Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240813

Harmonization of Data Across Cohorts Using Standard Terminologies

Ahjung BYUN[a], Sumi SUNG[b], Jiyeon YU[a], Eunsuk CHANG[c] and Hyeoun-Ae PARK[a,1]
[a]*Seoul National University, Seoul, Republic of Korea*
[b]*Chungbuk National University, Chungcheongbuk-do, Republic of Korea*
[c]*Republic of Korea Air Force, Republic of Korea*
ORCiD ID: Ahjung Byun https://orcid.org/0000-0001-5914-1527, Sumi Sung https://orcid.org/0000-0003-3897-4698, Jiyeon Yu https://orcid.org/0009-0002-5923-0903, Eunsuk Chang https://orcid.org/0000-0002-1350-3606, Hyeoun-Ae Park https://orcid.org/0000-0002-3770-4998

Abstract. Korean National Institute of Health initiated data harmonization across cohorts with the aim to ensure semantic interoperability of data and to create a common database of standardized data elements for future collaborative research. With this aim, we reviewed code books of cohorts and identified common data items and values which can be combined for data analyses. We then mapped data items and values to standard health terminologies such as SNOMED CT. Preliminary results of this ongoing data harmonization work will be presented.

Keywords. Cohort data, Data harmonization, Global standard health terminology, Bio big data, Interoperability

1. Introduction

There have been national and international initiatives to harmonize and standardize data across different population cohorts such as ORCHESTRA and CINECA [1,2]. Korean National Institute of Health (KNIH) manages 27 cohorts for health promotion and disease prevention of the Korean population. These cohorts, with different target populations developed over 20 years, collect different sets of data with different values. This hinders data reuse and collaborative research across cohorts. KNIH initiated data harmonization across cohorts to create a common database of standardized data elements for future collaborative research by mapping common data items and values to standard health terminologies. This initiative started with two cohorts which are most widely used for research in Korea - Korean Genome and Epidemiology (KoGES) and Cardiovascular and Metabolic Disease Etiology Research Center (CMERC). Data items and values of the two cohorts were extracted and mapped to SNOMED CT. This poster presents our findings related to the harmonization of cohort data using SNOMED CT.

[1] Corresponding Author: Hyeoun-Ae Park; E-mail: hapark@snu.ac.kr.

2. Methods

The process of mapping data items and values of the two cohorts to the standard health terminology was carried out using SNOMED CT mapping guidelines [3].

First, the purpose of the mapping is to create a common database of standardized data elements for future collaborative research studies. Source codes of the map are data items and values of the KoGES and CMERC cohorts. Target code of the map is SNOMED CT. Second, we reviewed the code books of the two cohorts and extracted common data items and values. Third, we translated extracted data items and values into English in order to use as search terms for SNOMED CT. We used English terms proposed by the Korean Medical Association and the Korean Medical Library Engine. Fourth, two experienced mappers participated in mapping data items and values to SNOMED CT. Fifth, the map was validated internally and externally. Internally, two mappers compared their mapping results and tried to reach consensus. Externally, when there was disagreement in the maps developed by the two mappers, a third terminology expert was invited to initiate the consensus process. Lastly, if there were data items and values that were not mapped to SNOMED CT, we added the unmapped concepts as new concepts to SNOMED CT Korean Extension.

3. Results

A total of 624 data items and 414 values were identified from the two cohorts. Nine data groups with 336 data items were found to be common between the two cohorts. Over 90% of data items and values were mapped to SNOMED CT. However, the majority of terms describing detailed values of data items in the lifestyle management domain could not be mapped. The data items and values not mapped to standard terminology were added as new concepts to SNOMED CT Korean Extension. Additionally, reference sets for the two cohorts have been developed with SNOMED CT concepts mapped to data items and values of the two cohorts.

4. Discussion and Conclusions

Harmonization of cohort data was attempted by mapping data items and values of the two cohorts to global health terminology, SNOMED CT. This work will be expanded to other cohorts managed by KNIH. Researchers will be able to conduct collaborative research using harmonized data across cohorts in the near future.

References

[1] Rinaldi E, Stellmach C, Rajkumar NM, Caroccia N, Dellacasa C, Giannella M, Guedes M, Mirandola M, Scipione G, Tacconelli E, Thun S. Harmonization and standardization of data for a pan-European cohort on SARS-CoV-2 pandemic. NPJ Digital Medicine. 2022 Jun 14;5(1):75.

[2] CINECA-Common Infrastructure for National Cohorts in Europe, Canada, and Africa [Internet]. The European Union; c2022 [cited 2024 May 24]. Available from: https://www.cineca-project.eu/

[3] Sung S, Park HA, Jung H, Kang H. A SNOMED CT Mapping Guideline for the Local Terms Used to Document Clinical Findings and Procedures in Electronic Medical Records in South Korea. JMIR Medical Informatics. 2023 Mar;11(1):e46128, doi: 10.2196/46127