific purpose. This purpose can be the completer functionalities, for example for ambiguous abbreviations (in French, *TV* can mean *tachycardie ventriculaire* or *toucher vaginal*), or when the expression is gender specific, such as all expressions relating to *Prostata*. Some dimensions are related to national classifications, such as the Swiss classification for surgical interventions (CHOP) [34], or international ones, such as ICPC-2 or ICD-10, including their various versions (several releases of ICD 10 for example). Finally, some are internal to the organization, such as a specific identification for surgery requiring a surgical theatre, with the resources needed, used for the surgical theatre logistics and management at HUG. Expressions can have no or several entries in any specific dimensions.
Encoding has been made by domain experts. For example, the ICD-10 and CHOP classifications have been made by a coding expert of the billing division of HUG, SNOMED-CT encoding by a physician, ICPC-2 encoding by an outpatient physician, etc. Several dimensions, such as chronic/acute, gender specificity or syntactical dimensions have been done by medical students.
Dimensions described here are not exhaustive, but representative. The coding of the dimensions is a complex activity, mostly towards keeping a global coherence. In this work, the strategy is to have a clear definition of a dimension and aim at reaching the best quality of representation of that dimension, regardless of the others except the expression itself. The objective being that a specific expression that can be represented in that dimension, must be represented with the highest precision possible for that dimension, respecting only the rules specific to it. This strategy has many advantages. It allows to keep the intention of the dimension being coded at best and allows the encoding work to be distributed among several actors, domain experts or students, according to their competences specific to that dimension and their understanding of the expression. At the end, a specific expression can be understood differently and with a different granularity, according to the perspective of the dimension used, or seen as the sum of some or all dimensions.

General classifications

International classifications:

ICD-10 is the basis for billing of inpatient stays in Switzerland. Once every two years, the Swiss Confederation publishes its own version of the ICD-10 classification which is a translation of the ICD-10 German Modification (ICD-10 GM) which is a slightly modified translation of the ICD-10 released by the World Health Organization [35]. Every expression in the list was first encoded with the ICD-10 WHO version to evaluate gaps in the list and perform subset definitions for specific use-cases. Secondly the list was encoded into the Swiss ICD-10 GM version. The encoding was

performed using the official coding rulebook for hospital stays in Switzerland [36]. This dimension has been added in the aim of performing automatic coding of inpatient stays for billing, prediction tools on problems versus diagnosis or support of pathways. Several versions of ICD-10 are encoded, according to years, or to the source, including the WHO original ICD-10.

ICPC-2 is a classification used to encode general practice clinical activities and primary care. It belongs to the WHO Family of International Classifications [21]. This classification was chosen by the clinicians from the outpatient clinics for its ability to classify problems in simple categories relevant for care such as symptoms, diagnosis, screenings or procedures. Beside the activities of outpatient clinics, including research, this classification is used to generate alerts when adding multiple problems with the same ICPC-2 encoding.

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a terminology of more than 340,000 concepts and 1 million relationships [7,37]. It is described as the most comprehensive clinical healthcare terminology in the world and has become central for semantic interoperability. It has been chosen to be the United States standard for encoding of diagnoses and problem lists [38]. SNOMED CT includes powerful features such as the combination of concepts (post-coordination) or the expression constraint language which can be used to perform complex queries on SNOMED CT encoded data. SNOMED-CT is one of the pillars for the semantically driven activities for data science at HUG and allows to connect many different aspects of the EHR, such as problem lists, formularies, and other structured data. It allows complex queries such as, every problem related to an organ, or including an inflammatory process. Due to the size and complexity of the terminology, the encoding of tens of thousands of expressions in SNOMED CT requires a significant amount of time and experience. The encoding of the expressions is using only single or multiple pre-coordinated elements, a step towards fully post-coordinated expressions.

National classifications:

The Swiss classification for surgical interventions (CHOP) [34] is used to encode and bill surgical interventions. It was added with similar goals as the ICD-10 GM. Every expression in the list that can be mapped to a CHOP code was mapped and updates implemented annually when new version is released.

Internal classifications

Several internal classifications are used in specific contexts, which are illustrated hereafter.

Department/specialty specific lists: Adult and pediatric emergency departments are using specific problem lists, which were included in the process. Most of the time, these lists are derived from ICD-10. The appropriate dimensions were added, including the specialty preferences. The adaptations are systematically validated by experts of the specialty. The same process has been applied with several specialty, such as oncology, neurosurgery, etc.

Clinical decision support: Some expressions and dimensions have been added specifically to support computerized provider order entry, exemplified with the *Antibiotic prescription support* to improve choice of antibiotics, monitor and lower antibio-resistances. The expressions related to that list were added, properly encoded, and their belonging to problems related to antibiotic prescription added in a new dimension, so that it could be used in several modules of the EHR.

Surgical intervention list. One key development enabled the usage of the list as the unique source of expressions for surgical intervention planning and documenting. When an intervention is planned in the hospital, it triggers a chain of events that will lead to the intervention. Operating room must be booked, staff must be appointed, specific devices and materials must be ordered etc. This process historically was separated in silos, medical, paramedical or logistic with separated lists. The list of surgical interventions used for the operating room planning was manually integrated into the common list as a new dimension. This integration was made by specialties and is still ongoing. It allowed the common list to become the single source of expressions for surgical intervention

planning.

Nutrition and Dietetic diagnoses list. The most recent development of the list focused on the diagnoses used by the dieticians and nutritionists, which was a list of expressions extracted from the Terminologie Internationale de Diététique et de Nutrition (TIDN) [39]. These expressions were curated and integrated as a new dimension, making the common list the single source of expressions for the nutritionist and dieticians of the hospital.

Other dimensions

Other specific dimensions are useful for numerous purposes. The gender specificity dimension defines if an expression is gender-specific, such as: "vasectomy". The intervention dimension defines that there is an act performed and differentiates from interventions requiring surgery theatre. Multiple other dimensions are used for numerous purposes such as possible abbreviations of the expression, preferred terms, chronic/acute, etc.

Language

The expressions being in French, an English translation was made, and keywords of the expressions added in both French and English.

Results

Evolution of the list

The list of problems presented in this work had to "compete" with 17 specific, specialty vertical, local problem lists and was proposed as an additional choice to clinicians. They could freely choose between their "usual" lists and the new one. This competitive approach was a strong incentive to stick to the needs of clinicians and become their "preferred" list. Within the first year, the new list became the most used in most cases and the legacy lists were then removed. The two first years required frequent adjustments, but with a slowing down frequency up to the current situation which is on specific demand, such as covid-19, or monthly. Table 1 summarizes the major releases.

Table 1. Major releases, corpus size and comments

Date of release	Number of active problems	Modifications
January 2017	45,946	First production deployment.
September 2017	45,458	Partial integration of expression for surgery planning. Corrections of expressions.
January 2018	51,255	5,867 expressions created from legacy list usage and free-text entries.
February 2018	50,822	Integration of expressions for antibiotics prescription and monitoring project. Corrections of expressions.
May 2018	52,040	1,091 expressions created from legacy list usage and free-text entries.
November 2018	52,211	Integration of the list for adult emergency ward. Abbreviations system integration.
August 2019	51,824	310 expressions created on demand from users.
January 2020	52,956	Integration of expressions for surgery planning. Integration of a list of diagnoses used by dieticians and nutritionists. Integration of the list for pediatric emergency ward.
April 2020	52,958	Emergency adding of 2 expressions for SARS Cov2

		cases.
August 2020	20,120	Inactivation of 32,840 never used expressions. Preferred term system integration.

In January 2017, the list was deployed in the geriatric and general pediatric division of HUG as well as part of the rehabilitation medicine division and ambulatory primary care division. The list was then progressively deployed in new divisions. Table 2 summarize those deployments.

Table 2. Deployment of the list in new divisions by date

Date	**Division**
April 2017	Neurosurgery
May 2017	Neurology
May 2017	Visceral surgery
November 2017	Psychiatry (adult and pediatric)
November 2018	Rehabilitation
September 2019	Adult emergency
September 2020	Internal medicine
September 2020	Oncology
September 2020	Cardiology

For us, an important success indicator is that currently, three major divisions, internal medicine, geriatrics and rehabilitation, decided to remove free-text entry possibility, judging that the common list was sufficiently complete for their usage.

Table 3. Some descriptive statistics of the list

Type of expression or encoding	**Expressions (N= 20,120), n (%)**
Active expressions	20,120 (100)
Abbreviations	2,127 (10.57)
ICPC-2 encoding	20,120 (100)
ICD10 WHO 2008 encoding	11,860 (58.95)
ICD10 GM 2018 encoding	18,481 (91.85)
CHOP 2019 encoding	1,223 (6.08)
SNOMED CT encoding	9,222 (45.83)
Gender specificity encoding	805 (4)
Acute/chronic specificity encoding	8,013 (39.83)
Intervention encoding	1,855 (9.22)
Surgery planning	985 (4.9)
Antibiotic decision support	553 (2.75)
Adult emergency ward	1,108 (5.51)
Pediatric emergency ward	939 (4.67)
Nutrition and dietetics	139 (0.69)

In 4 years, 7,270 expressions were added from legacy lists, free-text or on users' request. After 3 years of usage, all 32,840 expressions that were never used, nor linked to any specific project, were inactivated from the source and deleted for the production. The current version of the list contains 20,120 active expressions. Evolution of the number of expressions in the list is shown in Figure 1.

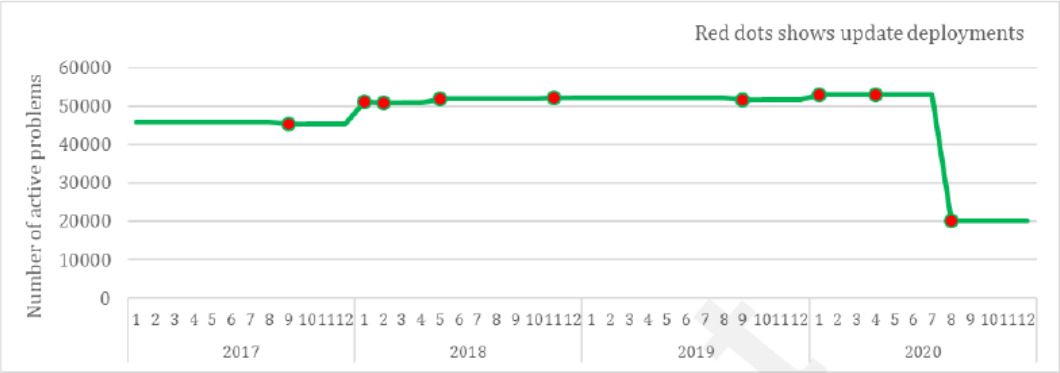


Figure 1 Evolution of the number of active expressions in the common list.

Usage of the list

After 4 years of usage, all problems created were extracted from HUG's data warehouse, representing 1,146,135 problems creations. Among them 59.69% (684,102/1,146,135) were chosen from the common list, 14.83% (169,970/1,146,135) from legacy lists and 25.48% (292,063/1,146,135) entered as free-text entries. Over the legacy list problems, 63.01% (107,095/169,970) were created during the first year. In December 2017, the month with the biggest free-text entries proportion, 37.47% (8,321/22,206) of the problems were created using this method. In December 2020, the last month of the observation period, 18.38% (4,547/24,738) of the problems were created using free text and 80.18% (19,836/24,738) using the list.
From the common list, 15,232 distinct expressions were used at least once. Figure 2 shows the absolute number of problems created by month and their origin. Legacy lists combine all problems coming from the 17 legacy lists in production in HUG at the time of the first deployment and that are progressively abandoned.

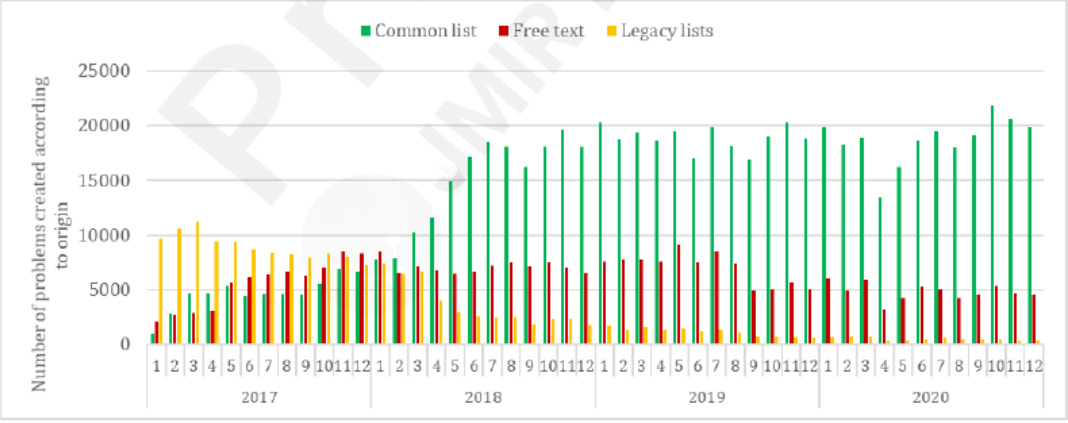


Figure 2 Number of problems created by month according to their origin.

Figure 3 displays the proportion of problems chosen in the common list versus legacy lists and free-text entries.

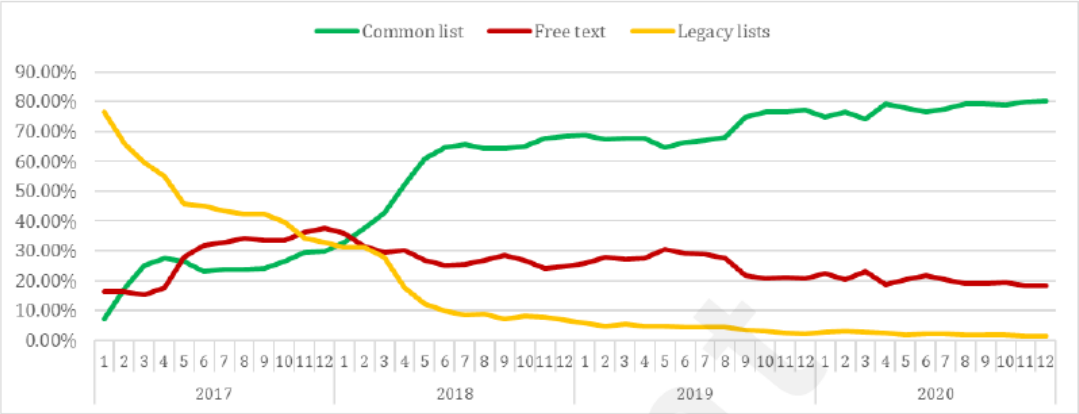


Figure 3 Proportion of problems created by month according to their origin.

The 20 most frequently used expressions in the common list are displayed in Table 4. Free-text entries and expressions from legacy lists are not included.

Table 4. The 20 most frequently used expressions of the common list over 4 years.

Expression	English translation	Number of uses
hypertension artérielle	arterial hypertension	16,974
insuffisance rénale aiguë	acute renal failure	6,391
hypercholestérolémie	hypercholesterolemia	5,219
accouchement normal d'un nouveau-né vivant par voie basse	normal vaginal delivery of a liveborn	5,045
appendicectomie	appendicectomy	4,550
hypertension artérielle traitée	treated arterial hypertension	4,230
décompensation cardiaque	cardiac decompensation	4,159
douleur thoracique	thoracic pain	3,707
hypokaliémie	hypokalemia	3,363
troubles cognitifs	cognitive disorder	3,323
hyponatrémie	hyponatremia	3,212
infection à SARS-CoV-2 (COVID19)	SARS-CoV-2 (COVID19) infection	3,118
fibrillation auriculaire	atrial fibrillation	3,055
diabète type 2	type 2 diabetes	3,051
insuffisance rénale chronique	chronic renal failure	3,002
malnutrition protéino-énergétique grave	serious protein-energy malnutrition	2,898
dyslipidémie	dyslipidemia	2,844
obésité	obesity	2,833
asthme	asthma	2,756
douleur abdominale	abdominal pain	2,749

Finally, the list was exploited for various research activities, training machine learning models using the various mappings, predicting billing codes of a stay using the ICD-10 encoding or for workload

predictions during the multiple waves of the pandemic. This work will not be discussed in this paper.

Discussion

Principal results

After 4 years of deployment and iterative improvements, a list of 20,120 active expressions mapped to more than 14 semantic dimensions was deployed in most major divisions of HUG and used for the creation of 684,102 new problems. Specific dimensions allowed the list to be used for various purposes such as surgical planification, decision support or nutrition and dietetic diagnosis.

Manually building a problem list is a time-consuming task and starting from the clinician language as a source of expressions is a double-edged sword. It aims at improving information precision to support clinicians to find easily and fast the most appropriate expressions representing best the conditions of patients. As a cost, it tends to increase noise at proposing numerous expressions with small variation, either syntactically, semantically or both, depending on the completer used. This effect can be mitigated using the features of the dimensions, either syntactically such as abbreviations and common variants, or semantically by using the aggregation properties of classificatory dimensions. This allows to search the entirety of the list while reducing the number of possibilities proposed to the most pertinent set. These tools will not be discussed in this work. Finally, statistics on the use of the list are important to improve it, such as progressively filtering out never used expressions or improving granularity in existing ones that are extended by free texts, for example.

We noticed that problems appearing frequently in practice tend to have multiple variations, with various levels of granularity or additional information, improving naturally the expressiveness of the list and the ease for clinicians to find the most appropriate element. On the other hand, rare problems tend to have less representations, if any in the list thus reinforcing the need to keep free text entries, to our opinion.

The many dimensions that have been encoded allow to compare the list of expressions and its coverage for the respective coverage of classifications. For example, taking ICPC-2 and ICD-10, the immediate observation is that the list contains elements that can be expressed in both classifications, but in many more lexical variants. On the other hand, many expressions of the classifications are not found in the list, for many of them not codable elements or unmet conditions in our setting. As a result, the list covers more than any of the classification separately, but only meaningful expressions. Moreover, it frees the care professionals from the task of knowing multiple classifications and their structures. This reduces the compression of information while keeping strong interoperable capabilities through the semantic dimensions.

The semantic dimensions are the major addition of this approach. They allow to bridge the need for various representations of a concept as expressed by the clinicians with the need for semantic interoperability. By encoding each expression into all relevant dimensions, it was possible to reuse the created problems for other goals, for example: extracting subsets related to a specific disease through the ICD-10 encoding, all patients that undergo a specific procedure using the CHOP encoding or more complex queries such as all problems that include an inflammation process through the SNOMED CT encoding. However, the maintenance cost of those dimensions is important. The more dimensions there is the more work it requires to add a new expression, since it must be encoded in possibly all of them. Moreover, classifications updates (such as a new version of ICD-10) sometimes require a full reading and update of the encoding.

The semantic dimensions linked to intra hospital use cases allowed the list to be used for multiple projects. Specific subsets for divisions such as the emergency wards were beneficial for convincing users to start using the common list. The surgical planning addition promoted the list as a central source of expressions and concept outside of the care domain. The role of the list as a central source of expressions for patient's problems is shown by the number of projects that included the addition of

a dimension to the list. In a sort of virtuous circle, the more the list was known, the more demands were made to adapt it to new needs.

As every project of this type, the final challenge is to convince users to use the module and teach them how to do so correctly. This has been heavily pushed in this work by the Medical and Quality Directorate, the team designing the problem list module in HUG. Teaching both in person and through videos helped disseminate the usage of the module in divisions that historically did not use it.

During the first year of deployment, the module was introduced and promoted in 4 new divisions of the hospital. This increased the number of users and the number of problems created. Those new users with no experience of the problem module are arguably the reason for the initial augmentation in the proportion of free-text entries seen in Figure 3. The diminution in problems created from legacy list are to put on the account of the progressive removal of those lists from the module. After this initial period, the proportion of free text diminishes progressively from 37.47% (8,321/22,206) in December 2017 to 18.38% (4,547/24,738) in December 2020, the lowest percentage on the full period. It is interesting to note that this period of one year also correspond to the time it took for the common list to become the most used method for creating problems.

This reduction of the proportion of free-text entries shows that the common list corresponds to the needs of the care professionals and that its adoption is progressing. While it is not possible to determine in which proportion this evolution is due to the content of the list, the functionalities of the problem module or the dissemination effort, it seems likely that it is a combination of the three and that only a transversal approach could succeed in this transition.

The situation before the deployment of the common list could seem preferable because the proportion of free-text entries was low and the usage of legacy list well established. However, the final situation is arguably better for several reasons. First, the legacy lists lacked proper semantic interoperability. They were manually modified version of existing classifications, with the limitations described before and the added complexity of manual, unverified modifications. They were not harmonized, and it was not possible to group or analyze problems from multiple lists without manual reading of the expressions. This prevented those lists of being used for other purposes as the common list allows.

The apparent decrease of the number of problems created in April and May 2020 is explained by the COVID-19 pandemic. Indeed, the HUG stopped their elective activity and shifted to treating only COVID-19 patients which reduced the number of patients with various problems and reduce the overall number of problems created.

Lessons learned

This work allowed to draw significant learnings for the building and implementation of a problem list. They are listed in Textbox 1.

Textbox 1: Key learnings

- Existing controlled vocabularies are too narrow or subject oriented to be used natively as problem lists.
- It is possible to build a problem list starting from the clinician's language to better match their needs.
- It is possible to reduce the expressivity needed for a problem list to a meaningful set of expressions used in practice.
- On purpose semantic dimension encoding allows secondary use of data.
- Internally building a list of expressions allows flexibility and quick adjustments when needed.

Limitations

Although the data and analysis included in this work were carefully carried out, some limitations worth noting. First, the evaluation data are analyzed as one source of problems created. However, this does not translate the complexity of the deployment of the list in the hospital. Indeed, the problem module is deployed in the EHR globally, but some divisions use it, and some do not. Inside those divisions some teams of residents are more used to the module than others. Additional data should be gathered to track the dissemination effort, the training provided, to understand when the module was adopted in which division and by whom.
Finally, the proportion of common list, free text and legacy list problems is only a proxy for the preferences of the users. It does not account for other elements such as division specific guidelines or orally transmitted habits. To credit the progression of the common list to its quality is a conclusion that should be confirmed by a closer evaluation, in partnership with the users.

Conclusion

Overall, there is still room for improvement when building and implementing a problem list into the production environment of care. Most of the existing efforts use terms from existing terminologies rather than focusing on the language used by clinicians. The perfect problem list that contains what the care professionals want and can be used for every other use-case is yet to be created.
The proposed approach breaks with common approaches for the building of problem lists by directly addressing the gap between existing controlled vocabularies and real clinicians' language when expressing patient's problem. Secondly it brings new perspectives for secondary use by encoding the expressions in various semantic dimensions, allowing specific usages of the list in the hospital and beyond.
By applying this approach, more than 50,000 expressions were manually curated into a common problem list integrated in the EHR. By iterative updates, the list was enriched and refined to 20,120 active expressions matching users' needs. More than 14 semantic dimensions were added to the list including 5 major classifications, and multiple dimensions internal to the hospital such as division specific adaptations, surgical planning, antibiotics prescription support, nutrition and dietetic diagnoses, etc. Those additions pushed the adoption of the common list as a central, harmonized source of expressions in the hospital. The recent decision of 3 major divisions of the hospital to remove the option to make free-text entries shows that the list correspond to the needs of the users.
Manually creating and updating a set of expressions directly extracted from clinical document has succeeded in HUG to engage users in transitioning from legacy systems to a new module including the common list. The overall number of problems created is increasing while the problems entered as free text are decreasing.
The manual work required to build and maintain the list is substantial in the three domains, maintenance of the expressions, development of the problem module and dissemination of its usage. However, this approach brings a solution for keeping data interoperable while not constraining the user and allowing multiple use cases. An evaluation of the impact of the list on the workload of clinicians and on the secondary uses of the produced data should be made to further validate the approach.

Acknowledgements

This project has been partly funded by the Evolving Language National Centre of Competence in Research of the Swiss National Fund, NCCR N-603-11-01.

Conflicts of Interest

None declared.

Abbreviations

EHR: Electronic Health Records
HUG: University Hospitals of Geneva
ICD-10: International Classification of Diseases, 10th revision
POMR: Problem Oriented Medical Record
SNOMED CT: Systematized Nomenclature of Medicine, Clinical Terms

References

1. Weed LL. Medical Records That Guide and Teach. New England Journal of Medicine Massachusetts Medical Society; 1968 Mar 14;278(11):593–600. PMID:5637758

2. Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [Internet]. [cited 2021 Jan 4]. Available from: https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf

3. Holmes C. The problem list beyond meaningful use. Part I: The problems with problem lists. J AHIMA 2011 Feb;82(2):30–33; quiz 34. PMID:21337850

4. Group AW. Problem List Guidance in the EHR. Journal of AHIMA American Health Information Management Association; 2011 Sep;82(9):52–58.

5. Elkin PL, Mohr DN, Tuttle MS, Cole WG, Atkin GE, Keck K, Fisk TB, Kaihoi BH, Lee KE, Higgins MC, Suermondt HJ, Olson N, Claus PL, Carpenter PC, Chute CG. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. Proc AMIA Annu Fall Symp 1997;500–504. PMID:9357676

6. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. Proc AMIA Symp 1998;795–799. PMID:9929328

7. SNOMED International SNOMED CT Browser [Internet]. [cited 2021 Jan 6]. Available from: https://browser.ihtsdotools.org/?

8. SNOMED Home page [Internet]. SNOMED. [cited 2021 Jan 7]. Available from: https://www.snomed.org/

9. 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;2003:699–703. PMID:14728263

10. Penz JFE, Brown SH, Carter JS, Elkin PL, Nguyen VN, Sims SA, Lincoln MJ. Evaluation of SNOMED coverage of Veterans Health Administration terms. Stud Health Technol Inform 2004;107(Pt 1):540–544. PMID:15360871

11. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, Speroff T. 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. PMID:16770974

12. The CORE Problem List Subset of SNOMED CT® [Internet]. U.S. National Library of

Medicine; [cited 2021 Feb 25]. Available from: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html

13. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decis Mak 2005 Aug 31;5:30. PMID:16135244

14. Solti I, Aaronson B, Fletcher G, Solti M, Gennari JH, Cooper M, Payne T. Building an Automated Problem List Based on Natural Language Processing: Lessons Learned in the Early Phase of Development. AMIA Annual Symposium Proceedings American Medical Informatics Association; 2008;2008:687. PMID:18999050

15. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform 2008 Sep;77(9):602–612. PMID:18280787

16. Meystre S, Haug P. Improving the sensitivity of the problem list in an intensive care unit by using natural language processing. AMIA Annu Symp Proc 2006;554–558. PMID:17238402

17. Meystre SM, Haug PJ. Comparing natural language processing tools to extract medical problems from narrative text. AMIA Annu Symp Proc 2005;525–529. PMID:16779095

18. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. Proc AMIA Symp 1998;280–284. PMID:9929226

19. ICD-10 Version:2019 [Internet]. [cited 2021 Jan 7]. Available from: https://icd.who.int/browse10/2019/en

20. Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. Journal of Biomedical Informatics: X 2019 Jun 1;2:100002. [doi: 10.1016/j.yjbinx.2019.100002]

21. WHO | International Classification of Primary Care, Second edition (ICPC-2) [Internet]. WHO. World Health Organization; [cited 2020 Nov 25]. Available from: http://www.who.int/classifications/icd/adaptations/icpc2/en/

22. LOINC from Regenstrief [Internet]. LOINC. [cited 2021 Jan 7]. Available from: https://loinc.org/

23. Wright A, McCoy AB, Hickman T-TT, Hilaire DS, Borbolla D, Bowes WA, Dixon WG, Dorr DA, Krall M, Malholtra S, Bates DW, Sittig DF. Problem list completeness in electronic health records: A multi-site study and assessment of success factors. Int J Med Inform 2015 Oct;84(10):784–790. PMID:26228650

24. Wang EC-H, Wright A. Characterizing outpatient problem list completeness and duplications in the electronic health record. J Am Med Inform Assoc 2020 Aug 1;27(8):1190–1197. PMID:32620950

25. Hier DB, Pearson J. Two algorithms for the reorganisation of the problem list by organ system. BMJ Health Care Inform 2019 Dec;26(1). PMID:31848142

26. Hammond KW, Helbig ST, Benson CC, Brathwaite-Sketoe BM. Are Electronic Medical

Records Trustworthy? Observations on Copying, Pasting and Duplication. AMIA Annu Symp Proc 2003;2003:269–273. PMID:14728176

27. Kreuzthaler M, Pfeifer B, Vera Ramos JA, Kramer D, Grogger V, Bredenfeldt S, Pedevilla M, Krisper P, Schulz S. EHR problem list clustering for improved topic-space navigation. BMC Med Inform Decis Mak 2019 Apr 4;19(Suppl 3):72. PMID:30943968

28. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, McLoughlin KS, Ramelson H, Schneider L, Bates DW. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. J Am Med Inform Assoc 2012 Aug;19(4):555–561. PMID:22215056

29. Bredfeldt CE, Awad EB, Joseph K, Snyder MH. Training providers: beyond the basics of electronic health records. BMC Health Serv Res 2013 Dec 2;13:503. PMID:24295150

30. Bakel LA, Wilson K, Tyler A, Tham E, Reese J, Bothner J, Kaplan DW. A quality improvement study to improve inpatient problem list use. Hosp Pediatr 2014 Jul;4(4):205–210. PMID:24986988

31. Klappe ES, de Keizer NF, Cornet R. Factors Influencing Problem List Use in Electronic Health Records-Application of the Unified Theory of Acceptance and Use of Technology. Appl Clin Inform 2020 May;11(3):415–426. PMID:32521555

32. Simons SMJ, Cillessen FHJM, Hazelzet JA. Determinants of a successful problem list to support the implementation of the problem-oriented medical record according to recent literature. BMC Med Inform Decis Mak 2016 Aug 2;16:102. PMID:27485127

33. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care 2007 Jun;13(6 Part 1):277–278. PMID:17567224

34. statistique O fédéral de la. Classification suisse des interventions chirurgicales (CHOP) - Index systématique - Version 2021 | Publication [Internet]. Office fédéral de la statistique. 2020 [cited 2021 Jan 7]. Available from: /content/bfs/fr/home/statistiken/gesundheit/nomenklaturen/medkk/instrumente-medizinische-kodierung.assetdetail.13772935.html

35. Adaptation steps [Internet]. [cited 2020 Nov 25]. Available from: https://www.dimdi.de/dynamic/en/classifications/icd/icd-10-gm/history/adaptation-steps/

36. statistique O fédéral de la. Manuel de codage médical. Le manuel officiel des règles de codage en Suisse - Version 2020 | Publication [Internet]. Office fédéral de la statistique. 2019 [cited 2020 Nov 25]. Available from: /content/bfs/fr/home/statistiken/gesundheit/nomenklaturen/medkk/instrumente-medizinische-kodierung.assetdetail.9927930.html

37. SNOMED CT Starter Guide - SNOMED CT Starter Guide [Internet]. [cited 2021 Jan 7]. Available from: https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide

38. Agrawal A, He Z, Perl Y, Wei D, Halper M, Elhanan G, Chen Y. The readiness of SNOMED

problem list concepts for meaningful use of electronic health records. Artif Intell Med 2013 Jun;58(2):73–80. PMID:23602702

39. Academy of nutrition and dietetic. Guide de poche du manuel de référence de la terminologie internationale de diététique et de nutrition (TIDN): terminologie normalisée pour le processus de soins en nutrition. Presses de l'Université Laval;

Preprint JMIR Publications

7.2.3. Article 3

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

XSL•FO
RenderX

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.

XSL•FO
RenderX

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

XSL•FO
RenderX

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

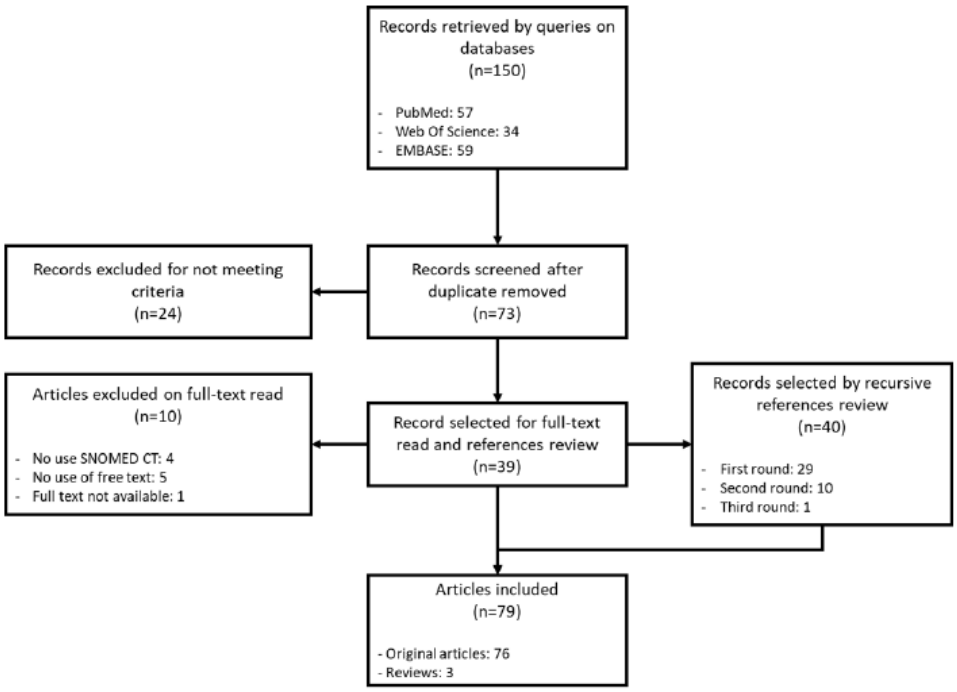
Article Selection

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.



XSL•FO
RenderX

Figure 2. Number of publications per year of publication.

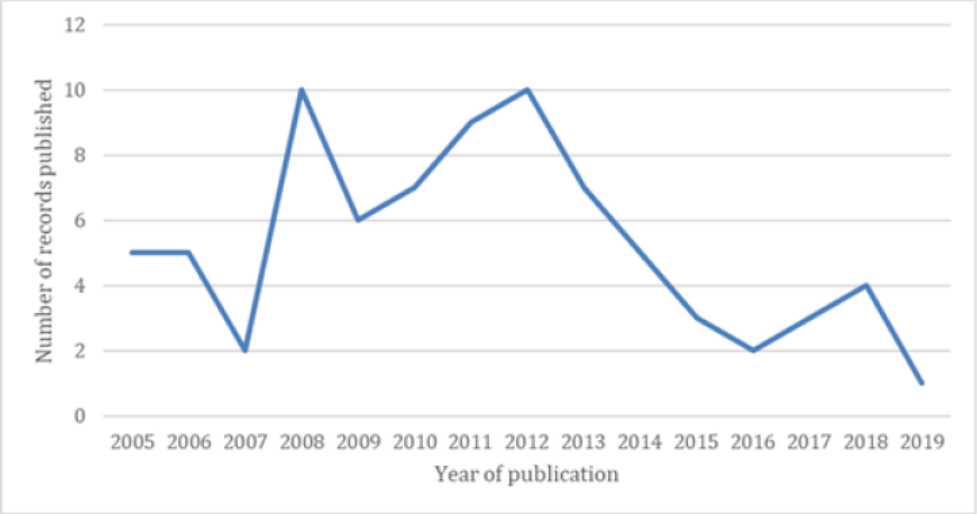


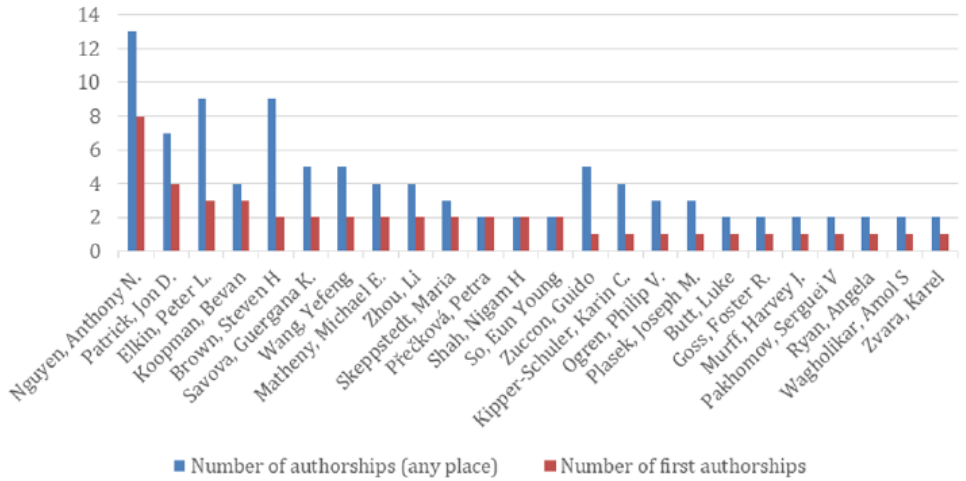
Table 1. Journals and conferences having more than 1 article in the selection.

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.

XSL•FO
RenderX

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)

XSL•FO
RenderX

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.

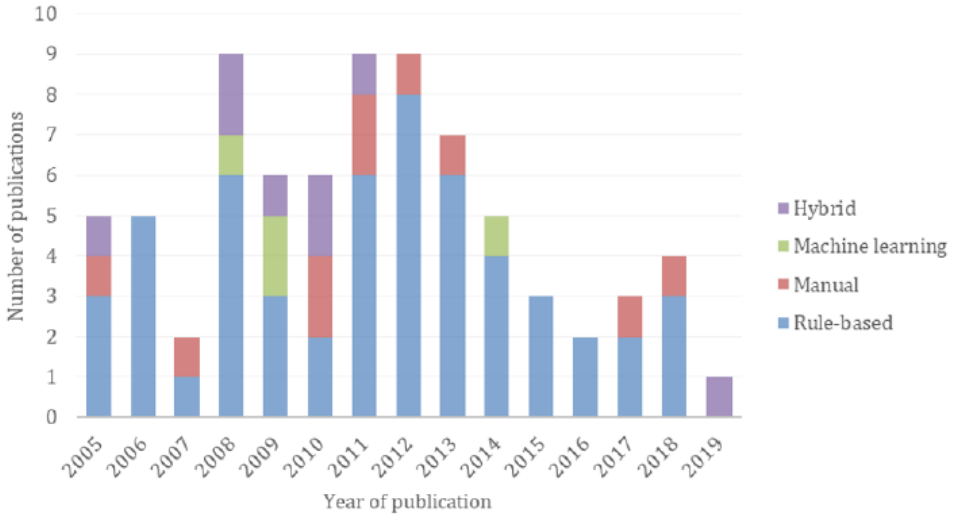
Table 6. Method used for mapping free-text data to Systematized Nomenclature of Medicine Clinical Terms.

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.

XSL•FO
RenderX

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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.

XSL•FO
RenderX

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.

XSL•FO
RenderX

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]

XSL•FO
RenderX

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

XSL•FO
RenderX

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]

XSL•FO
RenderX

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]

XSL•FO
RenderX

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]

XSL•FO
RenderX

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]

XSL•FO
RenderX

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:

©Christophe Gaudet-Blavignac, Vasiliki Foufi, Mina Bjelogrlic, Christian Lovis. Originally published in the Journal of Medical Internet Research (http://www.jmir.org), 26.01.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on http://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

7.2.4. Article 4

Original Paper

Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases With Extracted Entities and Their Relations

Vasiliki Foufi[1,2*], PhD; Tatsawan Timakum[3*], BA, MA; Christophe Gaudet-Blavignac[1,2*], BSc CS, MMed; Christian Lovis[1,2], MD, MPH, FACMI; Min Song[3], PhD

[1]Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland
[2]Faculty of Medicine, University of Geneva, Geneva, Switzerland
[3]Department of Library and Information Science, Yonsei University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Min Song, PhD
Department of Library and Information Science
Yonsei University
50 Yonsei-ro, Seodaemun-gu
Seoul, 120-749
Republic of Korea
Phone: 82 22123 2405
Fax: 82 2393 8348
Email: min.song@yonsei.ac.kr

Abstract

Background: Social media platforms constitute a rich data source for natural language processing tasks such as named entity recognition, relation extraction, and sentiment analysis. In particular, social media platforms about health provide a different insight into patient's experiences with diseases and treatment than those found in the scientific literature.

Objective: This paper aimed to report a study of entities related to chronic diseases and their relation in user-generated text posts. The major focus of our research is the study of biomedical entities found in health social media platforms and their relations and the way people suffering from chronic diseases express themselves.

Methods: We collected a corpus of 17,624 text posts from disease-specific subreddits of the social news and discussion website Reddit. For entity and relation extraction from this corpus, we employed the PKDE4J tool developed by Song et al (2015). PKDE4J is a text mining system that integrates dictionary-based entity extraction and rule-based relation extraction in a highly flexible and extensible framework.

Results: Using PKDE4J, we extracted 2 types of entities and relations: biomedical entities and relations and subject-predicate-object entity relations. In total, 82,138 entities and 30,341 relation pairs were extracted from the Reddit dataset. The most highly mentioned entities were those related to oncological disease (2884 occurrences of cancer) and asthma (2180 occurrences). The relation pair anatomy-disease was the most frequent (5550 occurrences), the highest frequent entities in this pair being cancer and lymph. The manual validation of the extracted entities showed a very good performance of the system at the entity extraction task (3682/5151, 71.48% extracted entities were correctly labeled).

Conclusions: This study showed that people are eager to share their personal experience with chronic diseases on social media platforms despite possible privacy and security issues. The results reported in this paper are promising and demonstrate the need for more in-depth studies on the way patients with chronic diseases express themselves on social media platforms.

(J Med Internet Res 2019;21(6):e12876) doi: 10.2196/12876

KEYWORDS

social media; chronic disease; data mining

XSL•FO
RenderX

Introduction

Background

People are often concerned about their health status and a range of medical issues, especially when it comes to complex or chronic diseases that can take a long time to treat or monitor. Patients often desire easy access to information about diseases and symptoms to understand their condition and to facilitate self-management of diseases without total reliance upon interaction with a physician [1]. Patients with chronic diseases in particular use social media to seek and provide social, emotional, and practical support [2]. Therefore, social media information can influence patients' decisions to manage their chronic condition [3].

Social media platforms may support patients in their search for medical products or provide suggestions to promote healthy behavior and can improve health education as they allow people to write about their experiences with diseases, drugs, symptoms, and treatments. In recent years, social media platforms have grown quickly, with the public, patients, and health professionals sharing their experiences, looking for information, and interacting with others.

Currently, more than 74% of internet users connect to social media, and 42% of the internet users take advantage of social media for health information. Moreover, 32% of social media users in the United States share about their health care experiences and family's struggle stories and 29% search for health information via social media platforms to observe other patients' experiences with their diseases [3]. Furthermore, 51% of those who live with a chronic disease have used the internet for information about health topics such as details of a specific disease, medical procedures, drugs, medical devices, or health insurances [4].

With its growing number of users, social media has become a powerful tool that can promote information sharing about health care, provide feedback from users, and foster support systems [5]. In addition, the existence of social media platforms enables researchers to learn and discover the health experiences and feeling of patients and potentially discover new knowledge in health science. For example, user conversation content from health-related online forums, such as blogs, Twitter, and Facebook, has already been analyzed to find the clusters of breast cancer symptom [6], examine smoking [7], and understand the user discourse and describe social media interactions about obesity prevention [8]. In particular, Reddit has been used as a data source for similar studies [9-12].

The interactions between individuals on social media and the information they share constitute an important new source of data that can be used, on one hand, to understand the impact of drugs, diseases, and medical treatments on patients outside controlled clinical settings and, on the other hand, to comprehend health-related behavior.

Discovering public knowledge in social media text constitutes a challenge for researchers and health care providers. To achieve this goal, various text mining approaches, such as topic modeling, information extraction, and visualization, exist.

Biomedical Entity and Relation Extraction

In the era of biomedical text mining, bioentities and their relations have arisen as a challenge to discover new knowledge. To mine the huge amounts of unstructured data, automatic information extraction tools have been conceived and developed based on several approaches. There are multiple systems developed for the identification and analysis of relations between diseases, drugs, and genes, such as Extraction of Drugs, Genes and Relations, a natural language system that extracts information about drugs and genes relevant to cancer from the biomedical literature [13]. Extraction of drug-disease treatment pairs from the published literature was also carried out [14,15]. To extract health social media information, adverse drug reactions and drug indications from a Spanish health forum were examined [16] using MeaningCloud [17], a multilingual text analysis engine based on a distant-supervision method to detect relations between drugs and side effects and used them to classify the relation instances.

PKDE4J2.0 is a system that extracts bioentities and their relations with the aim to discover biomedical scientific knowledge. It is based on a dictionary to automatically tag bioentities according to their types and a set of predefined rules used for relation extraction. PKDE4J2.0 can be applied for knowledge search, knowledge network construction, and knowledge inference [18]. PKDE4J1.1 was used to investigate drug-disease interactions in article abstracts from PubMed Central for making drug-symptom-disease triples [19]. This tool was also applied in biomedical literature to extract biomedical verbs to present a relation type between 2 entities [20] and on full-text papers to extract biological entities from diseases and genes and construct a knowledge network [21].

Health Information Extraction From Social Media Platforms

A large number of patients, caregivers, and health professionals use social media platforms to discuss mental health issues. They also constitute an important data source for researchers. Machine learning and statistical methods were used to discriminate online messages between depression and control communities using mood, psycholinguistic processes, and content topics extracted from the posts generated by members of these communities [22]. Users are interested in searching for treatment-related information, communicating with physicians to share their feelings about treatment effectiveness and side effects, discussing questions in health communities, and gaining knowledge about their conditions [23]. User-generated content from these platforms contains valuable information [24]. Their posts reflect what users think and feel about their medical experiences and often attract the attention of other patients, caregivers, and doctors.

Lu et al [25] mined data from online health communities and used text clustering integrating medical domain–specific knowledge to investigate patient needs and interests. Their results show that compared with existing methods, the addition of medical domain–specific features into their feature sets achieved significantly better clustering than was achieved without the addition of those features. Moreover, there were significant differences in hot topics on different kinds of disease

XSL•FO
RenderX

discussion platforms. Health-related posts on social media were analyzed to investigate the polarity of opinions online, performing sentiment analysis [26]. Medical terms, including those related to conditions, symptoms, treatments, effectiveness, and side effects, were extracted to generate a virtual document addressing each question raised by members of the community. Then latent Dirichlet allocation (LDA) was modified by adding a weighting scheme known as conditional LDA to cluster virtual documents with similar distributions of medical terms into a conditional topic (C-topic). Finally, the clustered C-topics were analyzed according to sentiment polarities and physiological and psychological sentiments. Identification of topics of patients' discussions on (1) Facebook about breast cancer and (2) cancerdusein.org was performed [27]. These topics were assigned to functional and symptomatic dimensions by applying LDA topic modeling and identified relations between the topics and the questionnaires.

Among others, Denecke [1] reported that "user-generated content on the web has become a new source of useful information to be added to the conventional methods of collecting clinical data."

In terms of biomedical information extraction, previous studies relied on formal research and individual case studies to identify biomedical information. These approaches include observations of changes in patients [28], meta-analysis of data from relevant databases [29], and surveys of cancer patients [30]. However, the scientific literature is generally limited to subscribers, and electronic medical records are not publicly available for reasons of patient privacy [31]. Moreover, these sources do not provide a complete understanding of how patients suffering from a chronic disease feel and how they express these feelings.

Using data from conversations between patients on social media platforms provides valuable information for researchers, physicians, and health care providers. This data source is different from, and complementary to, that obtained from conventional experimental methods.

Research Objectives

Therefore, a social media platform (Reddit) was chosen as the data source for this research that aimed to answer the following questions:

1. Which biomedical entities are prominent in the health social media platforms?
2. What types of entities are related in the corpus?
3. How do people express themselves about chronic diseases on social media platforms?

Methods

Data Collection

The data used for this research were extracted from disease-specific subreddits of the social news and discussion website Reddit [32]. Forums such as Reddit tend to have sharp contrast when compared with similar offline groups; for instance, people are likely to discuss problems that they do not feel comfortable to discuss face to face [33]. As of 2013, Reddit's official statistics included 56 billion page views, 731 million unique visitors, 40,855,032 posts, and 404,603,286 comments [34]. In particular, the subreddit about cancer numbers 22,429 subscribers and 75 posts per day [35]. These numbers demonstrate the external validity of Reddit. Another reason for having chosen Reddit as a data source is that the language of text posts is more structured than in other social media platforms such as Twitter.

Reddit's core functionality is the sharing of text-based posts with others who may or may not be members of the site. The subforum function allows the creation of designated spaces for users to congregate and interact with each other over a shared interest. Those subforums are called *subreddits*. A finite set of 19 subreddits related to chronic diseases was empirically selected for analysis. The choice of the specific subreddits was based on medical expertise and on the impact of these diseases on the quality of everyday life of patients.

As the main goal was the detection of relations between entities and of the way people suffering from chronic diseases express themselves in social media and not the study of characteristics of specific chronic diseases, the posts from the 19 subreddits were merged in a single dataset.

All of these subreddits host public content. In this research, no populational study has been performed. The study focuses on the expression of feelings and not on the identity of people sharing their experiences. From each post, only the title of the post and the body or textual content was extracted without additional information related to their authors.

The study was submitted to the Swiss Ethical Committee who concluded to a decision of nonconsideration provided that the collected data are not identifiable.

Lexicosemantic Resources

Lexicosemantic resources were constructed and incorporated into the tool. These resources included a list of stop words and biomedical dictionaries of diseases, drugs, anatomy, procedures, symptoms, side effects, and findings created from clinical health care terminologies such as the Systematized Nomenclature of Human and Veterinary Medicine - Clinical Terms [36], the National Library of Medicine's controlled vocabulary thesaurus [37], the Gene Ontology knowledgebase [38], the Kyoto Encyclopedia of Genes and Genomes database [39], and the DrugBank database [40]. Semantic relations properties were attributed to 4558 biomedical verbs extracted from the Unified Medical Language System [41].

The dictionaries were enriched with lemmas extracted from the corpus; for instance, *chemo*, *AML* (acute myeloid leukemia), *take care, support*, and *fight*.

Description of the Tool

In this research, the PKDE4J version 2.0 tool [42] was used. This text mining system consists of 2 modules: entity extraction and relation extraction.

Entity Extraction Module

This module integrates dictionary-based entity extraction and rule-based relation extraction into a highly flexible and extensible framework. The Stanford CoreNLP pipeline [43]

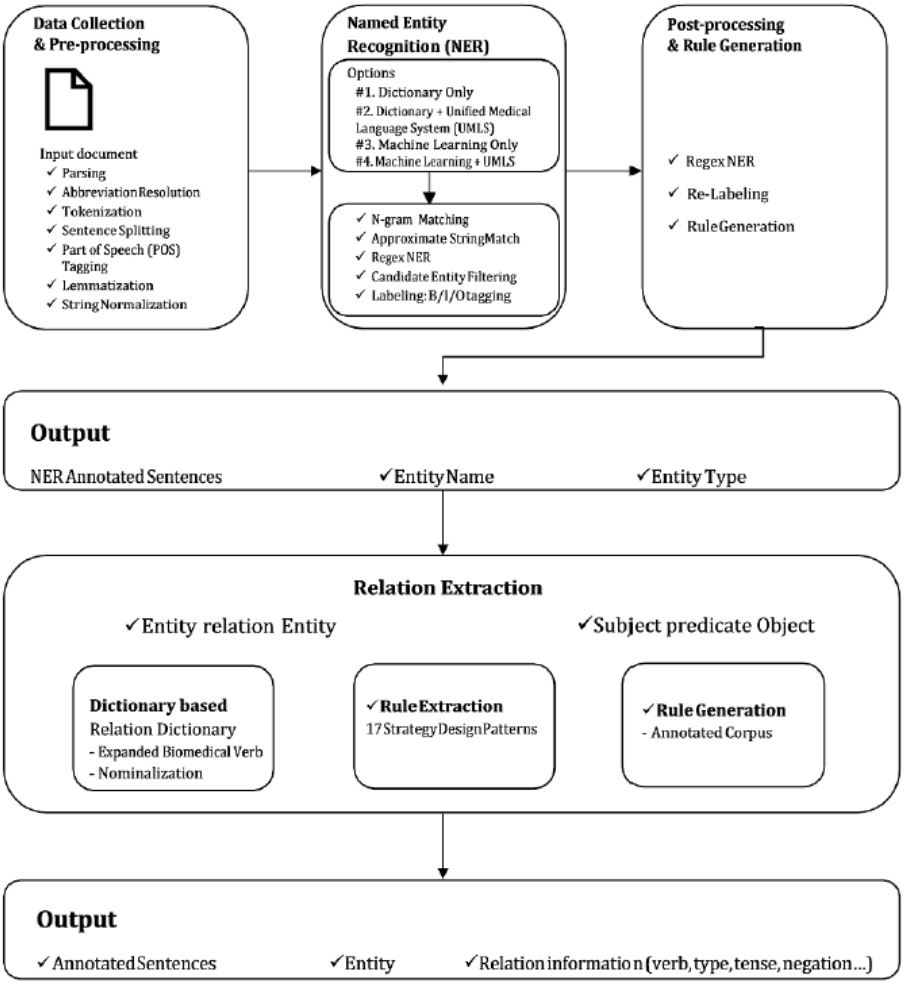
XSL•FO
RenderX

was modified to make it suitable for advanced dictionary-based entity extraction. The entity extraction module consists of 4 major submodules: preprocessing, dictionary loading, entity annotation, and postprocessing. PKDE4J can analyze entities and relations from both structured and unstructured text.

Relation Extraction Module

The relation extraction workflow identifies directed qualified relations starting from sentences from which 2 or more entities have been extracted by the entity extraction module. The relation extraction module takes a list of verbs and nominalization terms that are employed to identify relations of interest. After extracting entities from a sentence, further relation extraction algorithms are executed to construct rules for the extraction of relations of entities. A set of 20 dependency parsing–based rules is at the core of the relation extraction module and provides an ontologically enriched structure for sentences by annotating edges with dependency types. To extract relations, the system identifies a verb, which may be located between entities and contains relational characteristics, then, it checks the bioverb list to determine the relation between the entities (Figure 1) [44].

Figure 1. The workflow of the PKDE4J text mining system.



XSL•FO
RenderX

Visualization

The Gephi platform [45] was used to visualize the network of chronic diseases in the corpus. To build a graph, the k- shortest paths routing algorithm was applied. The graph visualization tool was then used to map the chronic disease entities. A PageRank of terms was computed to rank the important entities in the network; therefore, entities ranked highly by PageRank have the highest impact.

Validation of Entity Extraction

To evaluate the performance of the tool on entity extraction, 1000 posts randomly selected from the entire corpus were manually validated. The entities were evaluated as correct or incorrect based on the following specific guidelines.

Findings and Symptoms

This category refers to a phenomenon that is experienced by a person or described by a clinician and cannot be considered as a disease in the context, for example, "This news makes me feel anxiety."

Disease Names

This category refers to an abnormal condition of a human, animal, or plant that causes discomfort or dysfunction [46]. As also mentioned in the previous category, the context helps to distinguish between a disease and a symptom or finding. For example, in the sentence "After trying which dosage is good, my insomnia is thankfully gone again," *insomnia* refers to a disease, whereas in the sentence "I have had symptoms of insomnia within the last months," *insomnia* describes a symptom/finding.

Side Effects

This category includes a symptom/finding or a disease that is caused by a treatment in the context. For example:

> *Since beginning treatment have woken with bouts of nausea...*

Procedure

Procedure refers to any intervention carried on someone related to physical mental or social health. For example:

> *...treatment which would include surgery and radiation/chemotherapy according to his oncologist*

Results

Data Collection

A dataset of 17,624 text posts was semiautomatically collected using crawlers accessing public streams. Table 1 shows the subreddits used for this research, the number of posts per subreddit, and the proportion of corpus representation of each subreddit:

Table 1. Sources used for the data collection.

Subreddit name	Number of posts	Proportion of posts from each subreddit in the corpus
r/cancer	5210	26.9
r/MultipleSclerosis	1902	9.8
r/rheumatoid	1783	9.2
r/CrohnsDisease	1722	8.9
r/Asthma	1600	8.3
r/testicularcancer	1384	7.1
r/Parkinsons	1042	5.4
r/Hashimotos	1022	5.3
r/Alzheimers	927	4.8
r/breastcancer	794	4.1
r/braincancer	623	3.2
r/pancreaticcancer	397	2.1
r/lymphoma	387	2.0
r/leukemia	223	1.2
r/kidney	107	0.6
r/multiplemyeloma	104	0.5
r/thyroidcancer	63	0.3
r/lungcancer	41	0.2
r/skincancer	15	0.1
Total	19,346	100

XSL•FO
RenderX

After sorting the data corpus, duplicate posts and those with no relevant meaning, such as advertising posts and posts containing only a hyperlink, were removed. The final corpus comprises 17,580 posts (2,137,115 tokens).

Biomedical Entity and Relation Extraction

The PKDE4J system performed named entity extraction and 2 types of relation extraction: relations between biomedical entities and between subjects, predicates, and objects on the sentence level. The system's output is a corpus annotated with entities and information about their relation.

The entities are either simple terms or complex structures referring to diseases, anatomy, procedures, findings, symptoms, side effects, or drugs. In total, PKDE4J extracted 82,138 entities from the Reddit dataset, as shown in Table 2. The entity names and entity types were allocated to the 7 categories of the biomedical dictionaries. The 10 most frequent entity names followed by the number of occurrences in the corpus are displayed in Table 3. It should be noted that the terms are given in the text in the form found in the corpus. Therefore, abbreviated terms have not been expanded.

As displayed in the table, 29,669 disease entities were extracted representing 1341 unique diseases; 19,956 anatomy entities, of which 369 are distinct anatomical terms; 11,549 procedures of 296 different types; 6256 symptoms entities describing 65 symptoms; 5351 entities representing side effects of 321 different types; and 35 different drug names (616 in total). The most highly represented diseases are oncological (*cancer*, *breast cancer*, *tumor*, *leukemia*, and *lymphoma*) or relate to asthma. The anatomy category contains a range of anatomical terms. Specifically, *blood* is the most frequent term. Other widely used anatomical terms are *back*, *brain*, *hand*, *hair*, *breast*, *chest*, *heart*, and *neck*. The procedures category comprises terms referring to chemical treatments (*chemo*), surgery, laboratory test (*blood test*), social interventions (*advice* and *listening*), and others.

The most frequent symptom mentioned in the corpus is *pain* (472 occurrences). *Fatigue*, *inflammation*, *nausea*, and *cough* are some of the symptoms commonly reported by patients or relatives in the dataset. In the side-effect category, the most frequent entities are *anxiety*, *stress*, *swelling*, *crying*, and *fear* followed by *disability* and *worry*. The most commonly reported drug is *prednisone* followed by *morphine*, *salbutamol*, and *tramadol*.

Validation of Entity Extraction

Among the 5151 extracted entities, 3682 were correctly labeled by the system, whereas 1469 were attributed with incorrect labels. The performance of the system was 71.48%.

Next, an error analysis was performed on the incorrectly labeled entities. Errors were classified into 3 categories:

1. Lexical errors (488/1469, 33.21%): the term *breast* is an anatomical term, but in the post, the compound term *breast cancer* appears. However, the system failed to extract the entire entity.
2. Dictionary errors (550/1469, 37.44%), for example, *air* and *aspergillus* were falsely listed as an anatomical term and as a drug name, respectively.
3. Ambiguous concepts (431/1469, 29.33%): the term *bleeding* could be either a disease name or a symptom.

Table 2. Entity extraction results.

Entity types	Diseases	Anatomy	Procedures	Findings	Symptoms	Side effects	Drugs
Extracted entities, n	29,669	19,956	11,549	8741	6256	5351	616
Entity names, n	1341	369	296	483	65	321	35

Table 3. Ten most frequent entities by type.

Diseases	Anatomy	Procedures	Findings	Symptoms	Side effects	Drugs
Cancer (2884)	Blood (1542)	Chemo (1914)	Related (521)	Pain (2648)	Anxiety (683)	Prednisone (417)
Asthma (2180)	Back (1034)	Treatment (1909)	Lump (359)	Fatigue (639)	Stress (373)	Morphine (33)
All (2163)	Brain (962)	Surgery (1909)	Suffering (333)	Inflammation (472)	Swelling (348)	Salbutamol (33)
Breast cancer (804)	Hand (656)	Advice (774)	Confused (305)	Scared (273)	Crying (245)	Tramadol (26)
Can (745)	Head (627)	Radiation (627)	Problem (304)	Nausea (244)	Mass (220)	MRSA[a] (14)
Tumor (631)	Hair (549)	Biopsy (366)	Attack (277)	Hurt (205)	Fear (215)	Aspirin (11)
Disease (563)	Breast (535)	Chemotherapy (338)	Energy (270)	Sore (202)	Disability (164)	Omeprazole (9)
TSH[b] (506)	Chest (511)	Blood test (268)	Terrified (266)	Numb (195)	Worry (142)	Seretide (6)
Depression (414)	Heart (503)	Infusion (199)	Tired (251)	Cutting (104)	Fall (129)	Citrus (5)
Lymphoma (348)	Neck (459)	Listening (158)	Follow up (249)	Tingling (101)	Discomfort (120)	Echinacea (5)

[a]MRSA: methicillin-resistant *Staphylococcus aureus*.

[b]TSH: thyroid stimulating hormone.

XSL•FO
RenderX

Relation Extraction

The system extracted 2 entities (entity 1 and entity 2) found in the same sentence and linked with a relation and then it attributed the type of entities. For instance, entity 1, *Borderline* (disease) co-occurs with *High blood pressure* (symptom) in the sentence. In total, 30,341 relation pairs were extracted, as shown in Table 4.

Of the 30,341 relation pairs, the most frequent entity relation pairs and their number of co-occurrences are shown in Table 5.

The relations between anatomy and disease entity types are the most frequent (5550 pairs). The pair disease-disease co-occurs 4668 times, and the pair anatomy-anatomy appears 3595 times.

Table 6 contains the 5 most frequent entities per relation pair.

Table 4. Example of entity relation extraction.

Analysis result	Entity 1	Entity 1 type	Entity 2	Entity 2 type	Sentence from post
Output 1	Borderline	disease	High blood pressure	symptom	Prior to that, I was fat mid forties male borderline high HDL, high blood pressure but ZERO issues with thyroid or immune issues.
Output 2	Optic neuritis	disease	Multiple sclerosis	symptom	She said that I have something called optic neuritis and that about half the time people get it and they don't know why but the other half its because someone has multiple sclerosis.
Output 3	Syndrome	disease	Nerve	anatomy	I had bilateral optic neuritis significantly worse in my left eye in Late August September and I was also simultaneously diagnosed with Browns Syndrome which they're not 100% convinced on as it may have been misdiagnosed 6th nerve palsy.

Table 5. Most frequent entities per relation pair.

Relation pair	Co-occurrences, n
anatomy-anatomy	3595
anatomy-disease, disease-anatomy	5550
anatomy-procedure, procedure-anatomy	1730
anatomy-symptom, symptom-anatomy	1227
anatomy-side effect, side effect-anatomy	1081
disease-disease	4668
disease-procedure, procedure-disease	2540
disease-finding, finding-disease	2128
disease-side effect, side effect-disease	1502
disease-symptom, symptom-disease	1080
finding-finding	303
finding-anatomy, anatomy-finding	1362
procedure-procedure	1023
procedure-finding, finding-procedure	430
side effect-side effect	256

XSL•FO
RenderX

Table 6. The 5 most frequent entities per relation pair.

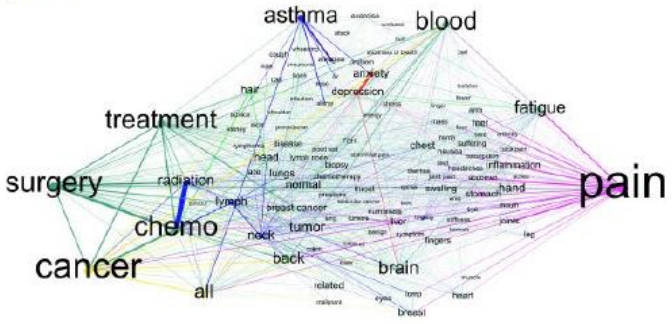
Rank	Pair of entity 1 and entity 2					
	Anatomy|Anatomy	Disease|Disease	Disease|Side effect	Disease|Anatomy	Disease|Procedure	Disease|Finding
1	Back|Hair	Cancer|ALL	Depression|Anxiety	Cancer|Lymph	Cancer|Surgery	Asthma|Attack
2	Neck|Lymph	Asthma|Allergy	Asthma|Anxiety	Tumor|Blood	Tumor|Surgery	Cancer|Suffering
3	Brain|Lungs	Tumor|Seminoma	Cancer|Fear	Asthma|Lungs	Breast cancer|Treatment	ALL|Follow up
4	Lungs|Lymph	ALL|Asthma	ALL|Swelling	ALL|Blood	Asthma|Advice	Depression|Suffering
5	Head|Hair	Depression|Fatigue	Aches|Anxiety	TSH|Blood	ALL|Treatment	Exercise|Muscle tension

The most frequent entities in the pair disease-anatomy is *Cancer|Lymph*, in the pair disease-disease is *Cancer|ALL*, and in the pair anatomy-anatomy is *Back|Hair*.

To summarize, once the most frequent entities were extracted, the results were processed according to the shortest path between each entity pair to produce the graph shown in Figure 2. Among 2561 nodes and 13,405 edges, this entity network shows that *pain* highly co-occurs with other entities in the network (biggest node, weighted at 0.022461), followed by *cancer* (PageRank score at 0.018057) and *surgery* (PageRank score at 0.015443).

The node *pain* has connections with other nodes, including *fatigue*, *inflammation*, *stomach*, *joints*, *cancer*, and *chemo*. The node *cancer* is strongly linked to *chemo*, *surgery*, *treatment*, *ALL (acute lymphoblastic leukemia)*, *anxiety*, and *blood*. The nodes of *surgery*, *chemo*, and *treatment* are linked to diseases and body parts. Finally, the entity nodes relating to mental health, such as *anxiety* and *depression*, also appear in the network and associate with other bioentity types. Table 7 shows the most frequent entities and the corresponding PageRank scores:

Figure 2. Biomedical entity network.



XSL•FO
RenderX

Table 7. The most frequent entities and the corresponding PageRank scores.

Label	PageRank score
Pain	0.022461
Cancer	0.018057
Surgery	0.015443
Chemo	0.014954
Treatment	0.014275
Blood	0.013841
Asthma	0.012554
All	0.010931
Brain	0.010311
Fatigue	0.009626
Back	0.008574
Radiation	0.008317
Tumor	0.007814
Neck	0.006968
Hand	0.006618
Lymph	0.006442
Normal	0.006176
Anxiety	0.006108
Head	0.005933
Hair	0.005762

Subject-Predicate-Object Entity Relation Extraction

The system extracted 69,263 subject or object entities. The top 10 entities are shown in Table 8. In total, 41,068 relations were extracted and the results were classified into 2 types of subjects: subject pronoun (*I, you, he, she, it, we*, and *they*) and subject noun (*treatment*). The relation pairs are divided into 19,645 pairs of subject pronoun-object entities and 21,423 pairs of subject noun-object entities.

Table 9 shows 2 examples of the subject-predicate-object relation extraction: the subject (for example *I*, *he*, *anyone*, *it*, and *asthma*), the predicate (verbs such as *have*, *get*, and *increase*), the object (terms such as *eczema*, *allergies*, *childhood asthma*, *my cough*, and *allergy shots*), and the sentence of the corresponding post.

Subject Pronoun-Predicate-Object

The subject pronoun-predicate-object relation extraction demonstrates that the most frequent subject pronoun is *I* (11,691 times, including *I've*, *I'm*, and *I'd*). Some examples are shown in Table 10.

Subject Noun-Predicate-Object

Table 11 shows some examples of subject noun-predicate-object relation extraction. Among the 21,423 relation pairs, the most frequent subject nouns are diseases such as *asthma* (272 occurrences), including phrases such as *asthma anxiety*, *asthma attacks*, *my asthma*, *my asthma and allergies*, *my asthma flare*, and *cancer* (226 occurrences).

XSL•FO
RenderX

Table 8. The top 10 occurrences of subjects and objects.

Entity	Count, n
I	10,314
It	2832
Pain	2341
She	1501
He	1323
Cancer	1060
Asthma	1015
They	900
Me	719
You	597

Table 9. Examples of subject-predicate-object relation extraction results.

Analysis result	Subject entity	Predicate	Object entity	Sentence
Output 1	I	Had	the skin allergy test	I had the skin allergy test done and it came back positive for almost every kind of pollen and mold, etc.
Output 2	my blue inhaler	Increases	my asthma	I noticed consistently my blue inhaler increases my asthma about 30% after using it and I believe was the cause of a recent very bad asthma attack.

Table 10. Example of subject pronoun-predicate-object relation extraction.

Subject	Predicate	Object	Sentence
I	take	the typical seretide	I take the typical seretide morning and night and ventolin when I need it.
I	have	a deep and painful cough	I have a deep and painful cough that's been leaving me with back, chest, and side pains.
You	ever take	allergy shots	Hey, to you asthmatics who have allergy induced asthma, did you ever take allergy shots.
He	was given	Prednisone	He was given prednisone for that as well.
She	has	Asthma	My sister, who lives with me, started complaining to me about it, saying that she doesn't want me doing that when her daughter is home because she has asthma and it smokes up the house.

XSL•FO
RenderX

Table 11. Example of subject noun-predicate-object relation extraction.

Subject	Predicate	Object	Sentence
Asthma	is becoming way more than just	a physical issue	Asthma is becoming way more than just a physical issue, it's taking a toll on my mental health.
Cancer	had spread to	her bones	The doctor told her that cancer had spread to her bones and that she'll have to have injections for it?
Fever	is indeed mentioned as	a side effect	Ive also been using Modulair Montelukast Sodium, and fever is indeed mentioned as a side effect on my leaflet.
Hives	are from	allergies	They're trying to tell me they are panic attacks but as far as I know hives are from allergies and they sometimes happen during my asthma attacks.
The depression	is occurring simultaneously with	the increased asthma symptoms	I noticed the depression is occurring simultaneously with the increased asthma symptoms and was wondering if there is a correlation and if anyone else has experienced this.
My milk allergy	was causing	my asthma	My milk allergy was causing my asthma.
My second course of prednisone	has been great for stopping	the wheezing	And I'm on my second course of prednisone which has been great for stopping the wheezing - even the rescue inhaler didn't help before.
The doctor	ruled out	pneumonia	Anyway, the doctor ruled out pneumonia and said I had caught a cold on the plane and it had triggered an asthma exacerbation.

Figure 3 demonstrates the most used subject and object entities. The results show that the most frequent subject is the pronoun *I* (PageRank score: 0.100619). The pronoun *It*, *She*, *He*, *They*, and *You* are also frequently used. Diseases, body parts, treatments, and symptoms are widely used as the subjects and/or objects as well as the possessive pronoun *my* (*my mother*, *my eyes*, and *my dad*). Table 12 presents the most frequent subject and object entities and the corresponding PageRank scores.

Social Media Language

Expressions that constitute specific terms developed on social media, such as *pm* (private message), *FWIW* (For What it's Worth) were identified in the corpus. "A common feature of microblog texts is the use of symbols in posts, such as the love-heart dingbat symbol" [47]. Emoticons such as 😁, and text-based emoticons such as *LOL* (laughing out loud), :), :-(, =), and :(are also frequent.

In addition, the corpus contains informal phrases such as "Rooting for you!" and "I'm still chugging along"; adjectives such as *loopy*, *drippy*, *dicey*, and *zonked*; and verbs such as *puke* that substitute for their equivalents in standard language.

Entities found in the Reddit corpus present numerous morphosyntactic variants. For example, the term *chemotherapy* was rarely found, but the short form *chemo* was frequently used. The disease name *Hodgkin's Lymphoma* is as *Hodgkin Lymphoma*, *Hodgkins Lymphoma*, *Hodgkin disease*, and *HL*. Similarly, the entity name *Mixed Cellularity Classical Hodgkin Lymphoma* is found as *Mixed Cellularity Hodgkin Lymphoma*, *Mixed Cellularity Hodgkins Lymphoma*, and *MCCHL*. Moreover, there are many abbreviated forms of entity names, such as *ALL*, *AML*, *BRCA2* (breast cancer type 2), *CLL* (chronic lymphocytic leukemia), *CML* (chronic myeloid leukemia), *COPD* (chronic obstructive pulmonary disease), *DCIS* (ductal carcinoma in situ), and *GERD* (gastroesophageal reflux disease). When these forms were included in the disease dictionary, the system managed to detect them. Some examples are presented in Table 13.

XSL•FO
RenderX

Figure 3. Subject and object entity network.

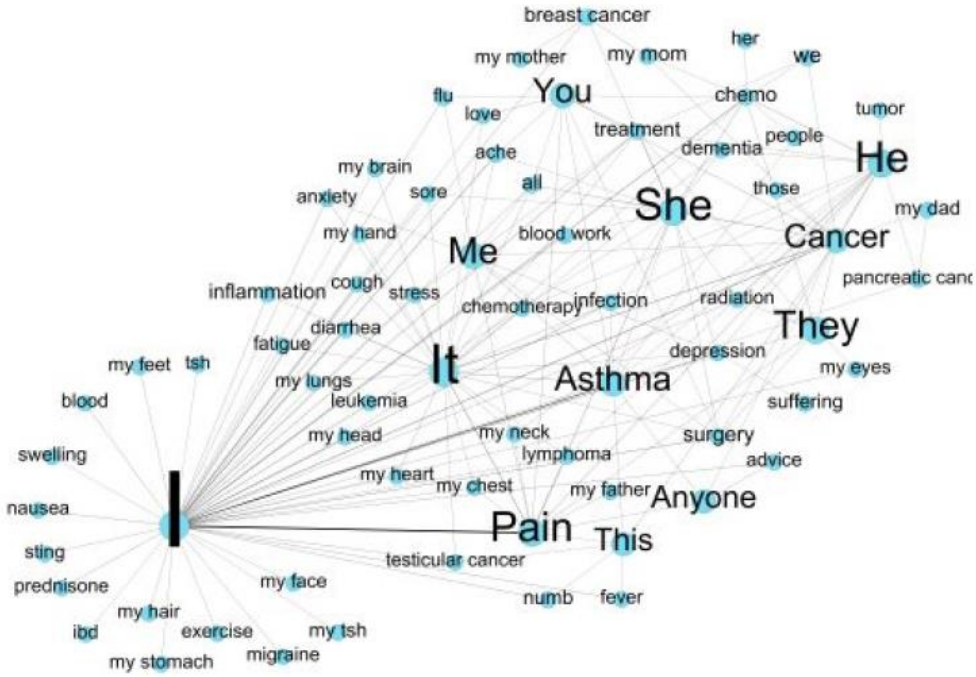


Table 12. The most frequent entities and the corresponding PageRank scores.

Label	PageRank score
I	0.100619
It	0.027234
Pain	0.020651
She	0.014206
He	0.012875
Cancer	0.00939
Asthma	0.009324
They	0.008927
Me	0.006934
You	0.005852

XSL•FO
RenderX

Table 13. Examples of disease entities in their abbreviated form.

Abbreviated form of entity name	Full entity name	Example from the corpus
ALL	Acute lymphoblastic leukemia	My boyfriend was diagnosed with ALL 2 years ago and stayed in remission after a few rounds of chemo.
AML	Acute myeloid leukemia	Diagnosed with AML this past Sept.
BRCA2	Breast cancer type 2	Her sister, my aunt, was diagnosed with breast cancer at 27 and was dead by 33 she tested positive for BRCA2 as well.
CLL	Chronic lymphocytic leukemia	My CLL is more of SLL, which is the same thing but presented in my lymph nodes.
CML	Chronic myeloid leukemia	25 years old, diagnosed with CML when I was 15.
COPD	Chronic obstructive pulmonary disease	Hes been smoking for over 40 years, has COPD and isnt in the greatest health generally overweight, inactive, etc.
DCIS	Ductal carcinoma in situ	We found out last week she has both DCIS and Invasive DCIS.
GERD	Gastroesophageal reflux disease	Sleep apnea can also worsen GERD, and GERD is known to worsen asthma.

Discussion

Principal Findings

In this paper, we collected user-generated chronic disease–related data from Reddit and extracted information pertinent to biomedical entities and their relations to examine the characteristics of the language used by users in this social media platform. Initially, the corpus was created by semiautomatically extracting posts from specific subforums of Reddit. Next, lexicosemantic resources from various sources were created. To perform the information extraction tasks—entity extraction and relation extraction—the PKDE4J text mining system was used. The system extracted 82,138 biomedical entities and 30,341 relations. These results indicate that the corpus contains a large amount of information.

Performance of the Tool

As described in the Results section, the system achieved a high performance in the named entity extraction task and the attribution of entity types (3682/5151, 71.48% extracted entities were correctly labeled). As already mentioned, the language used in Reddit is structured enough, with a satisfactory number of full sentences so the system managed to extract entities and their relations. The error analysis showed that the system failed to detect a number of entities or falsely attributed the entity type, because of lexical errors, to dictionaries' errors or to ambiguous concepts.

Entity Extraction

Entities prominent in the corpus refer to diseases, anatomical terms, procedures, findings, and symptoms. While interpreting the entities extracted from the corpus, it must be taken into account that the corpus was constructed by selecting subreddits created to share information about specific diseases. Therefore, it is expected that entities related to these diseases are the most likely to be represented. For instance, parts of the body affected by specific cancers, such as *breast* or *blood*, occur very frequently.

The most frequent disease entities in this corpus are oncologic diseases such as *cancer*, *ALL*, and *breast cancer*. Frequently mentioned nononcologic diseases are *asthma*, *depression*, and the generic entity *disease*. The entity *thyroid stimulating hormone* (TSH) is frequently mentioned, but it should be further classified in findings.

The most frequent anatomy entity is *blood*. This is explained primarily because of the numerous posts speaking about *leukemia* and *lymphoma*. Moreover, people often report the results of blood tests, a situation that increases the number of entities identified.

Terms tagged as *procedures* extracted from the corpus are mainly linked to oncologic diseases. About 2000 occurrences of *chemo* and *chemotherapy* were extracted. *Chemotherapy* is a significant procedure with numerous side effects. The fact that patients mention it at a high frequency shows that it is a treatment with a strong impact on quality of life and raises a lot of questions and worries for the patients involved. Social intervention procedures such as *listening* and *advice* are also frequent (see Table 3). This observation indicates that apart from technical information about treatment and surgeries, people also speak about the support they got during their disease or search for it in the community.

In medicine, symptoms can be difficult to differentiate from findings. This difference often resides in the context of the phenomenon. In the corpus, entities belonging to those categories as well as the side-effects category can be analyzed together to gain a better understanding of the results. Most frequent entities from these categories are closely related to the patient experiences and feelings. Concepts related to the feeling of fear are the most frequently present in this merged category: 7 out of 30 entities express feelings of fear or related with fear using the words *anxiety*, *stress*, *confused*, *scared*, *terrified*, *fear*, and *worry*. This is coherent with studies on cancer survivors that state the fear of cancer recurrence as almost universal among this population [48]. It appears that people with chronic conditions use social media to share feelings they have experienced. The chronic diseases selected in this corpus frequently imply severe impact on lifestyle and decrease life expectancy. Therefore, it is logical that *fear* and *anxiety* are prominent entities in the corpus.

XSL•FO
RenderX

Health-related quality of life in chronically ill patients is a known field in medical research since numerous years. Questionnaires such as the European Organisation for Research and Treatment of Cancer Quality of Life-C15-Palliative [49] or, more recently, the Functional Assessment of Cancer Therapy-General 7 [50] and Patient-Reported Outcomes Measurement Information System [51] are used routinely to assess it in those populations. When looking at the top concerns raised by patients suffering from cancer [50], it is interesting to note that they are in line with the top entities extracted from the corpus. More specifically, the most frequent nondisease entity extracted, *pain*, is a key item in multiple quality-of-life assessment questionnaires. This shows that the experiences that the patients share on social media platforms are coherent with what has been proven to have an impact on their life.

Overall, entities extracted from the corpus are coherent with similar studies conducted on health-related social media [27] and with validated evaluation of the quality of life of patients suffering from chronic diseases.

Relation Extraction

The relation extraction performed on this corpus shows that the most highly represented relation type identified is the *disease-anatomy* relation (5550 occurrences). The pairs most frequently representing a disease and its localization are *cancer-lymph*, *asthma-lungs*, and *tumor-blood*. This suggests that people using social media platforms to speak about their chronic diseases are willing to explain which disease they suffer from as well as the location of the disease. This propensity is probably linked to the fact that such subreddits are used to share life experiences and to find people with similar backgrounds. This commonality can be reassuring and informative for a person suffering from the same disease. To find these people and knowledge, it is valuable to share the nature of the disease and its anatomical location.

The second most frequent entity pair is *disease-disease* (4668 occurrences). Pairs of entities such as *cancer-ALL* (acute lymphoblastic leukemia) and *asthma-allergy* are frequent. This co-occurrence of diseases might be related to the fact that chronic diseases often lead to complications and to other diseases. For example, *asthma-allergy* was perceived from the sentence "have allergy induced asthma."

The third most frequent entity pair is *anatomy-anatomy* (3595 occurrences). When looking at specific occurrences of this pair, the pairs *neck-lymph*, *brain-lungs*, and *lungs-lymph* are frequent. Another entity pair, *head-hair* is related to people speaking about the side effects of chemotherapy.

Relations linking *diseases* to *procedures* are present at a high frequency in the corpus (2540 occurrences). When looking specifically in this category, it is clear that the most frequently identified disease in the corpus, *cancer*, is also most highly represented in those relations.

The relations extracted from the corpus demonstrate that patients with chronic diseases are willing to share detailed information about their health condition in a structured manner, describing thoroughly the disease, its location, the symptoms it caused, and the effect of treatment.

Subject-Predicate-Object Entity Relation Extraction

The language patterns of subject-predicate-object relations demonstrate important characteristics of health social media language. As is apparent in the outputs, subject pronouns and object pronouns were frequently mentioned and were used mostly in the singular first-person pronoun, such as *I*, *me*, and *my*. These patterns are related to the way individuals share personal or family experiences ("I-had-a bad cold or sinus infection," "Allergens-explains-my severe asthma," "It-is making-my heartburn," and "Anyone-develop-eczema") and feelings ("I have a history of testicular cancer in family so Im pretty scared bht im hoping its nothing"). Also, patients or relatives, after having described their problem, treatment, and possible effects, often ask for advice, as shown in the following sentence:

> *I noticed the depression is occurring simultaneously with the increased asthma symptoms and was wondering if there is a correlation and if anyone else has experienced this?*

Social Media Language

Data derived from clinical narratives and research papers differ significantly from social media content. The language and style used by the authors as well as the content are different. From a linguistic point of view, medical blogs usually consist of syntactically correct sentences but can contain verbless clauses or sentences without subjects [52]. Abbreviations, enumerations, and citations of conversations, medical terms, and opinion-related words are used frequently in medical blog posts and websites. As stated in the study by Korkontzelos et al [53], "in social media, users rarely use technical terms." Moreover, emoticons are very often used to convey emotion or to give contextual information to correctly understand a message (such as irony or sarcasm). The corpus processed in this research confirmed these observations.

Limitations

There are 2 major limitations of the PKDE4J tool with regard to the objectives of this paper. First, PKDE4J was initially developed for the processing of well-structured biomedical texts and not for social media text. This issue has a relatively less impact on this paper given that the entity extraction task is based on dictionaries. However, for tasks such as part-of-speech and sentence parsing needed for the extraction of relations, the informality of social media text poses a challenge. Second, the lack of terms from the dictionaries as well as lexical and semantic ambiguities lowered the performance of the system. For instance, abbreviations and acronyms can have multiple interpretations, and this can lead to ambiguities. In the current version of the system, these types of ambiguities are not handled. Consequently, all occurrences of *ALL* found in the corpus were extracted, even those that do not refer to the disease *Acute Lymphocytic Leukemia*. Also, the lexical unit *back* has sometimes been falsely recognized as a body part.

Conclusions

Data from social media platforms devoted to health can provide valuable information about the experiences of the patients involved. In this paper, we reported the application of an

XSL•FO
RenderX

information extraction approach using the PKDE4J tool to detect, extract, and visualize chronic disease entities and relations and to identify characteristics of the social media language in a corpus collected from Reddit.

In the Results section, we showed which disease entities are frequently mentioned and which are the most frequent relation pairs. Relation extraction demonstrated that the most frequent relation pair is the *disease-anatomy* pair and the subject-object relation pattern in the social media language is the use of the first-person pronoun provided that people share personal experiences.

Although data privacy and information sharing is becoming a major concern in research and legal frameworks, such as the General Data Protection Regulation law, have begun to set boundaries for the storage and sharing of information generated by users, it is interesting that despite those concerns, users are willing to share private health information in open social networks.

Further research should focus on the enrichment of dictionaries and adaptation of rules to common usages of social media language and the processing of emoticons for the sentiment analysis task. Finally, the identification of the type of semantic relations and the evaluation on the relation extraction results should be performed to assess the performance of the system in this task.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A3A2075114) and the University of Geneva, Switzerland.

Authors' Contributions

VF, TT, and CGB contributed equally to the manuscript.

Conflicts of Interest

CL is editor-in-chief for JMIR Medical Informatics.

References

1. Denecke K. Health Web Science: Social Media Data for Healthcare. New York: Springer International Publishing; 2015.
2. Patel R, Chang T, Greysen SR, Chopra V. Social media use in chronic disease: a systematic review and novel taxonomy. Am J Med 2015 Dec;128(12):1335-1350. [doi: 10.1016/j.amjmed.2015.06.015] [Medline: 26159633]
3. ReferralMD. 2017. 30 Facts & Stats on Social Media and Healthcare URL: https://getreferralmd.com/2017/01/30-facts-statistics-on-social-media-and-healthcare/ [accessed 2019-06-03] [WebCite Cache ID 78qwn2Dif]
4. Pew Research Center. Chronic Disease and the Internet URL: https://www.pewinternet.org/2010/03/24/chronic-disease-and-the-internet/ [accessed 2019-06-03] [WebCite Cache ID 78qx81xnX]
5. Moorhead S, Hazlett D, Harrison L, Carroll J, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23:e85. [doi: 10.2196/jmir.1933]
6. Marshall SA, Yang CC, Ping Q, Zhao M, Avis NE, Ip EH. Symptom clusters in women with breast cancer: an analysis of data from social media and a research study. Qual Life Res 2016 Mar;25(3):547-557 [FREE Full text] [doi: 10.1007/s11136-015-1156-7] [Medline: 26476836]
7. Myslín M, Zhu S, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013;15(8):e174 [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]
8. Chou WS, Prestin A, Kunath S. Obesity in social media: a mixed methods analysis. Transl Behav Med 2014 Sep;4(3):314-323 [FREE Full text] [doi: 10.1007/s13142-014-0256-1] [Medline: 25264470]
9. Sharma R, Wigginton B, Meurk C, Ford P, Gartner CE. Motivations and limitations associated with vaping among people with mental illness: a qualitative analysis of Reddit discussions. Int J Environ Res Public Health 2016 Dec 22;14(1):7 [FREE Full text] [doi: 10.3390/ijerph14010007] [Medline: 28025516]
10. Park A, Conway M. Tracking health related discussions on reddit for public health applications. AMIA Annu Symp Proc 2017;2017:1362-1371. [Medline: 29854205]
11. Pandrekar S, Chen X, Gopalkrishna G, Srivastava A, Saltz M, Saltz J, et al. Social media based analysis of opioid epidemic using Reddit. AMIA Annu Symp Proc 2018;2018:867-876 [FREE Full text] [Medline: 30815129]
12. Sumner SA, Galik S, Mathieu J, Ward M, Kiley T, Bartholow B, et al. Temporal and geographic patterns of social media posts about an emerging suicide game. J Adolesc Health 2019 Feb 25:1-7 (forthcoming). [doi: 10.1016/j.jadohealth.2018.12.025] [Medline: 30819581]
13. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000:517-528 [FREE Full text] [Medline: 10902199]
14. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics 2013 Jun 6;14(1):181. [doi: 10.1186/1471-2105-14-181]

XSL•FO
RenderX

15. Zhu Y, Song M, Yan E. Identifying Liver Cancer and Its Relations with Diseases, Drugs, and Genes: A Literature-Based Approach. PLoS One 2016 May 19;11(5):e0156091 [FREE Full text] [doi: 10.1371/journal.pone.0156091] [Medline: 27195695]
16. Segura-Bedmar I, Martínez P, Revert R, Moreno-Schneider J. Exploring Spanish health social media for detecting drug effects. BMC Med Inform Decis Mak 2015 Jun 15;15(2):S6 [FREE Full text] [doi: 10.1186/1472-6947-15-S2-S6] [Medline: 26100267]
17. MeaningCloud. Text Analytics – MeaningCloud text mining solutions URL: https://www.meaningcloud.com/ [accessed 2019-06-03] [WebCite Cache ID 78r622IFx]
18. Song M, Kim M, Kang K, Kim YH, Jeon S. Application of public knowledge discovery tool (PKDE4J) to represent biomedical scientific knowledge. Front Res Metr Anal 2018:1-16 [FREE Full text] [doi: 10.3389/frma.2018.00007]
19. Song M, Kang K, An JY. Investigating drug–disease interactions in drug–symptom–disease triples via citation relations. J Assoc Inf Sci Technol 2018 Jul 30:1355-1368 [FREE Full text] [doi: 10.1002/asi.24060]
20. Kim Y, Beak S, Song M. Constructing Linguistic Verb Source for Relation Extraction. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016 Presented at: CIKM'16; October 24-28, 2016; Indianapolis, Indiana, USA p. 2511-2512 URL: http://dtmbio.net/dtmbio2016/pdf/8.pdf
21. Amplayo R, Song M. ACL Member Portal. 2016. Building Content-driven Entity Networks for Scarce Scientific Literature using Content Information URL: https://aclweb.org/anthology/W16-5103 [accessed 2019-06-03] [WebCite Cache ID 78r7Ga6MV]
22. Nguyen T, Phung D, Dao B, Venkatesh S, Berk M. Affective and content analysis of online depression communities. IEEE Trans Affective Comput 2014 Jul 1;5(3):217-226. [doi: 10.1109/Taffc.2014.2315623]
23. Monnier J, Laken M, Carter CL. Patient and caregiver interest in internet-based cancer services. Cancer Pract 2002 Nov;10(6):305-310. [doi: 10.1046/j.1523-5394.2002.106005.x]
24. O'Neill B, Ziebland S, Valderas J, Lupiáñez-Villanueva F. User-generated online health content: a survey of internet users in the United Kingdom. J Med Internet Res 2014 Apr 30;16(4):e118 [FREE Full text] [doi: 10.2196/jmir.3187] [Medline: 24784798]
25. Lu Y, Zhang P, Liu J, Li J, Deng S. Health-related hot topic detection in online communities using text clustering. PLoS One 2013 Feb;8(2):e56221 [FREE Full text] [doi: 10.1371/journal.pone.0056221] [Medline: 23457530]
26. Yang F, Lee AJ, Kuo S. Mining health social media with sentiment analysis. J Med Syst 2016 Nov;40(11):236. [doi: 10.1007/s10916-016-0604-4] [Medline: 27663246]
27. Tapi NMD, Bringay S, Lavergne C, Mollevi C, Opitz T. What patients can tell us: topic analysis for social media on breast cancer. JMIR Med Inform 2017 Jul 31;5(3):e23 [FREE Full text] [doi: 10.2196/medinform.7779] [Medline: 28760725]
28. Carelle N, Piotto E, Bellanger A, Germanaud J, Thuillier A, Khayat D. Changing patient perceptions of the side effects of cancer chemotherapy. Cancer 2002 Jul 1;95(1):155-163 [FREE Full text] [doi: 10.1002/cncr.10630] [Medline: 12115329]
29. Sohl SJ, Schnur JB, Montgomery GH. A meta-analysis of the relationship between response expectancies and cancer treatment-related side effects. J Pain Symptom Manage 2009 Nov;38(5):775-784 [FREE Full text] [doi: 10.1016/j.jpainsymman.2009.01.008] [Medline: 19775863]
30. Aslam MS, Naveed S, Ahmed A, Abbas Z, Gull I, Athar MA. Side effects of chemotherapy in cancer patients and evaluation of patients opinion about starvation based differential chemotherapy. J Cancer Ther 2014;05(08):817-822. [doi: 10.4236/jct.2014.58089]
31. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014 Sep;83(9):605-623 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.06.009] [Medline: 25008281]
32. Reddit. URL: https://www.reddit.com/ [accessed 2019-06-03] [WebCite Cache ID 78r7x3juM]
33. Johnson GJ, Ambrose PJ. Neo-tribes: the power and potential of online communities in health care. Commun ACM 2006 Jan 1;49(1):107-113. [doi: 10.1145/1107458.1107463]
34. RedditBlog. Top Posts of 2013, Stats, and Snoo Year's Resolutions URL: https://redditblog.com/2013/12/31/top-posts-of-2013-stats-and-snoo-years-resolutions/ [accessed 2019-06-03] [WebCite Cache ID 78r86wPbY]
35. Subreddit Stats. /r/cancer stats URL: https://subredditstats.com/r/cancer [accessed 2019-06-03] [WebCite Cache ID 78r8J8oUb]
36. SNOMED. URL: http://www.snomed.org/ [accessed 2019-06-03] [WebCite Cache ID 78rD4S5yd]
37. National Center for Biotechnology Information. Medical Subject Headings URL: https://www.ncbi.nlm.nih.gov/mesh [accessed 2019-06-03] [WebCite Cache ID 78rDzIavh]
38. Gene Ontology (GO) Knowledge Base. URL: http://geneontology.org/ [accessed 2019-06-03] [WebCite Cache ID 78rEGtUWi]
39. GenomeNet. Kyoto Encyclopedia of Genes and Genomes (KEGG) disease database URL: https://www.genome.jp/kegg/disease/ [accessed 2019-06-05] [WebCite Cache ID 78tdhhRi4]
40. DrugBank. URL: https://www.drugbank.ca/ [accessed 2019-06-03] [WebCite Cache ID 78rFjahY6]
41. National Library of Medicine. Unified Medical Language System URL: https://www.nlm.nih.gov/research/umls/ [accessed 2019-06-03] [WebCite Cache ID 78rG0jTjh]

XSL•FO
RenderX

42. TSMM (Text & Social Media Mining) Lab - Yonsei University. PKDE4J URL: http://informatics.yonsei.ac.kr/pkde4j/ [accessed 2019-06-03] [WebCite Cache ID 78rGLBh9h]
43. Stanford CoreNLP. Introduction to pipelines URL: https://stanfordnlp.github.io/CoreNLP/pipelines.html [accessed 2019-06-03] [WebCite Cache ID 78rGV2Gpd]
44. Song M, Kim W, Lee D, Heo G, Kang K. PKDE4J: Entity and relation extraction for public knowledge discovery. J Biomed Inform 2015 Oct:320-332. [Medline: 26277115]
45. Gephi - The Open Graph Viz Platform. URL: https://gephi.org/ [accessed 2019-06-03] [WebCite Cache ID 78rGm1XPj]
46. Wiktionary. Disease URL: https://en.wiktionary.org/wiki/disease [accessed 2019-06-03] [WebCite Cache ID 78rGytBCe]
47. Zappavigna M. Academia.edu. Discourse of Twitter and Social Media: How we use language to create affiliation on the web URL: https://www.academia.edu/18311721/Discourse_of_Twitter_and_Social_Media_How_we_use_language_to_create_affiliation_on_the_web [accessed 2019-06-03] [WebCite Cache ID 78rH6DFwW]
48. Simard S, Savard J. Fear of Cancer Recurrence Inventory: development and initial validation of a multidimensional measure of fear of cancer recurrence. Support Care Cancer 2009 Mar;17(3):241-251. [doi: 10.1007/s00520-008-0444-y] [Medline: 18414902]
49. Groenvold M, Petersen M, Aaronson N, Arraras J, Blazeby J, Bottomley A, et al. EORTC QLQ-C15-PAL: the new standard in the assessment of health-related quality of life in advanced cancer? Palliat Med 2006 Mar;20(2):59-61. [Medline: 16613400]
50. Yanez B, Pearman T, Lis C, Beaumont J, Cella D. The FACT-G7: a rapid version of the functional assessment of cancer therapy-general (FACT-G) for monitoring symptoms and concerns in oncology practice and research. Ann Oncol Off J Eur Soc Med Oncol 2013 Apr;24(4):1073-1078. [Medline: 23136235]
51. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. nitial adult health item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS™) network: 2005–2008. J Clin Epidemiol 2010 Nov;63(11):1179-1194 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
52. Denecke K, Nejdl W. How valuable is medical social media data? Content analysis of the medical web. Inf Sci 2009 May 30;179(12):1870-1880. [doi: 10.1016/j.ins.2009.01.025]
53. Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, Gonzalez GH. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. J Biomed Inform 2016 Aug;62:148-158 [FREE Full text] [doi: 10.1016/j.jbi.2016.06.007] [Medline: 27363901]

Abbreviations

ALL: acute lymphoblastic leukemia
AML: acute myeloid leukemia
BRCA2: breast cancer type 2
CLL: chronic lymphocytic leukemia
CML: chronic myeloid leukemia
COPD: chronic obstructive pulmonary disease
C-topic: conditional topic
DCIS: ductal carcinoma in situ
FWIW: For What it's Worth
GERD: gastroesophageal reflux disease
LDA: latent Dirichlet allocation
LOL: laughing out loud
MRSA: methicillin-resistant *Staphylococcus aureus*
TSH: thyroid stimulating hormone

Edited by G Eysenbach; submitted 22.11.18; peer-reviewed by A Sarker, A Kulanthaivel, M Waring, E Van Den Broek-Altenburg; comments to author 21.12.18; revised version received 06.05.19; accepted 21.05.19; published 13.06.19

Please cite as:
Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M
Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases With Extracted Entities and Their Relations
J Med Internet Res 2019;21(6):e12876
URL: http://www.jmir.org/2019/6/e12876/
doi: 10.2196/12876
PMID: 31199327

XSL•FO
RenderX

©Vasiliki Foufi, Tatsawan Timakum, Christophe Gaudet-Blavignac, Christian Lovis, Min Song. Originally published in the Journal of Medical Internet Research (http://www.jmir.org), 13.06.2019. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on http://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

7.2.5. Conference Article 1

710 Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)
© 2018 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-852-5-710

Automatic Annotation of French Medical Narratives with SNOMED CT Concepts

Christophe GAUDET-BLAVIGNAC [a,1], Vasiliki FOUFI [a],
Eric Wehrli [b], and Christian LOVIS [a]
[a] *Division of Medical Information Sciences, Geneva University Hospitals and University of Geneva*
[b] *Laboratoire d'Analyse et de Technologie du Langage, University of Geneva*

Abstract. Medical data is multimodal. In particular, it is composed of both structured data and narrative data (free text). Narrative data is a type of unstructured data that, although containing valuable semantic and conceptual information, is rarely reused. In order to assure interoperability of medical data, automatic annotation of free text with SNOMED CT concepts via Natural Language Processing (NLP) tools is proposed. This task is performed using a hybrid multilingual syntactic parser. A preliminary evaluation of the annotation shows encouraging results and confirms that semantic enrichment of patient-related narratives can be accomplished by hybrid NLP systems, heavily based on syntax and lexicosemantic resources.

Keywords. Interoperability, narrative data, SNOMED CT, NLP

1. Introduction

Medical data is composed of both structured and unstructured data. Narratives are a type of unstructured data that contains crucial semantic information but not readily usable by computers. Moreover, narratives constitute a challenge for interoperability between healthcare systems, hospitals and departments.[1] SNOMED CT (henceforth SCT), created in 2002, constitutes a terminology organized as a directed graph with concepts as nodes and relationships as edges. Currently, SCT contains more than 350,000 concepts. Property rights and developments are held by SNOMED International (London, UK). The goal of SNOMED International is to develop SCT and to ensure that it becomes the "most comprehensive and precise common global language for health terms in the world"[2]. Since SCT supports post-coordination, i.e. a formal grammar that can associate existing concepts, qualifiers, and predicates, it has similar properties to a natural language. In this paper, a method for automatic annotation of French medical narratives with SCT codes is proposed.

[1] Corresponding author. E-mail : christophe.gaudet-blavignac@unige.ch

2. Related work

Processing medical data with various terminologies, and recently SCT, has been a research focus of other studies as well[3], [4]. Those studies have pursued two kinds of goals. The first goal was the classification of documents, such as pathology or radiology reports[5], [6] in categories related to the disease mentioned in the text. The second one was a more general information retrieval task that aimed at extracting codes or annotating free text with concepts[7]. Commonly used terminologies are the ICD10[8], the UMLS[9] or SCT[10]. In some cases, preliminary work involves creating a subset of those terminologies in relation to a specific goal. It is the case with the UMLS because its Metathesaurus contains more than 100 terminologies, classifications and thesauri.

The methods used to annotate or classify free-text documents vary. Rule-based methods need to be manually or semi-manually developed but require no training corpus and can produce very satisfying results when combined in a pipeline[11], [12]. On the other hand, machine learning and statistical methods, such as Naïve Bayes or Support Vector Machine, do not require the manual creation of rules. However, access to large gold standard corpora used as training sets is essential[13]. Hybrid NLP systems integrating both statistical and linguistic approaches have also been proven very efficient at NLP tasks targeting the medical language[13]. The work presented in this article differs in several ways from the studies previously mentioned. First, the language of the free-text documents used in those references is mostly English. Working with another language requires to translate the terms and adapt the rules to the specificities of the target language. Second and most important, the absence of syntactic-semantic parsing of the text to detect terms in different morphological or syntactic structures makes the method presented in this paper innovative. Our system performs analysis of free medical text in French on the morphological, syntactic and semantic level and annotates the recognized terms with SCT concepts simultaneously.

3. Method

In this research, SCT is approached as a natural language. Automatic annotation of narratives with SCT concepts therefore requires the processing of texts using NLP tools.

3.1. Tool

The tool used for this goal is the hybrid multilingual syntactic parser *Fips* [14]. It relies on generative grammar concepts and is made of a generic parsing module which can be refined to suit the specific needs of a particular language or sublanguage. The lexicon is one of the key components of the parser. It contains detailed morphosyntactic and semantic information, selectional properties, valency information, and syntactic-semantic features that influence the syntactic analysis. To achieve automatic annotation of medical narratives, modifications were needed to correctly process the specificities of the French medical language such as abbreviations or technical terms.

3.2. Creation of electronic dictionaries

Specific lexicons have been developed and incorporated in the parser:

a) A French medical language dictionary was created by extracting simple words and collocations from a corpus of discussions of 11,000 discharge summaries from the internal medicine division of the University Hospitals of Geneva during 2012 to 2014. In its current version, the lexicon comprises 4,454 simple words and 5,640 collocations (groups of words) manually processed.
b) A SCT dictionary. To perform automatic annotation of French narratives with SCT codes, the SCT terminology was added as a new language in the parser. 173,067 SCT concepts and their equivalent code were entered in this dictionary.
c) A bilingual French-SCT dictionary. In the aim of automatic annotation, the target language (SCT) must be linked to the source language (French medical language) in a bilingual dictionary. In the current version of the system, 5,842 medical terms have been mapped to SCT concepts.

3.3. Automatic annotation

In this research, the automatic annotation procedure consists of parsing the initial text and recognizing medical terms. Then, the system looks up the dictionaries (both monolingual and bilingual) and proceeds to the SCT code attribution. Terms in medical terminologies can be affected by syntagmatic and paradigmatic variation to different degrees or may be too precise or complex to actually be used in electronic health records[15]. By providing syntactic analysis and a proper recognition of collocations, the parser can detect concepts regardless of the specific morphological or syntactic form under which they appear in the text. Table 1 shows an example of a sentence annotated with SCT concepts:

Table 1. Example of SCT annotation

Initial phrase	SCT Annotation
En raison des douleurs abdominales, un traitement de morphine iv est débuté et les traitements habituels du patient sont poursuivis	{21522001 | douleur abdominale |}, {373529000 | morphine |}, {255560000 | intraveineux |}, {40451002 | habituel |}, {116154003 | patient |}, {266714009 | poursuivre le traitement |}

We can observe that the system is capable of recognizing structures in various forms, i.e. *iv,* the abbreviated form of *intraveineux* 'intravenous'. It can also identify complex structures even if their constituents do not follow the canonical order and are found in different positions, i.e. the verbal collocation *poursuivre un traitement* 'continue a treatment', *les traitements ... sont poursuivis* 'the treatments … are being continued'.

4. Results

4.1. Automatic annotation

Automatic annotation using the syntactic parser was performed on a corpus of 11,000 discharge summaries. Table 2 below displays the results of the automatic annotation procedure.

Table 2. Automatic annotation of a corpus of 11,000 discharge summaries

Words	4481,191
Annotated terms	892,787
Unique SCT concepts	7,569
Annotated terms per sentence	4,17

4.2. Preliminary evaluation

A preliminary evaluation was completed on a small corpus of five discharge summaries (1,820 words) written by 4 different clinicians, chosen randomly. The corpus was first de-identified (i.e. Protected Health Information (PHI) was removed) and then manually annotated with SCT concepts by one expert. The concepts used for the annotation were selected from the set of codes that are incorporated in the parser's SCT dictionary. The same corpus was processed by the parser and 421 medical terms were automatically annotated. Then, a comparison of the two outputs was performed manually in order to evaluate the system. The performance of the system is very encouraging since precision of 0.7173 and recall of 0.517 were achieved. However, an evaluation on a bigger corpus would allow a more precise measurement of the efficiency of the method.

5. Discussion

5.1. Annotation procedure

The rules used to annotate a narrative with SCT concepts are subject to debate. Since the terminology is structured as a graph with a treelike disposition, there are various levels of granularity for each concept. For instance, *douleur abdominale* 'abdominal pain' could be annotated with a unique SCT code (*21522001*, cf. Table 1) or could be annotated several more specific concepts (*22253000* | *douleur* 'pain' |, *277112006* | *abdominal* 'abdominal' |). At the current stage of the research, the annotation was performed choosing the concept that corresponded to the largest text structure.

5.2. Limitations and future work

Medical documents contain sensitive information, as a consequence access to corpora and in particular annotated corpora, is a well-known challenge in this field. This is especially true for languages other than English. The size of the evaluation corpus is one of the major limitations of this paper. In addition, evaluation of medical free-text annotation must be performed in a specific setting to affirm that the results are reliable. The manual annotation task, in particular, should be performed by at least two annotators not directly involved in the development of the automatic annotation tool to avoid bias. Having more than one annotator is important to compute the inter-annotator agreement and set an upper-bound on the annotation task. The annotation of French narratives with SCT concepts is a first step toward the ultimate goal which is the complete representation of patient-related narratives into a formal language. The next step in this research will be the processing of post-coordinated concepts according to the SCT compositional grammar. Post-coordination will enable the storage of the full information contained in the text into SCT post-coordinated sentences.

6. Conclusion

In this paper, a method to annotate French medical free-text with SCT concepts is proposed. This method relies on a syntactic-semantic parser specifically modified to meet the needs of this task. Lexico-semantic resources (monolingual and bilingual dictionaries as well as grammar rules) were constructed taking into consideration the specificities of the French medical language. A preliminary evaluation has shown encouraging results with a precision of 0.7173, a recall of 0.5171 and an F-score of 0.6009. Further research is needed to produce post-coordinated structures and full representation of medical narratives into SCT.

Acknowledgements

This research has been financed by the "Réseau Thématique Langage & Communication" from the University of Geneva.

References

[1] S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, and C. U. Lehmann, "Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress."
[2] "Support : SNOMED International." [Online]. Available: https://ihtsdo.freshdesk.com/support/home. [Accessed: 20-Mar-2017].
[3] J. Patrick, Y. Wang, and P. Budd, "An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology."
[4] P. Ruch, J. Gobeill, C. Lovis, and A. Geissbühler, "Automatic medical encoding with SNOMED categories.," *BMC Med. Inform. Decis. Mak.*, vol. 8 Suppl 1, p. S6, 2008.
[5] G. Zuccon *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. AMIA Summit Transl. Sci.*, vol. 2013, pp. 300–4, Jan. 2013.
[6] A. Nguyen, J. Moore, G. Zuccon, M. Lawley, and S. Colquist, "Classification of pathology reports for Cancer Registry notifications," *Stud. Health Technol. Inform.*, vol. 178, no. May 2014, pp. 150–156, 2012.
[7] M. Torii, K. Wagholikar, and H. Liu, "Using machine learning for concept extraction on clinical documents from multiple data sources.," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 18, no. 5, pp. 580–587, 2011.
[8] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson, "Automatic ICD-10 classification of cancers from free-text death certificates," *Int. J. Med. Inf.*, vol. 84, pp. 956–965, 2015.
[9] B. Riedl, N. Than, and M. Hogarth, "Using the UMLS and Simple Statistical Methods to Semantically Categorize Causes of Death on Death Certificates.," *AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp.*, vol. 2010, pp. 677–81, 2010.
[10] P. Ruch, J. Gobeill, C. Lovis, and A. Geissbühler, "BMC Medical Informatics and Decision Making Automatic medical encoding with SNOMED categories."
[11] D. De Meyere *et al.*, "Automatic annotation of medical reports using SNOMED-CT: a flexible approach based on medical knowledge databases," 2015.
[12] A. R. Aronson, "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program."
[13] R. J. Kate, "Towards Converting Clinical Phrases into SNOMED CT Expressions.," *Biomed. Inform. Insights*, vol. 6, no. Suppl 1, pp. 29–37, 2013.
[14] E. Wehrli and L. Nerima, "The fips multilingual parser," in *Language Production, Cognition, and the Lexicon*, Springer, 2015, pp. 473–490.
[15] C. Hansart, D. De Meyere, P. Watrin, A. Bittar, and C. Fairon, "CENTAL at SemEval-2016 Task 12: a linguistically fed CRF model for medical and temporal information extraction," *Proc. 10th Int. Workshop Semantic Eval. SemEval-2016*, pp. 1286–1291, 2016.

7.2.6. Conference Article 2

Reconnaissance et représentation automatiques de concepts médicaux français en SNOMED CT

Christophe Gaudet-Blavignac[a], Vasiliki Foufi[a], Eric Wehrli[b], Christian Lovis[a]

[a] Sciences de l'Information Médicale (SIMED), Hôpitaux Universitaires de Genève, Université de Genève
[b] Laboratoire Automatique de Traitement du Langage (LATL), Centre Universitaire d'Informatique, Université de Genève

Résumé

L'interopérabilité des données médicales est primordiale pour le partage et la réutilisation de ces données à des fins de contrôle qualité, de recherche, ou d'aide à la décision. Une première étape vers cette interopérabilité est la transformation de données en un langage formalisé. SNOMED CT présente les caractéristiques d'un langage naturel et permet une représentation formelle de la majorité des concepts médicaux. Ce travail porte sur la représentation automatique de textes médicaux en concepts SNOMED CT post-coordonnés. Pour effectuer cette tâche, un outil de traitement automatique du langage naturel a été adapté et des ressources lexico-sémantiques ont été créées. L' évaluation de 1 500 expressions simples et post-coordonnées montre une précision de 85% pour les expressions simples et de 71,2% pour les expressions post-coordonnées.

Mots-clés :

Semantics, SNOMED CT, Natural Language Processing

Introduction

La quantité de données générées par les structures de santé augmente chaque année. Des efforts existent depuis de nombreuses années afin de rendre ces données interopérables et réutilisables. Des initiatives telles que le Health Information Technology for Economic and Clinical Health (HITECH) Act aux États-Unis à travers les objectifs du « Meaningful use » [1] ont essayé de pousser les établissements à intégrer un certain nombre de pratiques ou de standards afin d'accompagner la transition numérique des hôpitaux. Des organisations telles que HL7 [2] définissent des spécifications techniques pour permettre l'interopérabilité des données médicales. Si ces acteurs agissent du côté technique, l'International Health Terminology Standards Development Organisation (IHTSDO), récemment renommée SNOMED International [3], œuvre pour définir une ressource sémantique unique pour les données médicales. En publiant SNOMED CT, cette organisation espère développer le « langage global de la santé ».

Il devient impératif de proposer des solutions permettant le partage et la réutilisation de ces données à des fins de contrôle qualité, de recherche, ou d'aide à la décision.

Malgré les initiatives citées plus haut, force est de constater que les données médicales sont loin d'être facile à standardiser et à réutiliser. Si de nombreux éléments du dossier médical sont maintenant entrés de manière structurée (laboratoires, échelles thérapeutiques, prescriptions), de nombreux documents ou des segments sont encore sous forme de texte libre. [4] Ces textes contiennent néanmoins d'importantes informations cliniques et devraient pouvoir être exploités pour une meilleure automatisation des processus hospitaliers et pour une réutilisation des données.

Afin de pouvoir réutiliser ces données textuelles, il est nécessaire de les transformer en un format approprié à un traitement informatique sans toutefois dénaturer leur contenu sémantique. Pour pallier au problème de la dénaturation de la donnée qu'amène la transformation d'un texte en une liste de concepts, il est nécessaire d'utiliser un standard permettant une composition des concepts. Ce standard doit représenter la complexité du texte libre au moyen d'un langage formel, utilisable par une machine et incluant une sémantique forte. En produisant des concepts post-coordonnés, il est possible de représenter des expressions de plus en plus complexes, avec le but final de représenter chaque phrase du document par un seul concept post-coordonné. Cette approche permet d'aborder le défi de la standardisation comme une tâche de traduction automatique plutôt que comme une reconnaissance d'entités nommées. Par ailleurs ce résultat peut être obtenu en s'appuyant sur des méthodes classiques de projections sur dictionnaires, associées à une post-coordination reposant sur la grammaire cible et l'analyse syntactico-sémantique de la phrase dans le langage source.

Si de nombreux travaux se sont attelés à la tâche de reconnaissance d'entités nommées dans les textes biomédicaux au moyen de classifications standards [5–7], en ce qui concerne SNOMED CT, rares sont ceux qui vont jusqu'à la post-coordination. Certains travaux ont produit des concepts post-coordonnés [8–10], cependant ces combinaisons sont faites manuellement ou sans utiliser les relations et la grammaire de SNOMED CT.

D'autres travaux utilisent des méthodes couramment utilisées en traduction automatique telle que les réseaux neuronaux récurrents pour des tâches de codage automatique des entités nommées [11]. Cependant ils nécessitent de grands corpus annotés et ne proposent pas encore de concepts post-coordonnés. En outre, les problèmes éthiques liés aux documents médicaux rendent leur utilisation et leur partage difficile. [12] Pour finir, la majorité de travaux dans ce domaine porte sur des textes issus de la littérature biomédicale anglo-saxonne [13] et les systèmes de traitement du langage naturel disponibles concernent majoritairement l'anglais, la langue la plus utilisée dans la recherche.

Dans cet article, une méthode pour représenter le texte médical français en des concepts post-coordonnés en SNOMED CT est proposée. Le développement des ressources lexico-sémantiques nécessaires à son application y est décrit. Cette méthode est appliquée à un corpus et une validation des résultats est effectuée.

Méthode

Afin d'aborder le problème de l'interopérabilité des données comme une tâche de représentation formalisée de concepts, une série d'opérations doit être effectuée :

a) Une terminologie compositionnelle doit être choisie.
b) Les termes doivent être reconnus et analysés au niveau morphologique.
c) Une analyse syntactico-sémantique des phrases doit être produite.
d) Une annotation de la phrase avec des concepts de la terminologie doit être effectuée.
e) Les concepts doivent être articulés entre eux en respectant la grammaire compositionnelle de la terminologie pour produire des concepts post-coordonnés.

SNOMED CT

Issu de la fusion en 2002 des Clinical Terms Version 3, connus aussi sous le nom de Read Codes [14] et de la SNOMED Reference Terminology (SNOMED-RT) [15], SNOMED CT est une terminologie médicale contenant plus de 340 000 termes. [16] Elle est maintenue et éditée par SNOMED International, une organisation composée de représentants des états-membres du projet et payant une licence. Les concepts de SNOMED CT sont organisés en un graphe avec des arrêtes typées sémantiquement. Les arrêtes sont appelées des attributs et relient un concept à un autre. Chaque concept est au moins relié à un autre concept par un attribut « is-a ».
Une grammaire compositionnelle [17] est également fournie avec la terminologie et permet la combinaison de concepts afin d'exprimer des concepts absents de la terminologie. La création de ces nouveaux concepts s'appelle la post-coordination. Par exemple, le concept d' « éruption cutanée du pied droit » est absent de SNOMED CT, mais peut être exprimé par la combinaison des concepts « Eruption of skin » (SNOMED CT : 271807003) et « Structure of right foot » (SNOMED CT : 7769000). La grammaire SNOMED CT est définie dans le métalangage ABNF (Augmented Backus-Naur Form). Des spécifications supplémentaires liées aux concepts utilisables et à la portée des attributs sont également disponibles. Cette grammaire a été utilisée afin de produire une représentation du texte en SNOMED CT.

FipsSnomed

Dans cette étude, la représentation du texte en langage formalisé est approchée comme une tâche de traduction automatique via l'analyseur syntactico-sémantique Fips, développé pour des applications pratiques dans le domaine du traitement automatique du langage comme la traduction assistée par ordinateur ou le traitement de la parole [18]. Cet outil, né au sein de l'Université de Genève, se fonde sur les concepts de la grammaire générative et est composé d'un module d'analyse syntaxique générique, qui peut être raffiné afin de répondre à des besoins spécifiques à une langue ou à un sous-langage ; pour notre recherche, une version baptisée FipsSnomed a été développée en vue d'un traitement efficace des documents médicaux.

Analyse syntactico-sémantique

Comme l'analyseur Fips, FipsSnomed traite un corpus phrase à phrase. Pour chaque phrase, il construit une représentation syntaxique complète ou une séquence d'analyses partielles (en cas d'échec d'analyse complète), sous la forme d'une structure arborescente dont les extrémités (les feuilles) sont les items lexicaux (mots simples et composés). Un item lexical contient toute l'information issue de la base de données lexicale (mot, lexème, collocation), ainsi qu'un lien éventuel sur un concept SNOMED CT.

Analyse lexicale

Le lexique est un constituant fondamental de l'analyseur. Il contient des informations morphologiques, syntaxiques et sémantiques sur les entrées, ainsi que des informations de valence.
En plus des dictionnaires de français standard utilisés par le système pour l'analyse de textes, des dictionnaires électroniques de termes médicaux ont été développés. En particulier :

a) Un dictionnaire de 4 454 termes médicaux simples et 5 640 collocations (termes à plusieurs constituants) a été créé semi-manuellement. Les lemmes ont été extraits d'un corpus de 11 607 segments narratifs de lettres de sortie issu de l'étude présentée dans [19].
b) Un dictionnaire de termes SNOMED CT. Afin d'annoter de manière automatique les textes médicaux avec des codes, la terminologie SNOMED CT a été entrée dans l'analyseur multilingue en tant qu'une nouvelle langue naturelle. Plus de 173 000 concepts SNOMED CT et leur code équivalent ont été inclus dans le dictionnaire SNOMED CT.
c) Un dictionnaire bilingue Français-SNOMED CT. Pour effectuer l'annotation automatique et la représentation en concepts post-coordonnés, la langue cible (SNOMED CT) doit être liée à la langue source (Français médical) dans un dictionnaire bilingue. Dans la version actuelle du système, 5 842 termes médicaux ont été mis en correspondance avec des concepts SNOMED CT.

Annotation Automatique

L'annotation automatique, première étape de la représentation du texte, consiste en une analyse morphologique et syntaxique du texte initial et une reconnaissance des termes médicaux. Le système procède par la suite à une recherche dans les dictionnaires monolingues et multilingues afin d'attribuer un code SNOMED CT. Les concepts dans une terminologie médicale peuvent être sujets à des variations syntagmatiques et paradigmatiques à des degrés divers. Pour pallier à cela, l'analyseur effectue une analyse morphologique et syntaxique ainsi qu'une reconnaissance des collocations nominales et verbales, ce qui lui permet de reconnaitre des concepts et leurs variantes. Par exemple le concept « 266714009 | poursuivre le traitement | » apparaissant dans la phrase « le traitement habituel du patient est poursuivi » va être annoté en utilisant d'une part ce concept présent dans le système sous forme de collocation avec les informations syntaxiques correspondantes (groupe verbal composé du verbe « poursuivre » et d'un complément composé d'un article et du nom « traitement ») et d'une autre part l'analyse morphosyntaxique de la phrase dans laquelle le système reconnait le même groupe verbal mais conjugué et à la forme passive. La correspondance étant faite, le système assigne le code à la structure « le traitement […] est poursuivi ».

Post-coordination des concepts

FipsSnomed effectue la post-coordination SNOMED CT en se basant sur l'analyse morphosyntaxique qu'il a produite et sur la grammaire compositionnelle publiée par SNOMED CT. Elle est construite par lecture récursive de l'arborescence syntaxique. Pour chaque tête lexicale reliée à un concept SNOMED CT, une nouvelle annotation est initiée avec au minimum un concept non post-coordonné. Lorsque la lecture rencontre un terme portant un trait lexical spécifique lié à un attribut (« bodypart » lié à « finding site ») et correspondant à un concept SNOMED CT, ce concept est associé au concept en cours de construction par l'attribut concerné. Le concept en construction est en général la tête lexicale principale du syntagme nominal complet. Par exemple, pour le groupe nominal « ablation du rein droit », le système en construisant le concept « ablation » rencontre le concept « rein » qu'il lie par l'attribut « finding

site » au premier concept. Puis, il rencontre le concept « droit » qu'il lie à « rein » par l'attribut « latéralité », et ainsi de suite. Dans la version actuelle du système, trois types d'attributs sont générés : « finding site », « laterality » et « severity ».
Dans le cas de multiples concepts liés avec une conjonction de coordination comme par exemple « poumons gauche et droit », deux concepts post-coordonnés seront produits « poumon gauche » et « poumon droit ».

Evaluation

Afin d'évaluer l'annotation automatique ainsi que la post-coordination des concepts, un corpus de 11 607 segments narratifs issus de lettres de sortie a été utilisé. Chaque segment correspond à la partie résumant le séjour dans un document. Ce corpus a été traité par l'analyseur FipsSnomed avec les étapes de traitements suivantes : analyse morphosyntaxique des phrases, annotation automatique des concepts SNOMED CT, post-coordination des concepts. Du résultat de ce traitement, deux groupes aléatoires de résultats ont été extraits. Cette extraction a été faite par traitement du fichier texte contenant les résultats au moyen d'expressions régulières. La composition de ces deux groupes est la suivante : un groupe de 1 000 concepts annotés non post-coordonnés et un groupe de 500 annotations post-coordonnées. Pour ce dernier set, seul les annotations comportant des concepts correctement annotés ont été sélectionnées afin d'évaluer uniquement la justesse de l'attribut. De cette manière, les post-coordinations faites sur des concepts incorrectement annotés ne sont pas considérés dans l'évaluation. Ces annotations ont été évaluées par triplets composés de deux concepts reliés par un attribut. Certaines annotations étant composées de plusieurs post-coordinations, l'évaluation a porté sur 573 triplets au total. Un expert a ensuite validé manuellement la tâche effectuée par FipsSnomed sur ces deux sets. L'évaluation a été effectuée par comparaison de la sémantique du segment annoté dans le texte avec la sémantique du concept SNOMED CT assigné. Pour la post-coordination, la justesse de la relation créée entre deux concepts a été évaluée. La précision globale et spécifique à chaque type de concept et de relation a été ensuite calculée.

Résultats

Le corpus a été traité par l'outil sur un ordinateur avec système d'exploitation Windows 10, un processeur Intel Core i5-6300U 2,40GHz, et une mémoire RAM de 8,00 Go. La durée du traitement a été de 5 heures, 5 minutes et 48 secondes.

Annotation Automatique

Le tableau 1 montre des exemples factices annotés au moyen de FipsSnomed :

Tableau 1 – Exemple d'annotation SNOMED CT

Phrase initiale	Annotation SNOMED CT
En raison des douleurs abdominales, un traitement de morphine iv est débuté et les traitements habituels du patient sont poursuivis.	{21522001 | douleur abdominale |}, {373529000 | morphine |}, {255560000 | intraveineux |}, {40451002 | habituel |}, {116154003 | patient |}, {266714009 | poursuivre le traitement |}
Le patient présente une hypoesthésie du pied.	{116154003 | patient |}, {397974008 | hypoesthésie | : 363698007 | finding site | = 259051005 | pied |}

Le tableau 2 présente les résultats de l'annotation automatique de FipsSnomed sur le corpus. Ces résultats représentent le nombre de concepts qui ont été annotés dans le texte puis le nombre de post-coordinations qui ont été créées entre ces concepts annotés. Le nombre de post-coordinations représente donc le nombre de groupes de concepts liés par des attributs.

Tableau 2 – Résultats de l'annotation automatique

Mots	4 481 191
Phrases	213 808
Nombre de phrases moyen par lettre	18,42
Taux d'analyses de phrases complètes	58,13% (124 284)
Concepts annotés	950 158
Post-coordinations créées	24 530
Concepts SNOMED CT uniques	7 569
Nombre moyen d'annotations par phrases	4,17

Evaluation

Concepts

Le tableau 3 résume les résultats de l'annotation des concepts non post-coordonnés.

Tableau 3 – Evaluation de l'annotation des concepts non post-coordonnés

Type de concept (semantic tag)	Nombre de concepts	Nombre de concepts corrects	Précision en %
qualifier value	405	352	86,91
attribute	167	145	86,83
finding	72	52	72,22
person	59	59	100,00
procedure	55	39	70,91
disorder	53	52	98,11
substance	45	43	95,56
observable entity	38	30	78,95
morphologic abnormality	25	20	80,00
environment	20	7	35,00
body structure	15	13	86,67
physical object	11	5	45,45
occupation	10	10	100,00
specimen	5	5	100,00
organism	4	4	100,00
event	4	4	100,00
situation	3	3	100,00
physical force	2	0	0,00
social concept	2	2	100,00
product	2	2	100,00
navigational concept	1	1	100,00
cell structure	1	1	100,00
medicinal product	1	1	100,00
Total général	**1 000**	**850**	**85,00**

Post-coordination

Le tableau 4 résume les résultats de l'annotation des concepts post-coordonnés en fonction du type de relation.

Tableau 4 – Evaluation de la post-coordination

Type de relation	Nombre de post-coordination	Nombre de post-coordinations correctes	Précision en %
Severity	143	94	65,73
Laterality	233	188	80,69
Finding Site	197	126	63,96

Total général	**573**	**408**	**71,20**

Discussion

Annotation automatique

Sur le corpus décrit dans la méthode, l'analyseur FipsSnomed a produit 950 158 annotations, que ce soit des annotations simples ou post-coordonnées. Cela représente plus de 4 annotations par phrase. Ce premier résultat tend à montrer que l'analyseur parvient à détecter un nombre satisfaisant de concepts SNOMED CT dans du texte médical. Le nombre de post-coordinations étant de 24 530, cela représente 2,58% des annotations. Cette proportion doit être mise en perspective avec le fait que seuls 3 types de relations ont été implémentés dans le système ce qui représente 0,24% des relations présentes dans SNOMED CT (1 209).

Concepts

Les résultats d'annotations (voir tableau 1) montrent que le système est capable de reconnaître des structures de formes variées comme des abréviations (iv-intraveineux) ou des termes dont les constituants ne sont pas dans leur forme canonique ou leur ordre syntaxique habituel. La collocation verbale « poursuivre un traitement » est reconnue même sous la forme « les traitements … sont poursuivis ».
La précision globale de l'annotation des concepts est de 85%. Ce résultat est très encourageant et montre que l'analyseur parvient à annoter un texte français avec des concepts SNOMED CT en respectant la sémantique du texte.
Les concepts les plus représentés dans le set d'évaluation sont les « qualifier value ». Ces concepts sont souvent des adjectifs comme « stable », « pulmonaire », « thérapeutique », « nouveau ». Ces concepts sont détectés avec une précision de 86,91% ce qui est en harmonie avec les résultats globaux.
Les attributs sont la deuxième classe de concepts la plus représentée. Ces concepts représentent des relations en SNOMED CT. Ils ne devraient apparaître que dans une post-coordination. Cependant leur détection dans le texte pourra permettre une post-coordination plus globale quand le nombre d'attributs sera étendu.
Les concepts de type « disorder » représentent environ 10% des concepts évalués avec une précision de 98,11%. Il est très important qu'un système qui annote les concepts médicaux soit précis sur les maladies annotées car cette information est cruciale dans un dossier médical. D'un autre côté, la catégorie « finding » qui représente les signes et symptômes observés a une précision de 70,91%. Ce score plus faible peut s'expliquer par la nature polysémique de certains mots français comme « suivi », « crise » ou « tension » et également par la nature très large de cette catégorie qui peut contenir un très grand nombre de concepts.
La catégorie « person » est détectée avec une précision de 100%. Cela est dû en partie à la présence fréquente du mot « patient » dans les textes mais montre aussi que le système est capable de reconnaitre les mots décrivant des intervenants de la santé tels que les pneumologues ou les infectiologues.

Post-coordination

La tâche de la post-coordination montre une précision de 71,2% sur le set d'évaluation. Ce résultat varie en fonction de la relation ciblée. D'une manière générale, l'erreur la plus fréquente concerne l'assignation d'une relation sur un concept qui ne devrait pas le permettre comme c'est le cas de certains « qualifiers ». Les concepts « avec » ou « révélé » n'acceptent jamais une relation d'un des trois types implémentés dans le système. Ces erreurs pourraient être réduites en créant des règles de post-coordination plus strictes comme celles présentes dans le guide éditorial de SNOMED CT [20]. La relation « laterality » est celle qui obtient de meilleurs résultats avec 80,69% de précision. Cela s'explique par la nature plus simple de la détection d'une relation de latéralité. En effet, les concepts ciblés par cette relation sont au nombre de deux : « droit » et « gauche ». Pour la relation « finding site », la précision est de 63,96%. Les erreurs sont dues le plus souvent à une relation liant un concept à une localisation présente plus loin dans la phrase mais indépendante du premier concept. Par exemple : « Une tomodensitométrie abdominale révèle un foie cirrhotique » produit le triplet : « Tomodensitométrie : finding site : foie ». Cette erreur est due à une trop grande permissivité du système sur la création des post-coordinations. Afin d'améliorer cela il serait nécessaire de mieux restreindre les concepts pouvant être coordonnés en fonction de la relation syntaxique qu'ils entretiennent avec la tête lexicale. Finalement, la relation « severity » obtient une précision similaire à « finding site » avec 65,73% avec des erreurs similaires.

Cas ambigus

Le système arrive à lever quelques types d'ambiguïtés grâce à :

a) la reconnaissance des collocations répertoriées dans le lexique (p. ex. attribution de deux types de concepts différents au mot simple polysémique « membre » dans les cas de « membre de la famille » (person) et « membre inférieur » (body structure) ;
b) l'analyse syntaxique (p. ex. distinction entre le nom « patient » qui correspond au type de concept « person » et l'adjectif « patient »).

Par contre, dans les cas où une analyse contextuelle serait nécessaire pour lever l'ambiguïté, le système va produire une annotation potentiellement fausse. C'est par exemple le cas du concept « hospitalisation » qui désigne soit l'admission d'un patient à l'hôpital soit le séjour complet d'un patient à l'hôpital. Dans ce cas précis, le système dans sa version actuelle, assignera toujours le même concept qui sera le premier qui a été inséré dans son dictionnaire (en l'occurrence, « 32485007 Hospital admission (procedure) »).

Limitations

Sur le corpus utilisé, le taux de phrases disposant d'une analyse complète était de 58,13%. Même si le système n'a pas besoin d'une analyse complète pour annoter un concept, il est possible que cela ait impacté la détection de concepts apparaissant dans des portions de phrases complexes.
Le système ne comporte pas actuellement de fonctionnalités qui lui permettent de reconnaître des termes présentant des fautes d'orthographe. Les fautes d'orthographes vont résulter en une non reconnaissance du concept et l'absence d'annotation.
De plus l'évaluation proposée porte sur un petit nombre de concepts proportionnellement au corpus complet 1 000/950 158 concepts simples et 500/24 530 post-coordinations. Une évaluation plus complète est nécessaire pour confirmer ces résultats.
L'évaluation de la tâche ayant été faite par validation des résultats, seule la précision a pu être calculée. Afin d'évaluer de manière plus complète ce travail, il serait nécessaire de définir un gold standard qui permette de calculer le rappel et le F-score. De plus, l'évaluation ayant été faite par un seul expert, il est impossible de calculer l'accord inter-annotateur. Finalement, l'évaluation a été effectuée sur le corpus utilisé pour la création des ressources lexicales et l'adaptation de l'analyseur Fips au français médical. Une évaluation de la performance de FipsSnomed sur un corpus autre que le corpus d'entraînement par au moins deux experts serait nécessaire pour obtenir des résultats plus fiables.

Conclusions

Dans cet article une méthode est proposée pour la représentation des textes médicaux français en SNOMED CT. Elle est appliquée et évaluée sur un corpus clinique. Les résultats de l'annotation automatique et de la post-coordination sont très encourageants et montrent que l'approche du problème de l'encodage du texte en SNOMED CT comme une tâche de traduction en une représentation formelle via un outil de traitement automatique du langage naturel (TALN) permet de produire des concepts post-coordonnés. Les avantages de cette approche sont multiples. Premièrement, un outil de TALN comme l'analyseur syntactico-sémantique FipsSnomed, permet une détection des concepts quelles que soient leurs variations morphosyntaxiques. Ensuite, les informations syntaxiques obtenues par l'analyseur peuvent être utilisées pour la post-coordination. Finalement, cette approche symbolique ne requiert pas la création de corpus annotés, une tâche coûteuse et complexe surtout quand elle concerne des données sensibles telles que les documents cliniques.

Remerciements

Cette recherche a été financée par le Réseau "Langage et Communication" de l'Université de Genève.

Références

[1] C. Group, What is the HITECH ACT? | What HITECH Compliance Means, *Compliancy Group*. (n.d.). https://compliancy-group.com/what-is-the-hitech-act/ (accessed June 12, 2019).
[2] Health Level Seven International - Homepage | HL7 International, (n.d.). https://www.hl7.org/ (accessed June 12, 2019).
[3] SNOMED Home page, *SNOMED*. (n.d.). http://www.snomed.org/ (accessed June 14, 2019).
[4] S. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle, Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research, *IMIA Yearbook of Medical Informatics Methods Inf Med*. **47** (2008) 128–44. doi:me08010128.
[5] M.E. Matheny, F. FitzHenry, T. Speroff, J.K. Green, M.L. Griffith, E.E. Vasilevskis, E.M. Fielstein, P.L. Elkin, and S.H. Brown, Detection of infectious symptoms from VA emergency department and primary care clinical documentation, *International Journal of Medical Informatics*. **81** (2012) 143–156. doi:10.1016/j.ijmedinf.2011.11.005.
[6] P. Jindal, and D. Roth, Extraction of events and temporal expressions from clinical narratives, *Journal of Biomedical Informatics*. **46** (2013) S13–S19. doi:10.1016/j.jbi.2013.08.010.
[7] A.N. Nguyen, J. Moore, M.J. Lawley, D.P. Hansen, and S. Colquist, Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications, *Studies in Health Technology and Informatics*. **168** (2011) 117–124. doi:10.3233/978-1-60750-791-8-117.
[8] D.H. Lee, F.Y. Lau, and H. Quan, A method for encoding clinical datasets with SNOMED CT, *BMC Medical Informatics and Decision Making*. **10** (2010) 53. doi:10.1186/1472-6947-10-53.
[9] A.N. Nguyen, M.J. Lawley, D.P. Hansen, and S. Colquist, Structured pathology reporting for cancer from free text: Lung cancer case study, *Electronic Journal of Health Informatics*. **7** (2012). doi:10.1111/j.1743-7563.2009.01252.x.
[10] H. Liu, K. Wagholikar, and S.T.-I. Wu, Using SNOMED-CT to encode summary level data - a corpus analysis., in: AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science, United States, 2012: pp. 30–37.
[11] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh, Medical concept normalization in social media posts with recurrent neural networks, *Journal of Biomedical Informatics*. **84** (2018) 93–102. doi:10.1016/j.jbi.2018.06.006.
[12] R. Chevrier, V. Foufi, C. Gaudet-Blavignac, A. Robert, and C. Lovis, Use and Understanding of Anonymization and De- identification in the Biomedical Literature: a Scoping Review, (n.d.) 28. doi:10.2196/13484.
[13] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: opportunities and challenges, *J Biomed Semantics*. **9** (2018) 12. doi:10.1186/s13326-018-0179-8.
[14] Read Codes, *NHS Digital*. (n.d.). https://digital.nhs.uk/services/terminology-and-classifications/read-codes (accessed June 13, 2019).
[15] K.A. Spackman, K.E. Campbell, and R.A. Côté, SNOMED RT: a reference terminology for health care, *Proc AMIA Annu Fall Symp*. (1997) 640–644.
[16] S. Bhattacharyya, Introduction to SNOMED CT, Springer Berlin Heidelberg, New York, NY, 2015.
[17] SNOMED CT Compositional Grammar Specification and Guide, (n.d.). https://www.snomed.org/news-articles/snomed-ct-compositional-grammar-specification-and-guide (accessed November 3, 2017).
[18] E. Wehrli, Fips, A "Deep" Linguistic Multilingual Parser, in: 2007: pp. 120–127. https://aclweb.org/anthology/papers/W/W07/W07-1216/ (accessed June 12, 2019).
[19] C. Fehlmann, M. Louis Simonet, J.-L. Reny, J. Stirnemann, and K. Blondon, Associations between early handoffs, length of stay and complications in internal medicine wards: A retrospective study, *European Journal of Internal Medicine*. (2019) S0953620519302225. doi:10.1016/j.ejim.2019.07.003.
[20] IHTSDO, SNOMED CT Editorial Guide, *Development*. (2011) 2002–2011.

Adresse de correspondance
Christophe Gaudet-Blavignac
Christophe.Gaudet-Blavignac@unige.ch
Division of Medical Information Sciences
University Hospitals of Geneva & University of Geneva
Campus Biotech, Chemin des Mines 9
CH-1202 Genève, Suisse
Tél : +41 22 37 90837

8. Conclusions and perspectives

Throughout this work, a new approach for the semantic interoperability of clinical data built around three main hypotheses was developed and evaluated.

8.1. An interlingua for clinical data

The first part of this work focused on solving the common barriers encountered when trying to implement nationwide interoperability for structured data. To overcome the constraints that result from the choice of a single standard, data model or controlled vocabulary, we propose to use a compositional approach centered around semantics, with the usage of SNOMED CT primarily and other relevant controlled vocabularies when necessary. Hence the creation of a set of SNOMED CT encoded variables that can be reused, combined and linked to as many other controlled vocabularies as needed. Doing so, the semantics of the data is ensured and additional variables or controlled vocabulary bindings can be added to fit new use cases. This approach, while raising several challenges such as the choice of the correct granularity for the variables, solves the issue of the multiplicity of controlled vocabularies in healthcare. Moreover, it allows the usage of a descriptive formalism to store the data instead of enforcing a data model. This choice is radical in the sense that it does not answer the usual question of the storage and transfer of the data by providing a constraining definition of a container which would hold the data, but by defining a language that will be used to describe it. It brings the possibility of purpose-specific conversions into data models. The ETL procedures only need to be created from the formalism chosen to each data model, rather than from each data model to every other one.

The SPHN initiative and its three-pillars strategy provided a real-life implementation of this approach. By being successfully adopted by all university hospitals and polytechnical schools in Switzerland, it brought unprecedented possibilities such as the creation of a meta catalog of the variables used by multiple projects across multiple institutions, a follow-up of the compliance of each center with the strategy along with the amount of data shared and used. In the long run, it will leverage the capacity of Switzerland to reuse healthcare data further in secondary or tertiary applications. Its wide adoption lays the foundations for the creation of FAIR (findable, accessible, interoperable, reusable) data endpoints which could replace legacy data warehousing solutions based on relational data models, thus facilitating multicenter research for precision medicine in Switzerland.

Additionally, the work accomplished in this first part had multiple other outcomes. Firstly, it drew the attention of large institutions in Switzerland to the importance of semantics by shifting the interoperability debate from “choosing a standard and enforcing it” to “encoding the semantics of the data, then describing it for storage and transfer”. Indeed, while SNOMED CT is recommended by ehealth Suisse at the national level (226), the knowledge of this terminology was limited in the SPHN realm and this strategy helped to disseminate the idea that semantics were crucial for interoperability and that SNOMED CT was among the best candidates to address the issue.

Secondly, it influenced multiple projects in the SIMED and in the HUG. As an example, as the covid-19 pandemic hit the world, the activity of the SIMED was shut down entirely and redirected to support the effort of the HUG to fight the growing number of hospitalizations and the surging demand for resources. Suddenly, the ability to understand clinical data became crucial to evaluate the situation and predict its evolution. To give an overview of the situation and produce predictions, the SIMED team applied methods directly inherited of the three-pillars strategy, mapping the raw data to variables encoded in SNOMED CT.

Finally, after the first wave of the pandemic, the SIMED was mandated by the Medical Director of the HUG to gather covid-19-related data into a single database for research use. This database named CovidDB was created by a team managed by CL and CGB and designed to be semantically driven, without constraining the data in a specific model but enriching it with SNOMED CT encoding and other controlled vocabularies when relevant. This database is available for any research group of the HUG for feasibility studies and analysis of covid-19-related data.

Overall, this first hypothesis was largely confirmed. Firstly, by the success of the SPHN strategy in bringing a solution for multi-center, multi-community interoperability, then by the changes it triggered at the national, institutional and division level. It is now clear that SNOMED CT, along with a restricted set of relevant information representations can act as the universal semantic bridge for clinical data that we called interlingua.

8.2. A restricted set of useful concepts

The second part of this thesis assumed that while the proposed interlingua allowed a near limitless expressivity, only a small subset of those expressions was useful in practice. Therefore, it should be possible to target them, reducing the set of relevant concepts and semantic dimensions to a useful and manageable size.

The problems list created in the SIMED and deployed in production in 2017 was a perfect use case to test this hypothesis. Limiting the list to expressions found in clinical documents and further refining it according to its usage in the hospital allowed us to converge to a list of 20,120 expressions, encoded in SNOMED CT and multiple other semantic dimensions, that were all used at least once by a clinician or in a project and for which every semantic dimension was used for at least one secondary goal.

The results presented in Article 2 show that this approach was able to bring interoperability to this specific type of data without constraining clinicians to use a controlled vocabulary that does not fit their expectations but keeping only useful expressions. The common list allowed multiple projects that would not have been possible otherwise, creating a virtuous circle in which the more projects used the list and added value to it by proposing new semantic dimensions, the easier it was to use it for new purposes since the information graph created by the expressions and their metadata were rich.

Currently, there are two axes of evolution for this project. Firstly, the common list will continue to grow with each deployment, with new expressions and semantic dimensions. Secondly, the approach taken to build the list could be applied to other use cases beyond the patients' problems where information is commonly entered as short sections of free text in order to broaden the semantic interoperability of clinical data.

Overall, this second part confirms the second hypothesis and completes the approach described in the first part by showing that the expressivity of medical natural language can be reduced to a manageable set of useful semantically enriched expressions. Additionally, this project allowed to bridge the gap between clinicians' language and controlled vocabularies by only proposing expressions used in practice, releasing the clinicians from the task of encoding the information.

8.3. An automatic interlingua translation for clinical narratives

Finally, the last contribution of this thesis, detailed in the third article along with the two conference articles, focused on medical narratives, a type of data notoriously challenging for semantic interoperability. The hypothesis that representing the meaning of natural language in a machine-readable format could be framed as an automatic translation challenge needed multiple steps to be confirmed.

The first observation drawn from the scoping review was that if SNOMED CT had been widely used to process clinical free text, it was, in most cases, without using its advanced features. Indeed, it seems that post-coordinating concepts to better represent free text was seldom if ever attempted and never framing it as a translation task. This first step confirmed that the approach was innovative.

Exploratory experiments such as extracting SNOMED CT concepts from social media content, translating SNOMED CT concepts in French and German and manually translating French medical text into post-coordinated SNOMED CT sentences were essential steps to build knowledge of SNOMED CT and the translation task. This work gave sufficient insights to start the adaptation of an existing syntactico-semantic parser in the aim of creating an French-SNOMED CT automatic translator. The parser being largely rule-based with elements of statistical approach, it was in accordance with the results of the review. While the evaluation of post-coordination with conventional f-score metrics was not performed, the results in Conference Article 1 and 2 were promising with an f-score of 0.61 on the annotation of single concepts and a precision of 0.71 on the post-coordination with three defined SNOMED CT relationships. However, the complexity of converting SNOMED CT description model rules into French syntactic rules and implementing them in a large legacy codebase added to the fact that this tool was not finally applied on large corpora. Finally, the fact that the tool's code is not open source hinders necessary modifications and possible future updates from the end users. Therefore, a more sustainable in the long term tool is being considered.

This third part proved that approaching the problem of narrative clinical data as an automatic translation task into the interlingua could yield encouraging results. It highlighted key issues in the structure and semantic content of SNOMED CT and gave insights on how to resolve them. This work will be instrumental to multiple future projects at HUG and beyond. First, on the automatic translation task, the knowledge gathered during the parser adaptation will be extracted and transferred to a new NLP system to create a more robust solution for automatic SNOMED CT translation. Secondly, the experience gathered translating SNOMED CT in French and French into SNOMED CT is directly used in the frame of the official French translation group of SNOMED International of which the author of this thesis is an active member. This group recently released a new version of the common French translation as well as validated guidelines (238).

This work confirmed the last hypothesis and lays the ground for new progress in the semantic representation of clinical narratives.

8.4. General conclusions

The work presented in this thesis is the result of six years of work in a stimulating environment at the crossroads between healthcare, computer sciences and semantics. This uniquely interdisciplinary setting allowed the development and validation of a multi-dimensional approach to bring semantic interoperability to clinical data.

This approach emphasizes the challenges of bringing interoperability in a domain which covers multiple communities, types of data, standards and information systems. It confirms that no single solution exists and that targeted, semantically-centered approaches are necessary. From large national frameworks to very specific documents written in any natural language, interoperability must penetrate every layer of healthcare. The proposed solution is based firstly on strong semantics, by using compositional controlled vocabularies to create a computer readable interlingua without enforcing a data model, it then restricts the representation complexity to a useful set of concepts encountered in practice and finally exploits the compositional capabilities of the interlingua to represent complex narrative data by translating natural language in post-coordinated SNOMED CT sentences.

This approach defines a new, semantically interoperable landscape for clinical data that could leverage new opportunities proposed by the growth of personalized medicine and, in the long run, improve global interoperability.

9. Abbreviations

ADaM	Analysis Data Model
ADL	Archetype Definition Language
API	Application Programming Interface
BRIDG	Biomedical Research Integrated Domain Group
CDA	Clinical Document Architecture
CDISC	Clinical Data Interchange Standards Consortium
CDISC ODM	CDISC Operational Data Model
CEMs	Clinical Element Models
CHOP	Classification Suisse des interventions chirurgicales
CKM	Clinical Knowledge Manager
CRF	Case Report Form
CSI	Clinical Semantic Interoperability working group
CTV3	Clinical Terms Version 3
DCC	Data Coordination Center
DICOM	Digital Imaging and Communications in Medicine
DRG	Diagnoses Related Groups
EHR	Electronic Health Record
ETL	Extract, Transform and Load
FAIR	Findable, Accessible, Interoperable, Reusable
FHIR	Fast Healthcare Interoperability Resource
GATE	General Architecture for Text Engineering
GCS	Glasgow Coma Scale
HADES	Health Analytics Data-to-Evidence Suite
HIMSS	Healthcare Information and Management Systems Society
HITECH	Health Information Technology for Economic and Clinical Health
HITEx	Health Information Text Extraction
HL7	Health Level 7
HL7 v1	HL7 version 1
HL7 v2	HL7 version 2
HL7 v3	HL7 version 3
HTTP	Hypertext Transfer Protocol
HUG	University Hospital of Geneva
ICD	International Statistical Classification of Diseases and Related Health Problems
ICD-10	International Statistical Classification of Diseases and Related Health Problems tenth revision
ICD-10 AM	International Statistical Classification of Diseases and Related Health Problems tenth revision, Australian Modification
ICD-10 CM	International Statistical Classification of Diseases and Related Health Problems tenth revision, Clinical Modification
ICD-10 GM	International Statistical Classification of Diseases and Related Health Problems tenth revision, German Modification
ICD-11	International Statistical Classification of Diseases and Related Health Problems eleventh revision
ICD-9	International Statistical Classification of Diseases and Related Health Problems ninth revision
ICD-9 CM	International Statistical Classification of Diseases and Related Health Problems ninth revision, Clinical Modification
ICD-O-3	International Classification of Diseases for Oncology, 3rd edition

IEEE	Institute of Electrical and Electronics Engineers
IHE	Integrating the Healthcare Enterprise
ISO	International Organization for Standardization
JMIR	Journal of Medical Internet Research
LOINC	Logical Observation Identifiers, Names, and Codes
MedDRA	Medical Dictionary for Regulatory Activities
NCI	National Cancer Institute
NHS	National Health Service
NLP	Natural Language Processing
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OMOP CDM	OMOP Common Data Model
ONC	Office of the National Coordinator for Health Information Technology
OSI	Open Systems Interconnection
PACS	Picture Archiving and Communication Systems
PITAC	President's Information Technology Advisory Committee
RDF	Resource Description Framework
RIM	Reference Information Model
SDG	Sustainable Development Goals
SHARPn	Strategic Health IT Advanced Research Projects consortium
SIMED	Division of Medical information Sciences
SMART	Substitutable Medical Applications and Reusable Technologies
SNOMED	Systematized Nomenclature of Medicine
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SNOMED RT	Systematized Nomenclature of Medicine Reference Terminology
SNOP	Systematized Nomenclature of Pathology
SPHN	Swiss Personalized Health Network
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System
WHO	World Health Organization

10. Bibliography

1. Chan M, Kazatchkine M, Lob-Levyt J, Obaid T, Schweizer J, Sidibe M, Veneman A, Yamada T. Meeting the Demand for Results and Accountability: A Call for Action on Health Data from Eight Global Health Agencies. PLOS Medicine Public Library of Science; 2010 Jan 26;7(1):e1000223. [doi: 10.1371/journal.pmed.1000223]

2. Data Interoperability Collaborative [Internet]. [cited 2021 Mar 22]. Available from: https://www.data4sdgs.org/initiatives/data-interoperability-collaborative

3. Interoperability: A practitioner's guide to joining-up data in the development sector [Internet]. [cited 2021 Mar 22]. Available from: https://www.data4sdgs.org/resources/interoperability-practitioners-guide-joining-data-development-sector

4. Fehlmann C, Louis Simonet M, Reny J-L, Stirnemann J, Blondon K. Associations between early handoffs, length of stay and complications in internal medicine wards: A retrospective study. Eur J Intern Med 2019 Sep;67:77–83. PMID:31311699

5. Blondon K, Zotto MD, Rochat J, Nendaz MR, Lovis C. A Simulation Study on Handoffs and Cross-coverage: Results of an Error Analysis. AMIA Annu Symp Proc 2017;2017:448–457. PMID:29854109

6. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. IEEE Std 610 1991 Jan;1–217. [doi: 10.1109/IEEESTD.1991.106963]

7. Society (HIMSS) HI& MS. HIMSS Dictionary of Health Information and Technology Terms, Acronyms and Organizations. CRC Press; 2019. ISBN:978-1-351-10451-7

8. Braunstein ML. Healthcare in the Age of Interoperability: The Promise of Fast Healthcare Interoperability Resources. IEEE Pulse 2018 Dec;9(6):24–27. PMID:30452344

9. Benson T, Grieve G. Principles of health interoperability: SNOMED CT, HL7 and FHIR. Springer; 2016.

10. Blobel B. Introduction into advanced eHealth -- the Personal Health challenge. Stud Health Technol Inform 2008;134:3–14. PMID:18376028

11. Benson T, Grieve G. CDA – Clinical Document Architecture. In: Benson T, Grieve G, editors. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR Cham: Springer International Publishing; 2016. p. 283–301. [doi: 10.1007/978-3-319-30370-3_12]

12. Aghdam ZN, Rahmani AM, Hosseinzadeh M. The Role of the Internet of Things in Healthcare: Future Trends and Challenges. Comput Methods Programs Biomed 2021 Feb;199:105903. PMID:33348073

13. HTTP - Hypertext Transfer Protocol Overview [Internet]. [cited 2020 Dec 15]. Available from: https://www.w3.org/Protocols/

14. Samwald M, Fehre K, de Bruin J, Adlassnig K-P. The Arden Syntax standard for clinical decision support: experiences and directions. J Biomed Inform 2012 Aug;45(4):711–718. PMID:22342733

15. Atherton J. Development of the Electronic Health Record. AMA Journal of Ethics American Medical Association; 2011 Mar 1;13(3):186–189. [doi: 10.1001/virtualmentor.2011.13.3.mhst1-1103]

16. Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. J Med Syst 1983 Apr;7(2):87–102. PMID:6688267

17. McDonald CJ, Murray R, Jeris D, Bhargava B, Seeger J, Blevins L. A computer-based record and clinical monitoring system for ambulatory care. Am J Public Health 1977 Mar;67(3):240–245. PMID:842761

18. Evans RS. Electronic Health Records: Then, Now, and in the Future. Yearb Med Inform 2016 May 20;Suppl 1:S48-61. PMID:27199197

19. Revolutionizing Healthcare Through Information Technology - Data.gov [Internet]. [cited 2020 Jan 6]. Available from: https://catalog.data.gov/dataset/revolutionizing-healthcare-through-information-technology

20. Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [Internet]. [cited 2021 Jan 4]. Available from: https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf

21. 2015 Update to Congress on the Adoption of Health Information Technology [Internet]. [cited 2021 Jan 4]. Available from: /report-to-congress/2015-update-adoption-health-information-technology.php

22. Colicchio TK, Cimino JJ, Del Fiol G. Unintended Consequences of Nationwide Electronic Health Record Adoption: Challenges and Opportunities in the Post-Meaningful Use Era. J Med Internet Res [Internet] 2019 Jun 3 [cited 2021 Jan 4];21(6). PMID:31162125

23. WHO | Electronic health records [Internet]. WHO. World Health Organization; [cited 2021 Jan 4]. Available from: http://www.who.int/gho/goe/electronic_health_records/en/

24. Everson J, Patel V, Adler-Milstein J. Information blocking remains prevalent at the start of 21st Century Cures Act: results from a survey of health information exchange organizations. Journal of the American Medical Informatics Association 2021 Apr 1;28(4):727–732. [doi: 10.1093/jamia/ocaa323]

25. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006;1040. PMID:17238659

26. Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA, Avitabile JE, Liu J, Housman D, Palchuk MB. tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. AMIA Jt Summits Transl Sci Proc 2014;2014:96–101. PMID:25717408

27. Raisaro JL, Troncoso-Pastoriza JR, Misbach M, Sousa JS, Pradervand S, Missiaglia E, Michielin O, Ford B, Hubaux J-P. MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. IEEE/ACM Trans Comput Biol Bioinform 2019 Aug;16(4):1328–1341. PMID:30010584

28. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak [Internet] 2020 Oct 29 [cited 2021 Jan 4];20. PMID:33121479

29. Mahajan SM, Ghani R. Combining Structured and Unstructured Data for Predicting Risk of Readmission for Heart Failure Patients. Stud Health Technol Inform 2019 Aug 21;264:238–242. PMID:31437921

30. Kong H-J. Managing Unstructured Big Data in Healthcare System. Healthc Inform Res 2019 Jan;25(1):1–2. PMID:30788175

31. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA 2014 Jun 25;311(24):2479–2480. PMID:24854141

32. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 1;26(4):364–379. PMID:30726935

33. Chen ES, Sarkar IN. Mining the Electronic Health Record for Disease Knowledge. In: Kumar VD, Tipney HJ, editors. Biomedical Literature Mining [Internet] New York, NY: Springer; 2014 [cited 2021 Jan 4]. p. 269–286. PMID:24788272

34. Liu X, Wang Y, Ji J, Cheng H, Zhu X, Awa E, He P, Chen W, Poon H, Cao G, Gao J. The Microsoft Toolkit of Multi-Task Deep Neural Networks for Natural Language Understanding. arXiv:200207972 [cs] [Internet] 2020 May 15 [cited 2021 Jan 4]; Available from: http://arxiv.org/abs/2002.07972

35. Zhang Z, Wu Y, Zhao H, Li Z, Zhang S, Zhou X, Zhou X. Semantics-Aware BERT for Language Understanding. AAAI 2020 Apr 3;34(05):9628–9635. [doi: 10.1609/aaai.v34i05.6510]

36. Johnson SB, Bakken S, Dine D, Hyun S, Mendonça E, Morrison F, Bright T, Van Vleck T, Wrenn J, Stetson P. An electronic health record based on structured narrative. J Am Med Inform Assoc 2008 Feb;15(1):54–64. PMID:17947628

37. Ross MK, Wei W, Ohno-Machado L. “Big Data” and the Electronic Health Record. Yearb Med Inform 2014 Aug 15;9(1):97–104. PMID:25123728

38. Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the Degree of Concordance of Coded and Textual Data in Answering Clinical Queries from a Clinical Data Repository. J Am Med Inform Assoc 2000;7(1):42–54. PMID:10641962

39. Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inform Assoc 2010;17(6):671–674. PMID:20962129

40. Matheny ME, Fitzhenry F, Speroff T, Green JK, Griffith ML, Vasilevskis EE, Fielstein EM, Elkin PL, Brown SH. Detection of infectious symptoms from VA emergency department and primary care clinical documentation. Int J Med Inform 2012 Mar;81(3):143–156. PMID:22244191

41. Classification of Diseases (ICD) [Internet]. [cited 2021 Jan 4]. Available from: https://www.who.int/standards/classifications/classification-of-diseases

42. Jacobson BC, Gerson LB. The inaccuracy of ICD-9-CM Code 530.2 for identifying patients with Barrett’s esophagus. Dis Esophagus 2008;21(5):452–456. PMID:19125800

43. Hess LM, Zhu YE, Sugihara T, Fang Y, Collins N, Nicol S. Challenges of Using ICD-9-CM and ICD-10-CM Codes for Soft-Tissue Sarcoma in Databases for Health Services Research. Perspect Health Inf Manag [Internet] 2019 Apr 1 [cited 2021 Jan 4];16(Spring). PMID:31019431

44. Walsh P, Rothenberg SJ. Which ICD-9-CM codes should be used for bronchiolitis research? BMC Med Res Methodol [Internet] 2018 Nov 22 [cited 2021 Jan 4];18. PMID:30466396

45. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ, Saltz JH. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Med Care 2013 Aug;51(8 0 3):S30–S37. PMID:23774517

46. Dugas M, Thun S, Frankewitsch T, Heitmann KU. LOINC® Codes for Hospital Information Systems Documents: A Case Study. J Am Med Inform Assoc 2009;16(3):400–403. PMID:19261942

47. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017 Aug;26(1):38–52. PMID:28480475

48. GA4GH [Internet]. [cited 2021 Feb 15]. Available from: https://www.ga4gh.org/

49. The OBO Foundry [Internet]. [cited 2021 Feb 15]. Available from: http://www.obofoundry.org/

50. Gene Ontology Resource [Internet]. Gene Ontology Resource. [cited 2021 Feb 15]. Available from: http://geneontology.org/

51. ISO - International Organization for Standardization [Internet]. ISO. [cited 2021 Mar 23]. Available from: https://www.iso.org/home.html

52. ISO 13606 Standard - EHR Interoperability [Internet]. [cited 2020 Dec 22]. Available from: http://www.en13606.org/information.html

53. Family of International Classifications (FIC) [Internet]. [cited 2021 Mar 23]. Available from: https://www.who.int/standards/classifications/family-of-international-classifications-(fic)

54. Siegel EL, Channin DS. Integrating the Healthcare Enterprise: a primer. Part 1. Introduction. Radiographics 2001 Oct;21(5):1339–1341. PMID:11553841

55. Integrating the Healthcare Enterprise (IHE) [Internet]. IHE International. [cited 2020 Dec 29]. Available from: https://www.ihe.net/

56. IHE Process [Internet]. IHE International. [cited 2020 Dec 29]. Available from: https://www.ihe.net/about_ihe/ihe_process/

57. 14:00-17:00. ISO/IEC 7498-1:1994 [Internet]. ISO. [cited 2021 Mar 23]. Available from: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/02/02/20269.html

58. HL7 Standards Product Brief - HL7 Version 2 Product Suite | HL7 International [Internet]. [cited 2020 Dec 15]. Available from: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185

59. Benson T, Grieve G. The HL7 v3 RIM. In: Benson T, Grieve G, editors. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR Cham: Springer International Publishing; 2016. p. 243–264. [doi: 10.1007/978-3-319-30370-3_12]

60. PRM | CDISC [Internet]. [cited 2020 Dec 23]. Available from: https://www.cdisc.org/standards/foundational/protocol

61. SDTM | CDISC [Internet]. [cited 2020 Dec 23]. Available from: https://www.cdisc.org/standards/foundational/sdtm

62. ADaM | CDISC [Internet]. [cited 2020 Dec 23]. Available from: https://www.cdisc.org/standards/foundational/adam

63. CDISC Announces BRIDG Model for Research as Final ISO Standard | CDISC [Internet]. [cited 2021 Feb 18]. Available from: https://www.cdisc.org/cdisc-announces-bridg-model

64. Commissioner O of the. Study Data Standards Resources [Internet]. FDA. FDA; 2020 [cited 2020 Dec 23]. Available from: https://www.fda.gov/industry/fda-resources-data-standards/study-data-standards-resources

65. CDISC Terminology | CBIIT [Internet]. [cited 2020 Dec 23]. Available from: https://datascience.cancer.gov/resources/cancer-vocabulary/cdisc-terminology

66. Why Not Just Use SNOMED? | CDISC [Internet]. [cited 2020 Dec 23]. Available from: https://www.cdisc.org/kb/articles/why-not-just-use-snomed

67. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010 Nov 2;153(9):600–606. PMID:21041580

68. Data Standardization – OHDSI [Internet]. [cited 2020 Dec 19]. Available from: https://www.ohdsi.org/data-standardization/

69. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, Morris JA. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc 2010 Dec;17(6):652–662. PMID:20962127

70. Informatics OHDS and. Chapter 1 The OHDSI Community | The Book of OHDSI [Internet]. [cited 2020 Dec 19]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/

71. Benson T, Grieve G. Implementing Terminologies. In: Benson T, Grieve G, editors. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR Cham: Springer International Publishing; 2016. p. 189–221. [doi: 10.1007/978-3-319-30370-3_12]

72. HL7 Standards Product Brief - HL7 Messaging Standard Version 2.9 | HL7 International [Internet]. [cited 2020 Dec 15]. Available from: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=516

73. Beeler GW. HL7 version 3--an object-oriented methodology for collaborative standards development. Int J Med Inform 1998 Feb;48(1–3):151–161. PMID:9600415

74. Reference Information Model (RIM) Downloads | HL7 International [Internet]. [cited 2021 Mar 23]. Available from: http://www.hl7.org/implement/standards/rim.cfm

75. HL7 Standards Product Brief - CDA® R2.1 (HL7 Clinical Document Architecture, Release 2.1) | HL7 International [Internet]. [cited 2020 Dec 16]. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=515

76. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo Shvo A. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc 2006 Feb;13(1):30–39. PMID:16221939

77. Interoperability Toolkit releases [Internet]. NHS Digital. [cited 2020 Dec 16]. Available from: https://digital.nhs.uk/services/interoperability-toolkit/interoperability-toolkit-releases

78. Burke T. Common - Clinical Documents [Internet]. Australian Digital Health Agency Developer Centre; 2018 [cited 2020 Dec 16]. Available from: https://developer.digitalhealth.gov.au/products/common-clinical-documents

79. HL7 Standards Product Brief - HL7/ASTM Implementation Guide for CDA® R2 -Continuity of Care Document (CCD®) Release 1 | HL7 International [Internet]. [cited 2020 Dec 16]. Available from: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=6

80. Ferranti JM, Musser RC, Kawamoto K, Hammond WE. The Clinical Document Architecture and the Continuity of Care Record: A Critical Analysis. J Am Med Inform Assoc 2006;13(3):245–252. PMID:16501180

81. Shabo (Shvo) A. Clinical Document Architecture. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems [Internet] New York, NY: Springer; 2018 [cited 2020 Dec 16]. p. 452–454. [doi: 10.1007/978-1-4614-8265-9_59]

82. Benson T, Grieve G. Principles of FHIR. In: Benson T, Grieve G, editors. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR Cham: Springer International Publishing; 2016. p. 329–348. [doi: 10.1007/978-3-319-30370-3_12]

83. Smith B, Ceusters W. HL7 RIM: an incoherent standard. Stud Health Technol Inform 2006;124:133–138. PMID:17108516

84. Resources For Health: A Fresh Look Proposal [Internet]. Health Intersections Pty Ltd. 2011 [cited 2020 Dec 16]. Available from: http://www.healthintersections.com.au/?p=502

85. Overview - FHIR v4.0.1 [Internet]. [cited 2020 Dec 16]. Available from: https://www.hl7.org/fhir/overview.html

86. Connectathons - FHIR - Confluence [Internet]. [cited 2020 Dec 16]. Available from: https://confluence.hl7.org/display/FHIR/Connectathons

87. Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, Mandl KD, Murphy SN. SMART-on-FHIR implemented over i2b2. J Am Med Inform Assoc 2017 Mar 1;24(2):398–402. PMID:27274012

88. DICOM [Internet]. DICOM. [cited 2021 Mar 25]. Available from: https://www.dicomstandard.org

89. Pianykh OS. Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide [Internet]. 2nd ed. Berlin Heidelberg: Springer-Verlag; 2012 [cited 2020 Dec 29]. [doi: 10.1007/978-3-642-10850-1]ISBN:978-3-642-10849-5

90. Software Tools – OHDSI [Internet]. [cited 2020 Dec 19]. Available from: https://www.ohdsi.org/software-tools/

91. HADES [Internet]. [cited 2020 Dec 19]. Available from: https://ohdsi.github.io/Hades/

92. openEHR [Internet]. [cited 2020 Dec 20]. Available from: https://www.openehr.org/

93. Thomas Beale. Archetypes: Constraint-based domain models for future-proof information systems. Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer Seattle, Washington, USA: Northeastern University; 2002. p. 16–32.

94. Clinical Knowledge Manager [Internet]. [cited 2020 Dec 22]. Available from: https://ckm.openehr.org/ckm/

95. 14:00-17:00. ISO 13606-1:2019 [Internet]. ISO. [cited 2020 Dec 22]. Available from: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/78/67868.html

96. Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). J Biomed Inform 2015 Oct;57:88–99. PMID:26188274

97. We don’t need no Semantic Interoperability [Internet]. Health Intersections Pty Ltd. 2011 [cited 2021 Jan 12]. Available from: http://www.healthintersections.com.au/?p=155

98. Mougin F, Bodenreider O, Burgun A. Analyzing polysemous concepts from a clinical perspective: Application to auditing concept categorization in the UMLS. J Biomed Inform 2009 Jun;42(3):440–451. PMID:19303057

99. SNOMED CT Starter Guide - SNOMED CT Starter Guide [Internet]. [cited 2021 Jan 7]. Available from: https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide

100. El-Sappagh S, Elmogy M. An encoding methodology for medical knowledge using SNOMED CT ontology. Journal of King Saud University - Computer and Information Sciences 2016 Jul 1;28(3):311–329. [doi: 10.1016/j.jksuci.2015.10.002]

101. thesaurus [Internet]. [cited 2021 Apr 6]. Available from: https://dictionary.cambridge.org/fr/dictionnaire/anglais/thesaurus

102. The Unified Medical Language System (UMLS) [Internet]. U.S. National Library of Medicine; [cited 2021 Jan 19]. Available from: https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html

103. CLASSIFICATION | definition in the Cambridge English Dictionary [Internet]. [cited 2021 Jan 20]. Available from: https://dictionary.cambridge.org/us/dictionary/english/classification

104. Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: “Ontologies” in today’s biomedical information systems and the limits of OWL. Journal of Biomedical Informatics: X 2019 Jun 1;2:100002. [doi: 10.1016/j.yjbinx.2019.100002]

105. Cimiano P, editor. Ontologies. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications [Internet] Boston, MA: Springer US; 2006 [cited 2021 Jan 19]. p. 9–17. [doi: 10.1007/978-0-387-39252-3_2]

106. Gruber TR. Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies 1995 Nov 1;43(5):907–928. [doi: 10.1006/ijhc.1995.1081]

107. NANDA International Nursing Diagnoses | NANDA International, Inc [Internet]. 2020 [cited 2021 Apr 6]. Available from: https://nanda.org/publications-resources/publications/nanda-international-nursing-diagnoses/

108. International Classification of Primary Care, 2nd edition (ICPC-2) [Internet]. [cited 2021 Jan 7]. Available from: https://www.who.int/standards/classifications/other-classifications/international-classification-of-primary-care

109. International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) [Internet]. [cited 2021 Apr 6]. Available from: https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology

110. DSM-5 [Internet]. [cited 2021 Apr 6]. Available from: https://www.psychiatry.org/psychiatrists/practice/dsm

111. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 1974 Jul 13;2(7872):81–84. PMID:4136544

112. American Academy of Pediatrics, Committee on Fetus and Newborn, American College of Obstetricians and Gynecologists and Committee on Obstetric Practice. The Apgar score. Pediatrics 2006 Apr;117(4):1444–1447. PMID:16585348

113. Elovic A, Pourmand A. MDCalc Medical Calculator App Review. J Digit Imaging 2019 Oct;32(5):682–684. PMID:31025219

114. Adams NE. Bloom’s taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152–153. PMID:26213509

115. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res 2018 Jan 4;46(D1):D708–D717. PMID:29040670

116. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018 Nov;36(10):996–1004. PMID:30148503

117. TERMINOLOGY | definition in the Cambridge English Dictionary [Internet]. [cited 2021 Jan 20]. Available from: https://dictionary.cambridge.org/us/dictionary/english/terminology

118. CISMeF. HeTOP [Internet]. Centre Hospitalo-Universitaire de Rouen; [cited 2021 Jan 12]. Available from: https://www.hetop.eu/hetop/

119. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, Darmoni SJ. Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform 2011;166:129–138. PMID:21685618

120. Terminology Standards | HIMSS [Internet]. [cited 2021 Jan 31]. Available from: https://www.himss.org/terminology-standards

121. Datatypes - FHIR v3.0.2 [Internet]. [cited 2021 Jan 26]. Available from: http://hl7.org/fhir/stu3/datatypes.html#CodeableConcept

122. Informatics OHDS and. Chapter 5 Standardized Vocabularies | The Book of OHDSI [Internet]. [cited 2021 Jan 26]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/

123. Wolandscat. A FHIR experience: consistently inconsistent [Internet]. Woland's cat. 2019 [cited 2021 Jan 26]. Available from: https://wolandscat.net/2019/05/05/a-fhir-experience-consistently-inconsistent/

124. Wolandscat. FHIR v openEHR – concreta [Internet]. Woland's cat. 2018 [cited 2021 Mar 26]. Available from: https://wolandscat.net/2018/10/10/fhir-v-openehr-concreta/

125. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998 Nov;37(4–5):394–403. PMID:9865037

126. Benson T, Grieve G. Clinical Terminology. In: Benson T, Grieve G, editors. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR Cham: Springer International Publishing; 2016. p. 121–133. [doi: 10.1007/978-3-319-30370-3_12]

127. Medical Vocabularies - J. Cimino - J. Cimino [Internet]. [cited 2021 Jan 20]. Available from: https://confluence.ihtsdotools.org/display/DOCEG/Medical+Vocabularies+-+J.+Cimino

128. Jetté N, Quan H, Hemmelgarn B, Drosler S, Maass C, Moskal L, Paoin W, Sundararajan V, Gao S, Jakob R, Ustün B, Ghali WA, IMECCHI Investigators. The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. Med Care 2010 Dec;48(12):1105–1110. PMID:20978452

129. FAQ on purchasing, permissions, copyright, licences and translations [Internet]. [cited 2021 Jan 13]. Available from: https://www.who.int/standards/classifications/frequently-asked-questions

130. WHO | ICD-11 Timeline [Internet]. WHO. World Health Organization; [cited 2021 Jan 14]. Available from: http://www.who.int/classifications/icd/revision/timeline/en/

131. WHO releases new International Classification of Diseases (ICD 11) [Internet]. [cited 2021 Jan 14]. Available from: https://www.who.int/news/item/18-06-2018-who-releases-new-international-classification-of-diseases-(icd-11)

132. The Lancet null. ICD-11. Lancet 2019 Jun 8;393(10188):2275. PMID:31180012

133. Adaptation steps [Internet]. [cited 2020 Nov 25]. Available from: https://www.dimdi.de/dynamic/en/classifications/icd/icd-10-gm/history/adaptation-steps/

134. IHPA. ICD-10-AM/ACHI/ACS [Internet]. IHPA; 2019 [cited 2021 Jan 12]. Available from: https://www.ihpa.gov.au/what-we-do/icd-10-am-achi-acs-classification

135. ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification [Internet]. 2020 [cited 2021 Jan 12]. Available from: https://www.cdc.gov/nchs/icd/icd10cm.htm

136. Cartwright DJ. ICD-9-CM to ICD-10-CM Codes: What? Why? How? Adv Wound Care (New Rochelle) 2013 Dec;2(10):588–592. PMID:24761333

137. Office of the Secretary, HHS. HIPAA administrative simplification: modification to medical data code set standards to adopt ICD-10-CM and ICD-10-PCS. Proposed rule. Fed Regist 2008 Aug 22;73(164):49795–49832. PMID:18958951

138. statistique O fédéral de la. Manuel de codage médical. Le manuel officiel des règles de codage en Suisse - Version 2020 | Publication [Internet]. Office fédéral de la statistique. 2019 [cited 2020 Nov 25]. Available from: /content/bfs/fr/home/statistiken/gesundheit/nomenklaturen/medkk/instrumente-medizinische-kodierung.assetdetail.9927930.html

139. Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies — SNOMED CT, LOINC, and RxNorm. Yearb Med Inform 2018 Aug;27(1):129–139. PMID:30157516

140. LOINC Table, Reports, and Users' Guide [Internet]. LOINC. [cited 2021 Jan 19]. Available from: https://loinc.org/downloads/loinc-table/

141. 2019 Annual Report [Internet]. LOINC. [cited 2021 Jan 19]. Available from: https://loinc.org/annual-reports/year-2019/

142. Drenkhahn C, Ingenerf J. The LOINC Content Model and Its Limitations of Usage in the Laboratory Domain. Stud Health Technol Inform Netherlands; 2020 Jun 16;270:437–442. PMID:32570422

143. Fiebeck J, Gietzelt M, Ballout S, Christmann M, Fradziak M, Laser H, Ruppel J, Schönfeld N, Teppner S, Gerbel S. Implementing LOINC - Current Status and Ongoing Work at a Medical University. Stud Health Technol Inform 2019 Sep 3;267:59–65. PMID:31483255

144. Hauser RG, Quine DB, Ryder A, Campbell S. Unit conversions between LOINC codes. J Am Med Inform Assoc 2017 Jun 19;25(2):192–196. PMID:28637208

145. Bhattacharyya SB. SNOMED CT History and IHTSDO. In: Bhattacharyya S, editor. Introduction to SNOMED CT [Internet] Singapore: Springer; 2016 [cited 2020 Feb 24]. p. 19–23. [doi: 10.1007/978-981-287-895-3_3]

146. Campbell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. Proc Annu Symp Comput Appl Med Care 1992;354–358. PMID:1482897

147. Wang AY, Barrett JW, Bentley T, Markwell D, Price C, Spackman KA, Stearns MQ. Mapping between SNOMED RT and Clinical terms version 3: a key component of the SNOMED CT development process. Proc AMIA Symp 2001;741–745. PMID:11825284

148. Bentley T, Price C, Brown P. Structural and lexical features of successive versions of the Read Codes. Proceedings of the Annual Conference of the Primary Health Care Specialist Group, Worcester 1996. p. 91–103.

149. SNOMED Home page [Internet]. SNOMED. [cited 2021 Jan 7]. Available from: https://www.snomed.org/

150. SNOMED - Members [Internet]. SNOMED. [cited 2020 Oct 27]. Available from: https://www.snomed.org/our-customers/members

151. SNOMED CT concept model - SNOMED CT Glossary - SNOMED Confluence [Internet]. [cited 2019 Jun 12]. Available from: https://confluence.ihtsdotools.org/display/DOCGLOSS/SNOMED+CT+concept+model

152. Rodrigues JM, Schulz S, Mizen B, Trombert B, Rector A. Scrutinizing SNOMED CT's Ability to Reconcile Clinical Language Ambiguities with an Ontology Representation. Stud Health Technol Inform 2018;247:910–914. PMID:29678093

153. Cornet R, Schulz S. Relationship groups in SNOMED CT. Stud Health Technol Inform 2009;150:223–227. PMID:19745301

154. Rodrigues JM, Schulz S, Mizen B, Rector A, Serir S. Is the Application of SNOMED CT Concept Model sufficiently Quality Assured? AMIA Annu Symp Proc 2017;2017:1488–1497. PMID:29854218

155. SNOMED International SNOMED CT Browser [Internet]. [cited 2021 Jan 6]. Available from: https://browser.ihtsdotools.org/?

156. Root and Top-level Concepts - SNOMED CT Editorial Guide - SNOMED Confluence [Internet]. [cited 2021 Mar 26]. Available from: https://confluence.ihtsdotools.org/display/DOCEG/Root+and+Top-level+Concepts

157. Weida R, Bowie J, McClure RC, Sperzel D. Leveraging SNOMED CT with a General Purpose Terminology Server. KR-MED Citeseer; 2008.

158. SNOMED CT Compositional Grammar Specification and Guide [Internet]. [cited 2017 Nov 3]. Available from: https://www.snomed.org/news-articles/snomed-ct-compositional-grammar-specification-and-guide

159. A universal healthcare language on the way. BDJ Team Nature Publishing Group; 2018 Nov 2;5(1):1–1. [doi: 10.1038/bdjteam.2018.184]

160. Héja G, Surján G, Varga P. Ontological analysis of SNOMED CT. BMC Med Inform Decis Mak 2008 Oct 27;8 Suppl 1:S8. PMID:19007445

161. Patrick J, Wang Y, Budd P, Rector A, Brandt S, Rogers B, Herkes R, Ryan A, Vazirnezhad B. Developing SNOMED CT subsets from clinical notes for intensive care service. Health Care & Informatics Review Online Open Access 2008;

162. Wagholikar A.S., Lawley M.J., Hansen D.P., Chu K. Identifying symptom groups from Emergency Department presenting complaint free text using SNOMED CT. AMIA Annu Symp Proc 2011;2011((Wagholikar A.S.) The Australian e-Health Research Centre, Brisbane, Queensland, Australia.):1446–1453.

163. Lindberg D a. B, Humphreys BL, McCray AT. The Unified Medical Language System. Yearb Med Inform Georg Thieme Verlag KG; 1993;02(1):41–51. [doi: 10.1055/s-0038-1637976]

164. Murphy SN, Mendis M, Hackett K, Rajesh Kuttan, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, Chueh HC. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 2007 Oct 11;548–552. PMID:18693896

165. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010 Apr;17(2):124–130. PMID:20190053

166. Chapter 13. Identity Management (IM) Cell Install - Developers Getting Started With i2b2 - i2b2 Community Wiki [Internet]. [cited 2020 Dec 27]. Available from: https://community.i2b2.org/wiki/display/getstarted/Chapter+13.+Identity+Management+%28IM%29+Cell+Install

167. Chapter 9. Ontology Management (ONT) Cell Install - Developers Getting Started With i2b2 - i2b2 Community Wiki [Internet]. [cited 2020 Dec 27]. Available from: https://community.i2b2.org/wiki/display/getstarted/Chapter+9.+Ontology+Management+%28ONT%29+Cell+Install

168. i2b2 FHIR Cell - i2b2 FHIR Cell - i2b2 Community Wiki [Internet]. [cited 2020 Dec 27]. Available from: https://community.i2b2.org/wiki/display/FCC

169. Cunningham H, Humphreys K, Gaizauskas R, Wilks Y. GATE–a TIPSTER-based general architecture for text engineering. Proceedings of the TIPSTER Text Program (Phase III) 1997.

170. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6:30. PMID:16872495

171. Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. AMIA Annu Symp Proc 2009 Nov 14;2009:442–446. PMID:20351896

172. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008 Feb;15(1):14–24. PMID:17947624

173. Apache Solr - [Internet]. [cited 2020 Dec 27]. Available from: https://lucene.apache.org/solr/

174. Potenzone R. The Software of the Foundation [Internet]. i2b2 tranSMART Foundation. 2019 [cited 2020 Dec 27]. Available from: https://i2b2transmart.org/platform-solutions-and-roadmap/

175. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform 2012 Aug;45(4):763–771. PMID:22326800

176. Oniki TA, Zhuo N, Beebe CE, Liu H, Coyle JF, Parker CG, Solbrig HR, Marchant K, Kaggal VC, Chute CG, Huff SM. Clinical element models in the SHARPn consortium. J Am Med Inform Assoc 2016 Mar;23(2):248–256. PMID:26568604

177. Cohen KB, Demner-Fushman D. Biomedical Natural Language Processing [Internet]. John Benjamins; 2014 [cited 2021 Jan 20]. [doi: 10.1075/nlp.11]ISBN:978-90-272-7106-8

178. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. IMIA Yearbook of Medical Informatics Methods Inf Med 2008;47(1):128–44. PMID:18660887

179. Zuccon G, Wagholikar AS, Nguyen AN, Butt L, Chu K, Martin S, Greenslade J. 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. PMID:24303284

180. Xu J, Zhang Y, Wu Y, Wang J, Dong X, Xu H. Citation Sentiment Analysis in Clinical Trial Papers. AMIA Annu Symp Proc 2015;2015:1334–1341. PMID:26958274

181. Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. J Healthc Inf Manag 2008;22(3):52–56. PMID:19267032

182. Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 2003 Sep 1;19(13):1699–1706. PMID:12967967

183. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;17 Suppl 1:S74-82. PMID:11472995

184. Cohen KB, Xia J, Roeder C, Hunter LE. Reproducibility in Natural Language Processing: A Case Study of Two R Libraries for Mining PubMed/MEDLINE. LREC Int Conf Lang Resour Eval 2016 May;2016(W23):6–12. PMID:29568821

185. 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;18(5):540–543. PMID:21846785

186. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res 2019 May 31;21(5):e13484. PMID:31152528

187. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. PMID:27219127

188. Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, Kola J, Roberts I, Setzer A, Tapuria A, Wheeldin B. The CLEF corpus: semantic annotation of clinical text. AMIA Annu Symp Proc 2007 Oct 11;625–629. PMID:18693911

189. Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association 2007 Sep 1;14(5):550–563. [doi: 10.1197/jamia.M2444]

190. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association 2010 Sep 1;17(5):514–518. [doi: 10.1136/jamia.2010.003947]

191. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association 2013 Sep 1;20(5):806–813. [doi: 10.1136/amiajnl-2013-001628]

192. DBMI Portal [Internet]. [cited 2020 Dec 27]. Available from: https://portal.dbmi.hms.harvard.edu/

193. Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc 2020 Oct 1;27(10):1529–1537. PMID:32968800

194. Oemig F, Blobel B. Natural Language Processing Supporting Interoperability in Healthcare. In: Biemann C, Mehler A, editors. Text Mining: From Ontology Learning to Automated Text Processing Applications [Internet] Cham: Springer International Publishing; 2014 [cited 2021 Jan 26]. p. 137–156. [doi: 10.1007/978-3-319-12655-5_7]

195. Farr C. Epic's CEO is urging hospital customers to oppose rules that would make it easier to share medical info [Internet]. CNBC. 2020 [cited 2021 Apr 6]. Available from: https://www.cnbc.com/2020/01/22/epic-ceo-sends-letter-urging-hospitals-to-oppose-hhs-data-sharing-rule.html

196. Hochman M. Electronic Health Records: a "Quadruple Win," a "Quadruple Failure," or Simply Time for a Reboot? J Gen Intern Med 2018 Apr;33(4):397–399. PMID:29404945

197. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLoS One [Internet] 2019 Feb 19 [cited 2020 Dec 19];14(2). PMID:30779778

198. SNOMED - Plan to further enhance SNOMED CT's capabilities using Concrete Domains [Internet]. SNOMED. [cited 2020 Dec 14]. Available from: https://www.snomed.org/news-and-events/articles/planned-transition-concrete-domains

199. Chomsky N. Syntactic structures. Oxford, England: Mouton; 1957. p. 116.

200. Khorrami F, Ahmadi M, Sheikhtaheri A. Evaluation of SNOMED CT Content Coverage: A Systematic Literature Review. Stud Health Technol Inform 2018;248:212–219. PMID:29726439

201. Campbell JR, Xu J, Fung KW. Can SNOMED CT fulfill the vision of a compositional terminology? Analyzing the use case for Problem List. AMIA Annu Symp Proc 2011;2011:181–188. PMID:22195069

202. Assess CT - Home [Internet]. [cited 2021 Mar 6]. Available from: https://assess-ct.eu/home/

203. Final Brochure [Internet]. [cited 2021 Mar 6]. Available from: https://assess-ct.eu/final-brochure/

204. Resource Description Framework (RDF): Concepts and Abstract Syntax [Internet]. [cited 2020 Dec 14]. Available from: https://www.w3.org/TR/rdf-concepts/

205. Category:Triple Store - Semantic Web Standards [Internet]. [cited 2021 Mar 6]. Available from: https://www.w3.org/2001/sw/wiki/Category:Triple_Store

206. Benson T, Grieve G. Why Interoperability Is Hard. In: Benson T, Grieve G, editors. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR Cham: Springer International Publishing; 2016. p. 19–35. [doi: 10.1007/978-3-319-30370-3_12]

207. 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. PMID:23828173

208. Hwang EJ, Park H-A, Sohn SK, Lee HB, Choi HK, Ha S, Kim HJ, Kim TW, Youm W. Mapping Korean EDI Medical Procedure Code to SNOMED CT. Stud Health Technol Inform 2019 Aug 21;264:178–182. PMID:31437909

209. Block L, Handfield S. Mapping Wound Assessment Data Elements in SNOMED CT. Stud Health Technol Inform 2016;225:1078–1079. PMID:27332492

210. Kersloot MG, van Putten FJP, Abu-Hanna A, Cornet R, Arts DL. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semantics 2020 Nov 16;11(1):14. PMID:33198814

211. Miñarro-Giménez JA, Martínez-Costa C, Karlsson D, Schulz S, Gøeg KR. Qualitative analysis of manual annotations of clinical text with SNOMED CT. PLoS One [Internet] 2018 Dec 27 [cited 2021 Mar 8];13(12). PMID:30589855

212. First review report of the International Advisory Board [Internet]. SPHN. 2019 [cited 2021 Jan 20]. Available from: https://sphn.ch/2019/12/20/iab-report/

213. Swiss Personalized Health Network. Report from the National Steering Board 2016 – 2019 [Internet]. Zenodo; 2021 Jan. Report No.: 16. [doi: 10.5281/ZENODO.4044123]

214. Infrastructure Development Projects [Internet]. SPHN. [cited 2021 Jan 27]. Available from: https://sphn.ch/network/projects/infrastructure-development-projects/

215. Driver Projects [Internet]. SPHN. [cited 2021 Jan 27]. Available from: https://sphn.ch/network/projects/driver-projects/

216. Clinical Data Semantics Interoperability Working Group Strategy [Internet]. SPHN. [cited 2021 Mar 9]. Available from: https://sphn.ch/document/csi_wg_strategy/

217. SPHN Dataset Release [Internet]. SPHN. [cited 2021 Mar 9]. Available from: https://sphn.ch/document/sphn-dataset/

218. Khenglawt V, Lalṭanpuia. Machine translation and its approaches. Atlantis Press; 2018 [cited 2021 Mar 12]. p. 141–145. [doi: 10.2991/msc-18.2018.22]

219. Vauquois B. Structures profondes et traduction automatique: le système du CETA. Editura Academiei Republicii Socialiste România; 1968;

220. Rethinking linguistic relativity. New York, NY, US: Cambridge University Press; 1996. p. viii, 488. ISBN:0-521-44433-0

221. Wolff P, Holmes KJ. Linguistic relativity. WIREs Cognitive Science 2011;2(3):253–265. [doi: https://doi.org/10.1002/wcs.104]

222. Gentner D, Goldin-Meadow S. Language in Mind: Advances in the Study of Language and Thought. MIT Press; 2003. ISBN:978-0-262-57163-0

223. Wittgenstein: Reality is shaped by the words we use [Internet]. Farnam Street. 2013 [cited 2021 Mar 10]. Available from: https://fs.blog/2013/01/reality-is-shaped-by-the-words-we-use/

224. 4. SNOMED CT Basics - SNOMED CT Starter Guide - SNOMED Confluence [Internet]. [cited 2021 Mar 10]. Available from: https://confluence.ihtsdotools.org/display/docstart/4.+snomed+ct+basics

225. primitive concept - SNOMED CT Glossary - SNOMED Confluence [Internet]. [cited 2021 Mar 10]. Available from: https://confluence.ihtsdotools.org/display/DOCGLOSS/primitive+concept

226. Bern eHealth S 3003. SNOMED CT - eHealth Suisse [Internet]. [cited 2021 Apr 29]. Available from: https://www.e-health-suisse.ch/fr/technique-semantique/interoperabilite-semantique/snomed-ct.html

227. brat rapid annotation tool [Internet]. [cited 2021 Mar 10]. Available from: https://brat.nlplab.org/index.html

228. Wehrli E. Fips, a “deep” linguistic multilingual parser. Proceedings of the Workshop on Deep Linguistic Processing - DeepLP ’07 2007;(June):120. [doi: 10.3115/1608912.1608931]

229. Wehrli E, Nerima L. The fips multilingual parser. Language Production, Cognition, and the Lexicon Springer; 2015. p. 473–490.

230. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, Mcewen SA. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. Research Synthesis Methods 2014;5(4):371–385. PMID:26052958

231. TALMED 2019 [Internet]. [cited 2021 Jan 30]. Available from: https://imia.limsi.fr/talmed2019/

232. Koller MT, Van Delden C, Müller NJ, Baumann P, Lovis C, Marti HP, Fehr T, Binet I, De Geest S, Bucher HC, Meylan P, Pascual M, Steiger J. Design and methodology of the Swiss Transplant Cohort Study (STCS): A comprehensive prospective nationwide long-term follow-up cohort. European Journal of Epidemiology 2013;28(4):347–355. PMID:23546766

233. Lovis C, Colaert D, Stroetmann VN. DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Technol Inform 2008;136:641–646. PMID:18487803

234. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, Karakoyun T, Ohmann C, Lastic P-Y, Ammour N, Kush R, Dupont D, Cuggia M, Daniel C, Thienpont G, Coorevits P. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015 Feb;53:162–173. PMID:25463966