Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220466

Deep SNOMED CT Enabled Large Clinical Database About COVID-19

Christophe GAUDET-BLAVIGNAC[a,b,1], Julien EHRSAM[a,b], Hugues TURBE[a,b], Daniel KESZTHELYI[a,b], Jamil ZAGHIR[a,b] and Christian LOVIS[a,b]
[a]*Division of Medical Information Sciences, University Hospitals of Geneva*
[b]*Department of Radiology and Medical Informatics, University of Geneva*

Abstract. In spring 2020, as the COVID-19 pandemic is in its first wave in Europe, the University hospitals of Geneva (HUG) is tasked to take care of all Covid inpatients of the Geneva canton. It is a crisis with very little tools to support decision-taking authorities, and very little is known about the Covid disease. The need to know more, and fast, highlighted numerous challenges in the whole data pipeline processes. This paper describes the decisions taken and processes developed to build a unified database to support several secondary usages of clinical data, including governance and research. HUG had to answer to 5 major waves of COVID-19 patients since the beginning of 2020. In this context, a database for COVID-19 related data has been created to support the governance of the hospital in their answer to this crisis. The principles about this database were a) a clearly defined cohort; b) a clearly defined dataset and c) a clearly defined semantics. This approach resulted in more than 28 000 variables encoded in SNOMED CT and 1 540 human readable labels. It covers more than 216 000 patients and 590 000 inpatient stays. This database is used daily since the beginning of the pandemic to feed the "Predict" dashboards of HUG and prediction reports as well as several research projects.

Keywords. COVID-19, Semantic Interoperability, SNOMED CT

1. Introduction

Since the beginning of 2020, the SARS-CoV2 pandemic is putting an important pressure on care systems throughout the world. In this context, the need for rapidly generating reliable evidence about a previously unknown disease is strong. As the volume and diversity of healthcare data is growing exponentially, the expectations on their use to reach this goal are high.

In Switzerland, five important epidemic waves hit the population since February 2020, the fifth being still ongoing in January 2022. [1] Those waves resulted in a dramatic increase of the demand for the healthcare system. In the canton of Geneva, the Geneva University Hospitals (HUG), a consortium of all public hospitals and multiple outpatient clinics, was designated to take care of COVID-19 inpatients. This resulted in an important shift in the activity of the hospital and required new tools to navigate and take decisions in these uncertain times. In the same period, an important rise in COVID-19 related research projects was observed, resulting in numerous demands for clinical data extraction and curation.

[1] Corresponding Author, Christophe Gaudet-Blavignac; E-mail: christophe.gaudet-blavignac@hcuge.ch

Large initiatives aimed toward healthcare data interoperability for research are already deployed in Switzerland, and international knowledge representation systems such as SNOMED CT are recommended to solve the semantic interoperability challenge of representing healthcare data. [2–4] However, this crisis highlighted several challenges limiting our ability to leverage this profusion of data to answer concrete questions related to public health measures or medical research. [5] Simple tasks such as counting cases, hospitalizations and deaths showed that if data exists, additional actions must be taken to make it technically and semantically actionable, such as the development of targeted ontologies. [6,7] Moreover, effective communication of this information is complex and must be adapted to efficiently disseminate evidence. [8]

The Division of Medical Information Sciences at HUG was mandated to create a database of COVID-19 related data. This database is used by the governance of HUG to support decision processes required by the pandemic. It is also used by several research projects and their respective IRB approvals. This work describes the creation of this database, and the approach taken to make clinical data related to COVID-19 actionable.

2. Methods

2.1. Design

The selected dataset was decided to be as inclusive as possible. Every patient tested for SARS-CoV2 regardless of the result or flagged in the electronic health record as COVID-19 positive or suspect of COVID-19 is included in the database. This approach is crucial to bring sustainability in a fast-evolving situation where knowledge and truth evolve every day. Indeed, the concept of positivity or suspicion of COVID-19 evolves with time, and each variant raises new challenges for detection and diagnostics. For example, the threshold for the positivity of a PCR is influenced by new detection methods such as salivary or nasal antigen testing and these add new categories of patients to be dealt with (positive to PCR but negative to antigen, etc.). Therefore, the most stable way of selecting cases to be included in the database is to include every patient tested for SARS-CoV2 as well as every suspicion or mention of COVID-19 in the Electronic Health Record (EHR).

For temporal coverage, it was decided that for each case included in the dataset, historical data available in HUG up to 2 years in the past would be extracted. This decision was based on the fact that, as new discoveries were made on the impact of SARSCoV2 infection, it became clear that patients were not equal in the severity of their disease. Therefore, as little was known about which preexisting condition, phenotypic or genotypic trait was relevant, it was decided that 2 years of full historical data would be extracted to be able to cluster cases and understand the underlying causes behind those variations.

The data model is derived from the source databases in HUG. It is structured around two key elements, the cases, which represent the patients, and the stays, whether inpatient or outpatient. The data is extracted from multiple sources depending on availability, the main one being the institutional data warehouse. When data is unavailable in the data warehouse, other sources are included, such as general consent information or structured radiologic reports. The database is updated hourly for a relevant subset of the data and every night for the complete dataset.

2.2. Semantic enrichment

Regarding the semantic interoperability of the database, a pragmatic approach was adopted. All data in the database would be encoded in an existing international standard, already encoded data would not be reencoded and SNOMED CT, when possible, would be preferred over other standards. This was because a clinical database is only actionable if the semantic of its content is clearly represented. The choice of SNOMED CT was made based on its coverage and the possibility to combine concepts.

To allow quick development of the database, only variables present in the dataset are encoded. Queries were created to generate the set of distinct clinical concepts present in the database. Those concepts are encoded by semantic experts using SNOMED CT. Each variable is linked to one or more SNOMED CT concepts according to the SNOMED CT Compositional Grammar. [9] For specific categories such as laboratory analysis or observations, French human readable labels are created.

Those semantic enrichments are integrated in the extract, transform and load process updating the database and updated regularly when new concepts appear in the data.

3. Results

3.1. Structure of the database

The development of the database began in March 2020. At the time of writing the database is still on and updating continuously. It now covers more than 200 000 patients and 590 000 inpatient stays. Its content is described in Table 1.

Table 1. Number of instances by category of data

Category	Number of instances
Patients	216 194
Laboratory results	35 069 657
Observations	161 785 476
Inpatient stays	590 610
Ambulatory consultations	479 973
Drug Prescriptions	3 278 824
Prescriptions other	26 178 856
Diagnosis codes	594 391
Procedure codes	566 476

The database contains 28 source tables containing raw data from each source. 29 materialized views contains the processed data, and the semantic encoding is stored in 10 mapping tables. Overall, the database's size is around 51 Gigabyte (Gb), the largest tables being observations (15Gb), laboratory analysis (8.5Gb) and medication orders (8.3Gb).

3.2. Variables and encoding

The encoding of the variables in SNOMED CT has been iteratively updated. It covers more than 28 000 variables and all the instances of laboratory analysis and observations.

Data already encoded in international standards such as ICD-10 [10] and ATC [11] have not been re-encoded due to the availability of mappings from those classifications to SNOMED CT. [12] Statistics about SNOMED CT encodings are listed in Table 2.

Table 2. Number of variables encoded in SNOMED CT

Variable	Number of SNOMED CT encodings
Laboratory analysis	4 826
Laboratory materials	1 089
Laboratory units	65
Observations	1 068
Formularies	125
Patient problems	13 680
Formularies' drop lists	4 223
Total	25 076

To simplify the navigation and understandability of its content, 1 540 human readable labels have been created for laboratory analysis and observations. Examples are shown in Table 3.

Table 3. Example of simplified labels for laboratory analysis and observations

Source data	French human readable label
SARS-CoV-2, ARN, PCR E-gene, ql (COBAS 6800)	Coronavirus - SARS-CoV-2 (COVID19), PCR
Ag-ADP 10µM	Agrégation plaquettaire
pv.upload.result.covid_19.positiveExternalDocumentIDs	Résultat test covid 19, externe positif
pv.ventilation.ohd.duration	Durée oxygénothérapie à haut débit

3.3. Database usage

Throughout the pandemic, this database has been used for multiple purposes. The main goal being to support the governance of the HUG in piloting the response to the pandemic. It is used daily to feed the "Predict" dashboards of HUG and to produce prediction reports.

Lastly, thanks to the inclusion of general consent information, the database, its simplified labels, and its semantic encoding have been used for data extraction for 5 research projects validated by the Institutional Review Board of HUG. Therefore, multiple types of users, from data scientists to caregivers doing research, can explore and understand the data.

4. Discussion

In this work, the creation of a database for COVID-19 related data in HUG is described. The three principles show important advantages for the creation of an actionable database for governance as well as for research.

By defining broad and simple selection criteria to define the cohort, covering 2 years of historical data, and encoding all data in international standards, this database can be used for multiple purposes. Its wide coverage both in term of included patients and selected variables helped to answer to the sustained flow of clinical questions raised by

the pandemic and to remain complete despite the variations in the definition of what a COVID-19 patient is.

Finally, the choice of SNOMED CT to encode every variable not already encoded in an international standard created a semantically interoperable representation of the information, usable to answer clinical questions. Moreover the ability of SNOMED CT to handle composition of codes through the Compositional Grammar was mandatory to correctly represent the 28 000 variables covered by the database. This semantic approach is compliant with the Swiss Personalized Health Network initiative that is currently deployed in Switzerland and is targeted at shaping the future of research.

5. Conclusions

Overall, this work showed that this three principles approach is pragmatic and can solve some of the challenges of health data interoperability. It succeeded in quickly filling the information needs of the hospital in a global crisis. A more systematic evaluation of this approach could further validate those results.

6. References

[1] COVID-19 Suisse | Coronavirus | Dashboard, (n.d.). https://www.covid19.admin.ch/fr/overview (accessed January 20, 2022).

[2] 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 Jun 24;9(6):e27591. doi: 10.2196/27591. PMID: 34185008; PMCID: PMC8277320.

[3] SNOMED, SNOMED International, (2006). https://www.snomed.org/news-articles/snomed-ct-compositional-grammar-specification-and-guide (accessed November 6, 2017).

[4] 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. doi: 10.2196/24594.

[5] 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. doi: 10.3389/fpubh.2020.583401. PMID: 33553088; PMCID: PMC7862946.

[6] Haber NA, Clarke-Deelder E, Feller A, Smith ER, Salomon JA, MacCormack-Gelles B, Stone EM, Bolster-Foucault C, Daw JR, Hatfield LA, Fry CE, Boyer CB, Ben-Michael E, Joyce CM, Linas BS, Schmid I, Au EH, Wieten SE, Jarrett B, Axfors C, Nguyen VT, Griffin BA, Bilinski A, Stuart EA. Problems with evidence assessment in COVID-19 health policy impact evaluation: a systematic review of study design and evidence strength. BMJ Open. 2022 Jan 11;12(1):e053820.

[7] de Lusignan S, Liyanage H, McGagh D, Jani BD, Bauwens J, Byford R, Evans D, Fahey T, Greenhalgh T, Jones N, Mair FS, Okusi C, Parimalanathan V, Pell JP, Sherlock J, Tamburis O, Tripathy M, Ferreira F, Williams J, Hobbs FDR. COVID-19 Surveillance in a Primary Care Sentinel Network: In-Pandemic Development of an Application Ontology. JMIR Public Health Surveill. 2020 Nov 17;6(4):e21434. doi: 10.2196/21434. PMID: 33112762; PMCID: PMC7674143.

[8] H. Turbé, V.G. Ruiz, M. Bjelogrlic, J. Rochat, and C. Lovis, Communicating on Multivariate and Geospatial Data supported by ergonomics criteria: COVID-19 case, in: 2020 Workshop on Visual Analytics in Healthcare (VAHC), 2020: pp. 4–16. doi:10.1109/VAHC53729.2020.00008.

[9] 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).

[10] Classification of Diseases (ICD), (n.d.). https://www.who.int/standards/classifications/classification-of-diseases (accessed January 4, 2021).

[11] WHOCC - Home, (n.d.). https://www.whocc.no/ (accessed December 14, 2020).

[12] SNOMED CT maps, *SNOMED*. (n.d.). https://www.snomed.org/snomed-ct/Use-SNOMED-CT/maps (accessed January 20, 2022).