CHIMERA: Automatic Concept Set Creation and Mapping to Standard OMOP Codes in ATLAS

Marcela Rivera[1], Shahithya Lalitha Prabakaran[2], Luisa Martínez-Sanchez[1], Satyajit Pande[3], Anna Ostropolets[4]

[1] Johnson and Johnson R&D Data Science and Digital Health, Barcelona, Spain

[2] Mu Sigma, Bengaluru, Karnataka, India

[3] Johnson and Johnson Data Science and Digital Health, Titusville, US

[4] Janssen R&D, Titusville, US

Acknowledgements: Alan Andryc[3], Chris Knoll[4]

Background

Developing concept sets for intricate cohorts poses a substantial challenge in RWE analyses. Currently, creation and mapping of concept sets to the OHDSI Standardized Vocabularies standard concepts in ATLAS is a labor-intensive process and the manual code entry might lead to errors. Programmatic concept set creation is faster but would preclude visual inspection or curation of the concept sets and is not user-friendly. To illustrate this, we took a published and validated cohort definition of Systemic Lupus Erythematosus (SLE) patients, based on standard and non-standard codes (ICD-9, ICD-10-CM, RxNorm, LOINC, CPT) (Walunas et al (2021)) and transformed the lists of codes into concept sets based on standard codes. The cohort definition included several concept sets of different domains such as diagnoses, laboratory values and measurements and more than 1000 codes needed to be manually mapped into standards. The process was complex, susceptible to errors and took an estimated 240h (Table 1 shows a subset of the concept sets). Therefore, scaling and standardizing concept set creation and mapping require automated solutions. In response, we introduce CHIMERA, a Streamlit-based tool that improves efficiency and reproducibility, empowering researchers to scale the cohort and phenotype development process more effectively.

Methods

The CHIMERA application leverages python code and R functions connecting to WebAPI[1]. CHIMERA features a user-friendly interface that allows users to upload lists of codes as input files (e.g., csv, Excel sheets), which the system processes to create and save concept sets in ATLAS. The automation pipeline includes steps for reading the input files, cleaning (handling null values and special characters, removing duplicates) and validating the input data, interpreting wildcards (*, %) to include ranges of source codes, mapping the source codes to their

[1] postConceptSetDefinition(), authorizeWebApi(), existsConceptSetName(), getConceptSetDefinitionsMetaData() and deleteConceptSetDefinition()

corresponding standards (based on Athena's concept and concept-relationship vocabulary tables available at https://athena.ohdsi.org/). CHIMERA allows users to create different types of concept sets, including sets with only standard codes, both standard and non-standard codes, or just the codes from the input file. CHIMERA automatically generates concept sets in ATLAS using WebAPI with appropriate naming conventions ("CHI_first2charactersOfATLASUserName_last2charactersOfATLASUserName_sheetname_keyword"), it checks for conflicts, as ATLAS doesn't allow multiple concept-sets with the same name. If a conflict is found, users are alerted and can either rename the sheet and reupload the file or deselect the conflicting sheet and proceed with others. The cleaned and mapped concept set table can also be downloaded as a CSV file. CHIMERA effectively manages the inclusion and exclusion of codes and their descendants. **Figure 1** describes the process flow of the application in detail.

The application is available in GitHub (https://github.com/mriver90/CHIMERA.git) and to internal users within the J&J organization (https://opendatasciencelab.jnj.com/modelproducts/6685406288c1222bccaafdcb). Upcoming features of CHIMERA would allow optional orphan codes recommendations for potential codes that might be missing, and end to end curation which include features for annotating inclusion or exclusion reasons for each code, modifying the automatic mappings and managing code descendants. Our goal is to scale the application across the OHDSI community, promoting standardization of concept code creation, mapping and curation and encouraging community contributions.

Results

In this software demo, we demonstrate how CHIMERA can be leveraged to automatically create, and map large concept sets in minutes, reducing the time of building concept sets by up to 95% compared to the manual process. For the SLE use case, the processing time was reduced from 240 hours to 45 minutes using CHIMERA (excluding the time required for manual curation of the resulting concept sets in both the manual and CHIMERA processes). A minimization of errors due to repetitive manual tasks can also be inferred, although this was not tested. The streamlined workflow provided by CHIMERA not only saves valuable time and resources but also enhances scalability, enabling the handling of larger and more complex (inclusion/exclusion and descendants rules, wildcards, or ranges of source codes, etc.) concept sets with ease. The resulting concept sets are directly saved in ATLAS, where the user can explore and modify if needed.

Table 1 showcases the time taken using CHIMERA to create a subset of 8 concept sets used in the published definition of Systemic Lupus Erythematosus (SLE) according to several criteria from the Systemic Lupus International Collaborating Clinics (SLICC) and from the American

College of Rheumatology (ACR)/European league against rheumatism (EULAR). The concept code lists in this use-case was referenced from the paper “Evaluation of structured data from electronic health records to identify clinical classification criteria attributes for systemic lupus erythematosus” by Walunas et al (2021).

Figure 2 illustrates the successful creation of the 8 concept sets in ATLAS using the CHIMERA application.

Conclusion

The CHIMERA application offers a robust solution to the challenges of manual concept set creation and mapping using ATLAS. By automating the process, CHIMERA significantly reduces the time and effort required while improving reproducibility and scalability. This innovation has a high potential to impact the OHDSI community by streamlining workflows and enabling more efficient cohort and phenotype development. Future developments will focus on further enhancing the application's features and expanding its applicability.

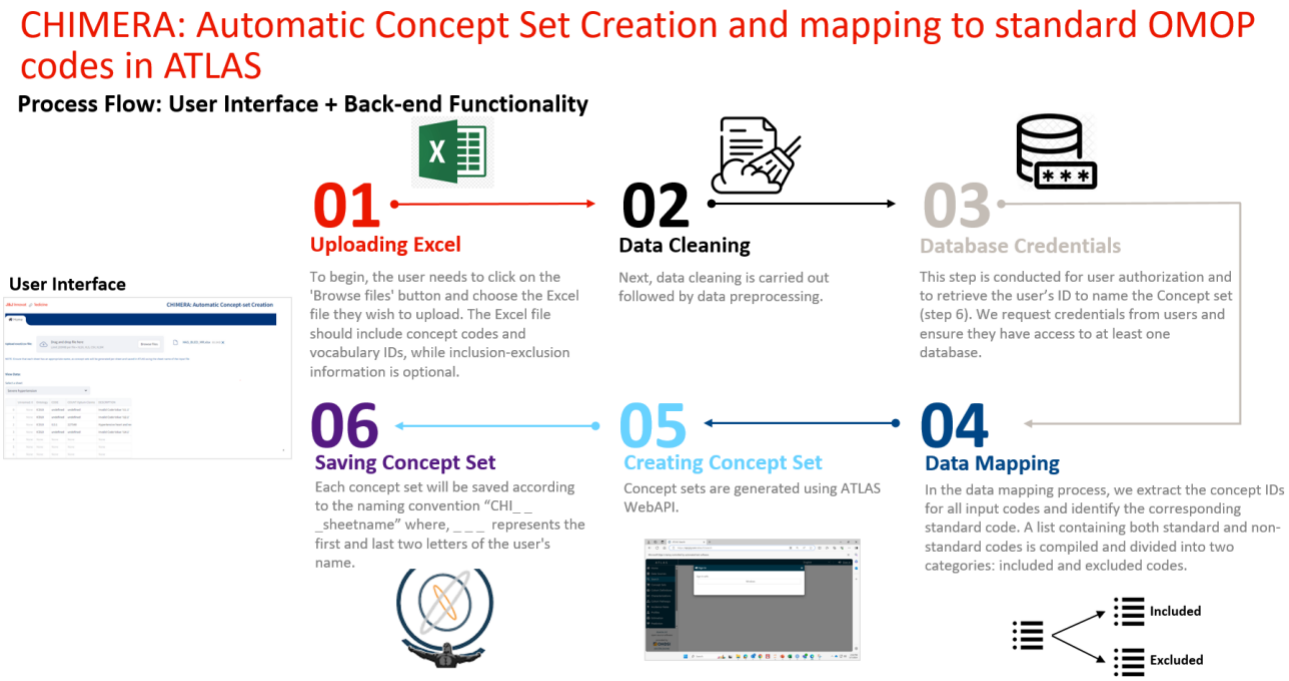

Figure 1. CHIMERA- Process Flow Diagram

Concept Set	No. of Source Codes	No. of Mapped Standard Codes	Source code + Standard code	Creation Time by CHIMERA
AcuteCutaneousLupus_SLICC_EULAR	33	17	50	4 minutes ***
Alopecia_SLICC_EULAR	16	7	23	
Anti-dsDNAAbs_SLICC.EULAR	10	10	10*	
Anti-NuclearAbs_ALL	10	10	10*	
Anti-PhospholipidAb_SLICC.EULAR	25	25	28**	
Anti-SmithAbs_SLICC.EULAR	13	13	13*	
Arthritis_ALL	159	60	219	
Neurological_Conditions_ACR	262	122	388**	
		Total	741	

NOTE:

*Source code given in the input file is already a standard code and doesn't have a mapping

**Certain source codes had a standard code mapping, while the rest were already standard codes

***Code list for all 8 concept sets in this table were loaded into CHIMERA, allowing for simultaneous creation of multiple concept sets

Table 1. Time efficiency of CHIMERA in the creation and mapping of a subset of 8 concept sets from a published definition of Systemic Lupus Erythematosus (SLE) by Walunas et al (2021). Systemic Lupus International Collaborating Clinics (SLICC), American College of Rheumatology (ACR), European league against rheumatism (EULAR)

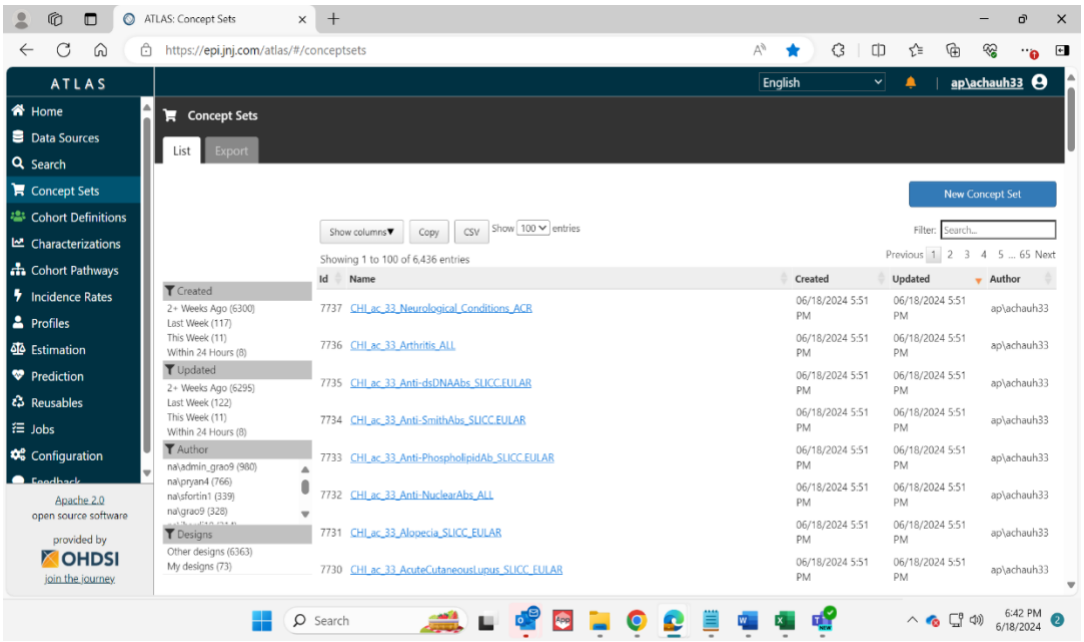

Figure 2. Concept Sets created using CHIMERA in ATLAS

References

1. postConceptSetDefinition() function of the ROhdsiWebApi package (HADES). Available from: Interacting With an OHDSI WebApi Instance • ROhdsiWebApi
2. Walunas TL, Ghosh AS, Pacheco JA, Mitrovic V, Wu A, Jackson KL, Schusler R, Chung A, Erickson D, Mancera-Cuevas K, Luo Y, Kho AN, Ramsey-Goldman R. Evaluation of

structured data from electronic health records to identify clinical classification criteria attributes for systemic lupus erythematosus. Lupus Sci Med. 2021 Apr;8(1):e000488. doi: 10.1136/lupus-2021-000488. PMID: 33903204; PMCID: PMC8076919.