Constructing vaccine vocabulary hierarchy using formal concept analysis

Adam Black, Yupeng Li, Denys Kaduk, Licong Cui, Rashmie Abeysinghe, Lixia Yao

Background

Vaccine concepts in the OMOP CDM Vocabulary lack a comprehensive and consistent hierarchy. In 2021 the vaccine vocabulary working group manually constructed a high-level vocabulary hierarchy for vaccine concepts. However, a manually curated hierarchy is difficult to maintain and scale up with proper quality control considering the large number concepts in existing vaccine vocabularies such as RxNorm and RxNorm Extension. An automated approach is needed to facilitate the creation of a high-quality and practically useful vocabulary hierarchy.

We utilized Formal Concept Analysis (FCA), a computational method for automatically creating concept hierarchies rooted in the mathematical theory of lattices, to build a vaccine vocabulary. The required inputs are simply the vaccine source codes with the vaccine attributes (e.g., indication, administration route, and dosing). All nodes/concepts, linkage/hierarchical relationships between nodes and mappings from source to standard concepts are generated automatically.

Methods

The FCA method requires all vaccine source codes along with their relevant attributes as input. Given a large number of vaccine source codes in multiple vocabularies (e.g. CVX, NDC, CPT, and ICD Procedure), we used a subset of CVX, HCPCS, CPT, and ICD Procedure codes with two vaccine attributes (indication and mechanism of action) for feasibility demonstration. An example of the input structure is shown in Table 1.

Table 1. Illustration of the input to the FCA tool

concept_id	concept_name	vocab	indication1	indication2	indication3	mechanism of action 1	mechanism of action 2
40213291	diphtheria, tetanus toxoids and pertussis vaccine	CVX	diphtheria	tetanus	pertussis	diphtheria toxoid	tetanus toxoid
40213190	trivalent poliovirus vaccine, live, oral	CVX	poliovirus			poliovirus live	
40213183	measles, mumps and rubella virus vaccine	CVX	measles	rubella	mumps		
40213168	measles and rubella virus vaccine	CVX	measles	rubella			
40213170	measles virus vaccine	CVX	measles				

Our tool first creates a table with all unique combinations of indication and mechanism of action in the input and then uses the FCA method to derive the hierarchical relationships among these concepts. Attributes that are not explicitly encoded in the decomposition of the input source codes are ignored. Two source codes with the same set of attributes are considered to be the same vaccine and mapped to the same concept in the new hierarchy.

Results

The tool generates three tables as output: a concept table with all combinations of disease and mechanism attributes that exist in the input (source codes), a concept_relationship table which contains the hierarchical relationships ("Is a" relationship_id), and a mapping table containing the "Maps to" relationships that can be used to map the source codes to the standard concepts. As new vaccine source codes are added to the input, the algorithm will create new nodes in the graph as needed to completely represent the set of attributes in the source data.

The resulting hierarchy contains all single and combination "disease" level concepts as well as all single and combination "mechanism" concepts. For example, the "haemophilus influenzae/hepatitis B vaccine" concept will include all vaccines with both "heamophilus influenzae" and "hepatitis B vaccine" disease attributes as descendants. All vaccine concepts are descendants of one top level "vaccine" concept. All code for this project will be made available on GitHub. (2)

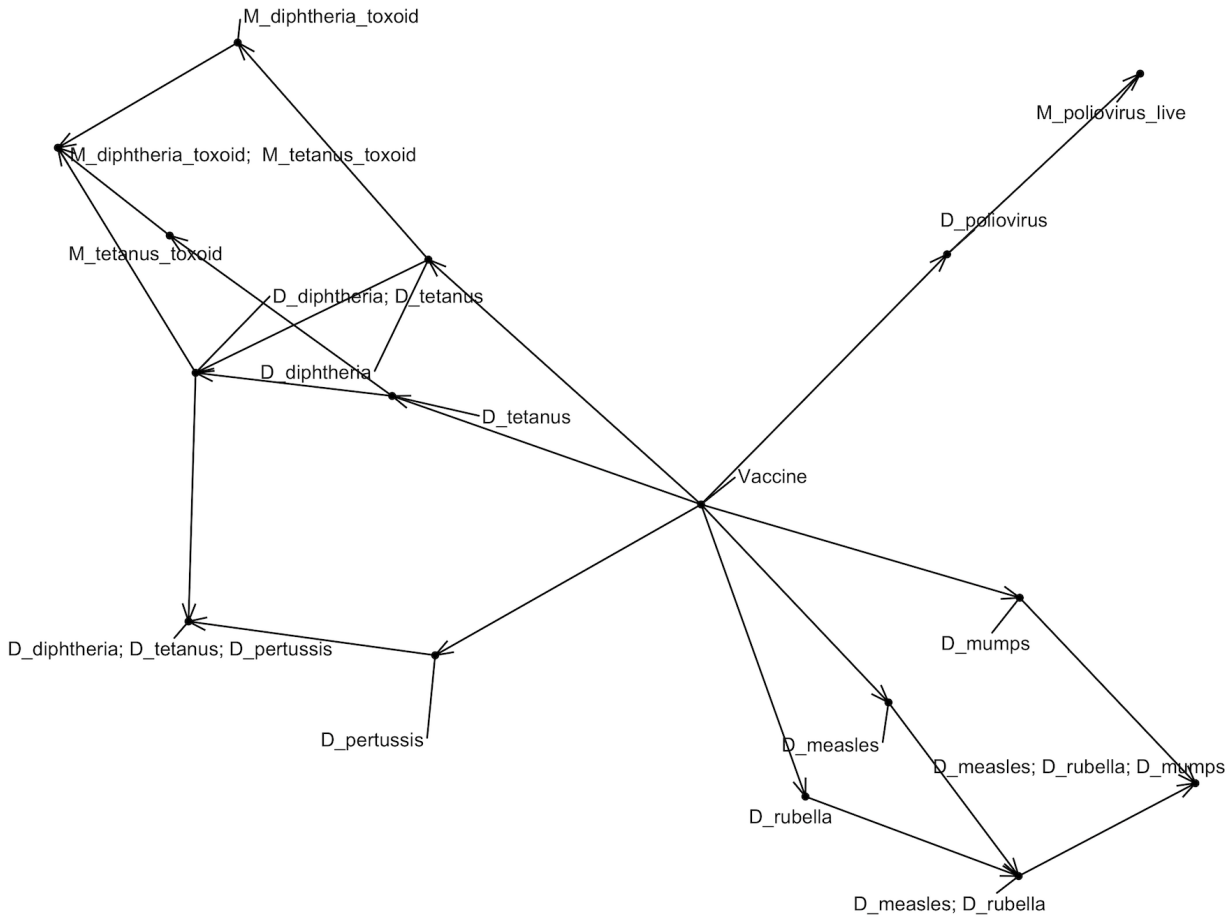

Figure 1. Example vaccine hierarchy automatically produced from decomposed vaccine source codes in Table 1

Conclusion

Due to all the various combinations that can occur in vaccines, it is difficult to manually curate and maintain a concept hierarchy that can faithfully represent all combinations. Our approach demonstrates the feasibility of an automated approach utilizing FCA to efficiently construct a high-level vocabulary hierarchy. In addition to being useful for the creation of a new vaccine hierarchy, this tool could be used to simplify the existing approach to building hierarchies for all drug concepts.

References

1. Ganter, Bernhard, and Rudolf Wille. Formal concept analysis: mathematical foundations. Springer Science & Business Media, 2012.
2. https://github.com/OHDSI/VaccineVocabulary