

The Need for Composite Concept Sets

Sigfried Gold^{1,2}, Harold P Lehmann¹, Lisa Eskenazi¹, Xiaohan Tanner Zhang¹, Stephanie Hong¹, Richard L. Zhu¹, Christopher G. Chute¹, on behalf of the N3C Consortium

¹ Johns Hopkins University, Biomedical Informatics and Data Science section, ² University of Maryland, College of Information Studies

Background

OHDSI/ATLAS cohort definitions provide a framework for selecting patient cohorts based on concept sets and temporal and conditional logic. When the cohort represents patients with a single clinical condition or event, the cohort definition will require a single set of codes signifying that phenomenon, but some clinical phenomena cannot be represented by a single code set. For example, combination medications: although one could restrict the concept set to codes for the combination medication, some health systems may not use codes for the combination but instead enter codes for each of the individual ingredients into the patient record separately.

This situation can also arise with post-coordinated expressions.¹ For instance, a record for ICD10CM code E11.10 (Type 2 diabetes mellitus with ketoacidosis without coma) would be converted at OMOP ETL to two SNOMED-CT concepts with OMOP concept IDs 4009303 and 201826 (Diabetic ketoacidosis without coma and Type 2 diabetes mellitus). If both of those concept IDs were present in a concept set, it would match many unintended patients: those with Type 2 diabetes, regardless of ketoacidosis and those with diabetic ketoacidosis without restricting to Type 2. In order to capture the intended patients, it is necessary either to resort to use of source concepts or to check for patients with concept IDs 4009303 and 201826 on the same day (or at the same time.)

Real-world problem

The National COVID Cohort Collaborative (N3C)², an open science community organized in response to the pandemic in 2020, provides a centralized OMOP repository of patient-level data from dozens of partners. As members of the Data Liaison team, we provide terminology assistance, services, and resources to support N3C researchers. In recent months we have worked on the development and curation of a few dozen medication concept sets to be used across research projects. The only way we are currently able to address problems like those described above — where identifying a clinical phenomenon of interest requires the presence of codes from more than one concept set¹ — is by requiring researchers to use cohort definition logic to assure that codes from each set occur on the same day. We attempt this by including explicit instructions in the metadata for such concept set, but we do not have complete confidence that users of these concept sets will know how to do this (especially when use of a concept set must be combined with more complex logic, as with washout periods or first occurrences) or even that they will read these instructions before using the concept sets.

Proposed solution

We propose the introduction of standards for the definition of composite code sets.² The definition of a

¹ Or the presence of codes from one concept set and absence of codes from another, as, for instance, if trying to identify patients who had received only a single ingredient medications for ingredients that are sometimes prescribed in combination drugs.

² Since we believe the standard we propose should be used beyond the OHDSI community, we use the more general term “code set” to encompass OHDSI concept sets, FHIR value sets, and other definitional objects that resolve to a

composite code set will accommodate the same metadata as a regular concept set, will refer to two or more pre-existing regular concept sets and the boolean operations that are intended to be applied to their results.. A composite code set for Kaletra (a combination of lopinavir and ritonavir), for instance, would be defined as the intersection of the individual-ingredient concept sets. Whatever tools (e.g., ATLAS) the user of this concept uses to construct their cohort definition would need to implement a feature for processing composite concept sets. et us assume we have two regular concept sets defined:

lopinavir: OMOP concept ID 1738170, includeDescendants: true(135 member concepts)

ritonavir: OMOP concept ID 1748921, includeDescendants: true (424 member concepts)

Both of these concept sets include all combination drugs that include either ingredient.

With regular code sets, in order to match patients having the combination, we would need a cohort definition or algorithm that matches any patients having records from both sets on the same day. Or, to match patients having lopinavir only, the algorithm would need to subtract patients having ritonavir on the same day.

To illustrate, if we have patient records like:

person 1, Kaletra, 1/1/2023

person 2, lopinavir, 1/1/2023

person 2, ritonavir, 1/1/2023

person 3, lopinavir, 1/1/2023

person 3, ritonavir, 1/5/2023

A composite concept set definition for Kaletra (or any combination of lopinavir and ritonavir on the same day), assuming we already have concept sets named 'lopinavir' and 'ritonavir' would be defined (along with metadata) as

intersection(lopinavir, ritonavir)

and would result in:

person 1, 1/1/2023

person 2, 1/1/2023

Whereas, a composite concept definition for lopinavir used by itself would be defined as

difference(lopinavir, ritonavir)

and would result in:

person 3, 1/1/2023

Composite code set definitions will need to be processed by cohort definition tools like the cohort tab in ATLAS³ or the Logic Liaison Template in the N3C Enclave.⁴ For researchers who write their own cohort definition code without assistance from tools like ATLAS, we might provide generators that could produce code in SQL (as a subquery or stored procedure) (or R or Python functions, etc.) that would apply the component code sets, apply the defined set operations, and return the results for incorporation in the larger cohort definition algorithm.

set of terminology codes.

Conclusion

Though most OHDSI study authors develop their own concept sets, large research collaboratives like N3C have shown the need for a library of concept sets for common comorbidities and medications, authored by terminology and clinical subject matter experts. For a clinical concept that cannot be represented by a single concept set and requires taking the intersection or difference of patient records matched by multiple concept sets, composite concept sets can simplify cohort definition and facilitate the correct use of concept set libraries, while being self-explanatory and straightforward at the same time. When ATLAS and other cohort definition tools accommodate composite concept sets, study developers using concept set libraries will be able to find and use concept sets for these complex phenomena as easily as they currently use simple concept sets today. Concept set library authors will be able to provide concept set definitions that are impossible to include in concept set repositories today and to trust that users will employ them correctly.

Acknowledgements

Research reported in this work was supported by the National Institutes of Health's National Center for Advancing Translational Sciences, Grant Number U24TR002306. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. N3C Data Use Request number: 9b45ea07-e1e0-497b-bbc1-8520b61e73b8. IRB00327758: National Clinical Cohort Collaborative (N3C): A national resource for shared analytics.

References

1. Gold S, Zhang T, Zhu RL, Hong S, Lehmann HP, Gabriel D, et al. ICD10–SNOMED mapping pitfalls: Post-coordinated expressions and concept sets. In: 2022 OHDSI Symposium Collaborator Showcase [Internet]. Bethesda, Maryland; 2022 [cited 2023 Feb 24]. Available from: <https://www.ohdsi.org/2022showcase-21/>
2. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021 Mar 1;28(3):427–43.
3. OHDSI/Atlas [Internet]. *Observational Health Data Sciences and Informatics*; 2020 [cited 2020 May 5]. Available from: <https://github.com/OHDSI/Atlas/wiki>
4. Data Liaisons and Logic Liaisons Services | N3C [Internet]. [cited 2023 Jun 13]. Available from: <https://covid.cd2h.org/liaisons>