Journal of Biomedical Informatics 134 (2022) 104162

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method

Rashmie Abeysinghe [a,1], Adam Black [b,1], Denys Kaduk [b,1], Yupeng Li [c,1], Christian Reich [d,e], Alexander Davydov [b], Lixia Yao [c,*], Licong Cui [f,*]

[a] Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX, USA
[b] Odysseus Data Services, Cambridge, MA, USA
[c] Merck & Co., Inc., Rahway, NJ, USA
[d] IQVIA, Cambridge, MA, USA
[e] Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA
[f] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

ARTICLE INFO

Keywords:
Vaccines
OMOP standardized vocabularies
Concept mappings
Mapping quality assurance

ABSTRACT

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) provides a unified model to integrate disparate real-world data (RWD) sources. An integral part of the OMOP CDM is the Standardized Vocabularies (henceforth referred to as the OMOP vocabulary), which enables organization and standardization of medical concepts across various clinical domains of the OMOP CDM. For concepts with the same meaning from different source vocabularies, one is designated as the standard concept, while the others are specified as non-standard or source concepts and mapped to the standard one. However, due to the heterogeneity of source vocabularies, there may exist mapping issues such as erroneous mappings and missing mappings in the OMOP vocabulary, which could affect the results of downstream analyses with RWD. In this paper, we focus on quality assurance of vaccine concept mappings in the OMOP vocabulary, which is necessary to accurately harness the power of RWD on vaccines. We introduce a semi-automated lexical approach to audit vaccine mappings in the OMOP vocabulary. We generated two types of vaccine-pairs: mapped and unmapped, where mapped vaccine-pairs are pairs of vaccine concepts with a "Maps to" relationship, while unmapped vaccine-pairs are those without a "Maps to" relationship. We represented each vaccine concept name as a set of words, and derived term-difference pairs (i.e., name differences) for mapped and unmapped vaccine-pairs. If the same term-difference pair can be obtained by both mapped and unmapped vaccine-pairs, then this is considered as a potential mapping inconsistency. Applying this approach to the vaccine mappings in OMOP, a total of 2087 potentially mapping inconsistencies were obtained. A randomly selected 200 samples were evaluated by domain experts to identify, validate, and categorize the inconsistencies. Experts identified 95 cases revealing valid mapping issues. The remaining 105 cases were found to be invalid due to the external and/or contextual information used in the mappings that were not reflected in the concept names of vaccines. This indicates that our semi-automated approach shows promise in identifying mapping inconsistencies among vaccine concepts in the OMOP vocabulary.

1. Introduction

Real-world data (RWD) is critical to evaluate the safety, effectiveness, and uptake of vaccines [1]. It has been used to detect rare or long-term adverse events from vaccine exposure that cannot be detected during clinical trials [2]. For example, Ray et al. used Electronic Health Record (EHR) data to investigate the risk of rheumatoid arthritis following tetanus, influenza, and hepatitis B vaccines and concluded that the study power was low even with a sample size of 1 million vaccinated people [3]. In another example, it was impossible to assess the efficacy of 4CMenB (Bexsero, GSK) and MenB-FHbp (Trumenba, Pfizer) in clinical trials prior to approval due to low incidence of invasive meningococcal disease. Both vaccines were approved based on immunogenicity tests alone and vaccine effectiveness had to be determined using RWD after inclusion of Meningococcal type B vaccination

* Corresponding authors.
E-mail addresses: lixia.yao@merck.com (L. Yao), licong.cui@uth.tmc.edu (L. Cui).
[1] Contributed equally

https://doi.org/10.1016/j.jbi.2022.104162
Received 19 May 2022; Received in revised form 13 July 2022; Accepted 9 August 2022
Available online 25 August 2022

into national immunization programs [4,5]. Vaccination administrations are captured in RWD using a variety of controlled vocabularies with different degrees of specificity. In the United States, vaccine records are often coded using the National Drug Code (NDC) [6], Healthcare Common Procedure Coding System (HCPCS) [7], Current Procedural Terminology (CPT) [8], and International Classification of Diseases Version 10 (ICD-10) [9,10]. Coding systems may vary depending on the data origin (e.g., insurance claims, EHRs and vaccine registries), insurance types (e.g., Medicare, Medicaid and commercial insurance), the setting where people receive the vaccine (e.g., pharmacies versus clinics), and the purpose of recording or study (e.g., billing, surveillance and quality control). These codes often do not specify the vaccine brand names and a single code may be used for multiple vaccine products. During the COVID-19 pandemic, it has become more important to differentiate vaccines developed by different manufacturers with disparate mechanisms of action (e.g., mRNA, vector, whole virus inactivated, and protein subunit vaccines) and composition (active components, conjugates, adjuvants, and preservatives), targeting distinct antigens, and using different manufacturing processes. These variations lead to different immunogenicity, effectivenesses, and adverse events. Clinically and semantically meaningful vaccine terms that discriminate different products and group them into categories (e.g., indicating the mechanism of action) are greatly needed to conduct health outcomes research and to address the needs of public health. In addition, the various vaccine coding systems currently in use represent different sets of semantic attributes at various levels of granularity, lack high-quality grouping, and lack mapping to each other, preventing researchers from exploiting the full potential of RWD.

The OMOP Common Data Model (CDM), a widely used RWD model developed by the Observational Health Data Sciences and Informatics (OHDSI) community, aims to address the issues of siloed data and small sample sizes by defining a general and flexible data structure and a series of preferred terminologies for diagnosis, medication, procedures, measurement, and other clinical events. The OMOP Standard Vocabularies (henceforth referred to as the OMOP vocabulary) form an integral part of the OMOP CDM, and consist of OMOP-internal vocabularies and external source vocabularies. The external source vocabularies adopted in the OMOP vocabulary include the Medical Subject Heading (MeSH) [11], International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) [12], Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [13], RxNorm [14], Logical Observation Identifier Names and Codes (LOINC) [15], and Current Procedural Terminology Fourth Edition (CPT4) [16]. Each concept in the OMOP vocabulary is assigned a domain such as "Condition", "Observation", "Drug", "Procedure", and "Visit", denoting the location of which the concept is expected to occur in the data tables of the OMOP CDM. Each concept is also assigned a unique OMOP concept ID that is distinct from the concept identifier or code provided by its source vocabulary [17]. For example, concept "*tetanus toxoid vaccine, inactivated*" from the source vocabulary RxNorm is assigned an OMOP concept ID: 529411, which is different from its RxNorm identifier 798306. Note that in the rest of the paper, unless otherwise specified, it is the OMOP concept ID that is specified within the parentheses following a concept name.

A central feature that the OMOP vocabulary provides is the mapping of concepts, which enables harmonization of equivalent concepts (i.e., concepts with the same meaning) among different source vocabularies. For a set of equivalent concepts representing the same meaning of a clinical event, one concept from a certain vocabulary is designated as the "standard concept", while the other concepts are specified as non-standard or source concepts and mapped to the standard one through the "Maps to" relationship [17]. For instance, concept "*Atrial fibrillation*" in the Condition domain is represented by MeSH code D001281, SNOMED CT code 49436004, ICD-9-CM code 427.31, ICD-10-CM code I48.91, CIEL code 148203, Nebraska Lexicon code 49436004, UK Biobank code 6-1471, and Read code G573000, where the SNOMED CT concept is designated as the standard concept and the others are specified as non-standard concepts and mapped to the standard concept.

When converting disparate source RWD datasets into the OMOP CDM, all non-standard concepts are mapped to equivalent standard OMOP concepts using the "Maps to" relationship. If an equivalent concept is not available, then the non-standard concept is mapped to a more generic standard concept [18]. Ideally the mappings should be made without information loss or mismatches. In reality, semantic incompleteness and erroneous mappings occur when the information from the non-standard concepts and standard concepts do not exactly match. For instance, the non-standard concept "*Adsorbed Tetanus Toxoid*" (35162837) is mapped to the standard concept "*tetanus toxoid vaccine, inactivated Injectable Solution*" (40086961). Tetanus toxoid is a purified preparation of inactivated tetanus toxin, and there are two types of preparation: fluid and adsorbed [19]. Therefore, the standard concept loses some granular information from the non-standard concept. There exists a standard concept "*tetanus toxoid, adsorbed*" (40213232), which could be a better targeting standard concept for the non-standard concept "*Adsorbed Tetanus Toxoid*" (35162837).

Various techniques have been developed to automatically perform such concept mappings between different terminologies or ontologies (so called ontology mapping or ontology matching) [20–24]. However, quality assurance of ontology mappings is an area that has rarely been investigated. In previous work, we performed a preliminary study to examine concept mappings between vaccine vocabularies used in the United States by leveraging the OMOP CDM [25,26]. We found that most vaccine administration events are recorded in RWD using imprecise procedure codes with limited or no brand information, whereas the OMOP CDM considers vaccination to be drug exposure. In our previous study, from the OMOP vocabulary, we retrieved 15,932 vaccine-related concepts that came from 32 source vocabularies and extracted 15,220 "Maps to" relations between the non-standard and standard vaccine concepts [26]. Among a collection of 1170 "Maps to" relations involving vaccine concepts occurring in five RWD datasets we had access to, it was found that 104 out of 1170 (8.89%) contain mapping inconsistencies by manual expert review. For instance, it was found that the non-standard concept "*hemophilus influenzae B, purified antigen conjugated; systemic*" (21601291) was incorrectly mapped to the standard concept "*Neisseria meningitidis*" (515671).

When the number of concepts involved is large, manual review of the mappings between source and standard concepts becomes difficult, time-consuming, and error-prone. Therefore, manual identification of mapping issues is unsustainable moving forward. Automated or semi-automated methods for quality assurance of vocabulary mappings are needed to efficiently identify potential mapping errors or direct human reviewers towards potential issues. In this paper, we develop a lexical approach to programmatically identify potential mapping issues among the OMOP vaccine concept mappings. To the best of our knowledge, this is the first work introducing a semi-automated method to assess the quality of mappings in the OMOP vocabulary.

2. Materials and methods

In this work, we leveraged the v5.0 release (29-OCT-2020) of the OMOP vocabulary, from which we extracted vaccine concept mappings. Then, we generated two types of vaccine-pairs: mapped and unmapped, where mapped vaccine-pairs refer to the pairs of vaccine concepts with a "Maps to" relationship, while unmapped vaccine-pairs refer to those pairs of vaccine concepts without a "Maps to" relationship. Representing each vaccine concept name as a set of words, we further generated a term-difference pair (i.e., name difference) for each vaccine-pair. If the same term-difference pair was obtained by both a mapped vaccine-pair and an unmapped vaccine-pair, then such a situation was considered as a potential mapping inconsistency. A random sample of such potential mapping inconsistencies were further manually investigated by domain experts to evaluate the effectiveness of our method and categorize the mapping issues.

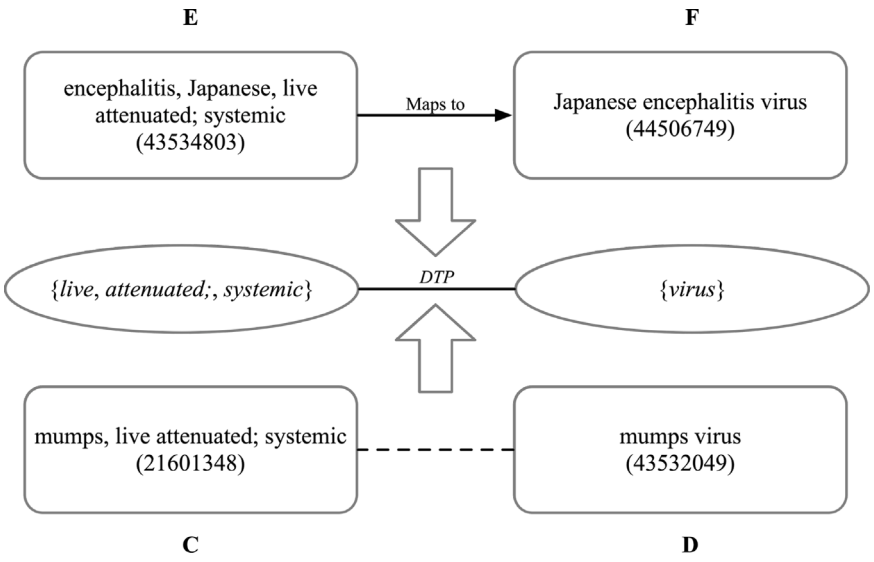

Fig. 1. Unmapped vaccine-pair C = "*mumps, live attenuated; systemic*" (21601348) and D = "*mumps virus*" (43532049) and mapped vaccine-pair E = "*encephalitis, Japanese, live attenuated; systemic*" (43534803) and F = "*Japanese encephalitis virus*" (44506749) both deriving the same TDP ({"*live*", "*attenuated*", "*systemic*"}, {"*virus*"}). Through manual review by domain experts, it was found that this indicates an erroneous existing mapping between the concepts E and F.

2.1. Extraction of vaccine concept mappings in OMOP vocabulary

The OMOP vocabulary is an integral part of the OMOP CDM [27]. The OMOP vocabulary contains over 100 source vocabularies. Concepts in the OMOP vocabulary represent clinical events and are the fundamental building blocks of the data records in the OMOP CDM. Given a clinical event, there may be multiple concepts from disparate source vocabularies representing the same meaning of the clinical event, where only one concept is designated as the standard concept and other concepts are considered as non-standard concepts (or source concepts) and mapped to the standard concept through the "Maps to" relationship. In previous work, we constructed a comprehensive list of vaccine concepts from the OMOP vocabulary using the iterative regular expression-based pattern matching on concept names, hierarchical relationships defined in the OMOP vocabulary, and manual review by experts [26]. Then we extracted all the vaccine concept mappings using the "Maps to" relationship where both the non-standard concept and standard concept were in the vaccine concept list. In this work, we reused the vaccine concept list and mappings to develop our lexical approach to automatically identify potential mapping inconsistencies.

2.2. Vaccine concept name representation

We represented each vaccine concept name as a set of words. More specifically, we boiled down the vaccine concept name into word tokens in lower case and modeled them as a set of words. For instance, the vaccine concept named "*Adsorbed Tetanus Toxoid*" (OMOP concept ID: 35162835) was represented as {"*adsorbed*", "*tetanus*", "*toxoid*"}. Note that since sets are unordered, the vaccine concept named "*Tetanus toxoid adsorbed*" (37396309) would generate the same set of words and share the same representation as "*Adsorbed Tetanus Toxoid*" (35162835).

2.3. Mapped and unmapped vaccine-pair generation

We constructed two collections of vaccine-pairs: mapped and unmapped. We generated mapped vaccine-pairs by direct retrieving of the "Maps to" relationships. More specifically, if a non-standard vaccine concept and a standard vaccine concept in our vaccine list were connected through the "Maps to" relationship in the OMOP vocabulary, then this pair of vaccine concepts was added to the mapped vaccine-pair list. For instance, the vaccine concepts "*Rubella Vaccine*" (45742246) and "*rubella virus vaccine*" (40213223) in the OMOP vocabulary form a mapped vaccine-pair. If a non-standard vaccine concept and a standard vaccine concept were not connected through the "Maps to" relationship, then this pair of vaccine concepts was added to the unmapped vaccine-pair list. For example, the vaccine concepts "*mumps, live attenuated; systemic*" (21601348) and "*mumps virus*" (43532049) form an unmapped vaccine-pair.

2.4. Vaccine difference pair generation

For each mapped or unmapped vaccine-pair, say (A, B), we generated a term-difference pair representing the name difference between A and B. Let $S(A)$ and $S(B)$ be the set-of-words of A and B respectively, we generated a term-difference pair (TDP) between A and B as follows:

$$TDP(A, B) = (\{x \mid x \in S(A) \text{ and } x \notin S(B)\},$$
$$\{x \mid x \in S(B) \text{ and } x \notin S(A)\}).$$

Stated in another way, the set differences $(S(A) - S(B), S(B) - S(A))$ form the TDP between A and B. For instance, consider the unmapped vaccine-pair C = "*mumps, live attenuated; systemic*" (21601348) and D = "*mumps virus*" (43532049), with $S(C)$ = {"*mumps*", "*live*", "*attenuated*", "*systemic*"} and $S(D)$ = {"*mumps*", "*virus*"} in Fig. 1. These generate a term-difference pair between C and D: $TDP(C, D)$ = ({"*live*", "*attenuated*", "*systemic*"}, {"*virus*"}). Similarly, the mapped vaccine-pair M = "*Rubella Vaccine*" (45742246) with $S(M)$ = {"*rubella*", "*vaccine*"} and N = "*rubella virus vaccine*" (40213223) with $S(N)$ = {"*rubella*", "*virus*", "*vaccine*"} in Fig. 2 generate $TDP(M, N)$ = ({}, {"*virus*"}). Note that the first set in this TDP is an empty set, since $S(M)$ is a proper subset of $S(N)$, .

2.5. Identifying potential vaccine mapping inconsistencies

Let (A, B) be an unmapped vaccine-pair and (X, Y) be a mapped vaccine-pair. If (A, B) and (X, Y) generate the same TDP, that is,

$$TDP(A, B) = TDP(X, Y),$$

then we suggest that there is a potential mapping inconsistency. Through manual review by domain experts, such inconsistencies could be generally classified into the following categories:

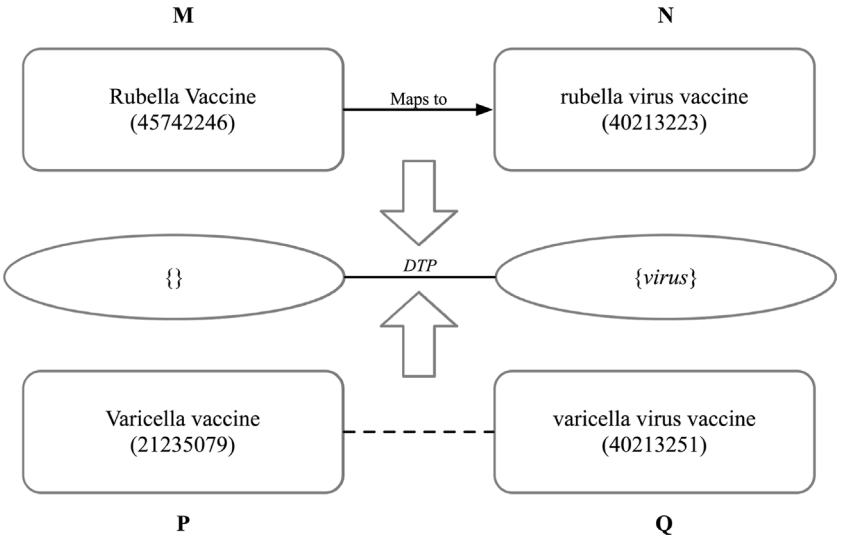

Fig. 2. Unmapped vaccine-pair P = "*Varicella vaccine*" (21235079) and Q = "*varicella virus vaccine*" (40213251) and mapped vaccine-pair M = "*Rubella vaccine*" (45742246) and N = "*rubella virus vaccine*" (40213223) both deriving the same TDP ({}, {"*virus*"}). Through manual review by domain experts, it was found that indicates a missing mapping between the concepts P and Q.

1. Erroneous existing mapping between X and Y;
2. Missing mapping between A and B;
3. Other issues around concepts A, B, X, and Y;
4. False positive (i.e., no mapping quality issue revealed).

The basic rationale of the classification of these categories is that the TDP may indicate a certain relation between the mapped concepts X and Y. If domain experts do not agree with such a relation and believe that the mapping between concepts X and Y is incorrect, then this scenario reveals an erroneous existing mapping between X and Y. If domain experts agree with such a relation and believe that it should also apply to the currently unmapped concepts A and B, then this scenario reveals a missing mapping between A and B. In other scenarios, domain experts may agree or disagree with such a relation, but they neither believe that A and B should be mapped nor believe that the mapping between X and Y is incorrect. In such cases, if domain experts can identify other issues regarding concepts A, B, X and Y, then it is classified into the category of "Other issues around concepts A, B, X, and Y"; otherwise, no mapping issue is revealed (i.e., it is a false positive suggested by our method).

For instance, in Fig. 1, the $TDP(C, D)$ = ({"*live*", "*attenuated*", "*systemic*"}, {"*virus*"}) obtained by unmapped vaccine-pair C = "*mumps, live attenuated; systemic*" (21601348) and D = "*mumps virus*" (43532049) is the same as the $TDP(E, F)$ obtained by mapped vaccine-pair E = "*encephalitis, Japanese, live attenuated; systemic*" (43534803) and F = "*Japanese encephalitis virus*" (44506749). Manual review of this mapping inconsistency by domain experts reveals an erroneous existing mapping between E = "*encephalitis, Japanese, live attenuated; systemic*" and F = "*Japanese encephalitis virus*". In Fig. 2, unmapped vaccine-pair P = "*Varicella vaccine*" (21235079) and Q = "*varicella virus vaccine*" (40213251) generate $TDP(P, Q)$ = ({}, {"*virus*"}), which is the same as the $TDP(M, N)$ generated by the mapped vaccines M = "*Rubella Vaccine*" (45742246) and N = "*rubella virus vaccine*" (40213223). Manual review of this inconsistency by domain experts reveals a missing mapping between P = "*Varicella vaccine*" and Q = "*varicella virus vaccine*".

2.6. Evaluation of potential vaccine mapping inconsistencies

To assess the effectiveness of our TDP method, a random sample of potential mapping inconsistencies were selected and manually reviewed by domain experts (authors AB, DK, and YL) to evaluate the validity of these mapping issues. The samples were picked such that no two samples contain the same TDP in order to avoid evaluating similar issues. The domain experts were provided with the mapped vaccine-pairs, unmapped vaccine-pairs, as well as the TDPs. Each sample was evaluated by two experts. The disagreements between experts were resolved through discussion and the final review for each sample was agreed upon by both reviewers who evaluated it.

Table 1
Vocabularies of non-standard and standard concepts involved in the mapped vaccine-pairs of potential mapping inconsistencies identified.

	Vocabulary	Number of unique concepts
Non-standard (or source) concepts	NDC	849
	RxNorm	106
	Gemscript	72
	dm+d	59
	VA Product	50
	Read	40
	RxNorm Extension	17
	SNOMED CT	17
	CIEL	13
	Nebraska Lexicon	12
	ATC	11
	MeSH	9
	CTD	5
	BDPM	4
	SPL	4
	JMDC	3
	NCCD	2
	CPT4	2
	AMT	1
	GGR	1
	HCPCS	1
Standard concepts	RxNorm	394
	RxNorm Extension	51
	CVX	23
	SNOMED CT	4

3. Results

We extracted 15,932 vaccine-related concepts from the OMOP vocabulary in previous work [26], from which 15,220 "Maps to" relations were identified among 14,570 vaccine concepts. Among these, 10,456

Table 2
Categorizing the mapping issues among the 95 cases with mapping inconsistencies. The "Other issues" have also been subcategorized.

Issue type	Number of pair
1. Erroneous existing mapping between mapped vaccine-pair	46
2. Missing mapping between unmapped vaccine-pair	6
3(1). Other issues: RxNorm upgrade	18
3(2). Other issues: Duplicated concepts in two vocabularies	17
3(3). Other issues: Standard vaccine concept in the Observation domain	8

Table 3
Ten examples of valid mapping inconsistencies verified by domain experts. Issue types: 1 = Erroneous existing mapping between mapped vaccine-pair; 2 = Missing mapping between unmapped vaccine-pair; 3(1) = Other issue: RxNorm upgrade (Concept X deprecated in RxNorm); 3(2) = Other issue: Concepts B & Y are duplicated standard concepts in two vocabularies; 3(3) = Other issue: Standard concept B is in the Observation domain.

Unmapped vaccine-pair		Mapped vaccine-pair		Issue
Concept A	Concept B	Concept X	Concept Y	type
First diphtheria vaccination (45428834)	Diphtheria antitoxin (19031041)	First diphtheria vaccination (45428834)	Diphtheria antitoxin (40213279)	1
Influenza vaccine (split virion, inactivated) suspension for injection 0.5 ml pre-filled syringes (37855865)	Haemophilus influenzae b (Ross strain) capsular polysaccharide meningococcal protein conjugate vaccine Injectable Solution (40045078)	Influenza vaccine (split virion, inactivated) suspension for injection 0.5 ml pre-filled syringes (37887008)	Haemophilus influenzae b (Ross strain) capsular polysaccharide meningococcal protein conjugate vaccine Injectable Solution (40045078)	1
Tetanus toxoid adsorbed 10 UNT/ML (529487)	Tetanus toxoid vaccine, inactivated 10 UNT/ML Injectable Solution (40822025)	Adsorbed Tetanus Toxoid (35162837)	Tetanus toxoid vaccine, inactivated Injectable Solution (40086961)	1
Varicella vaccine (21235079)	Varicella virus vaccine (40213251)	Rubella Vaccine (45742246)	Rubella virus vaccine (40213223)	2
Diphtheria/tetanus/pertussis dtpw vaccination (37876857)	Diphtheria toxoid vaccine, inactivated (529303)	Diphtheria/tetanus/pertussis dtpw vaccination (37898539)	Diphtheria toxoid vaccine, inactivated (529303)	2
0.5 ML Varicella-Zoster Virus Vaccine Live (Oka-Merck) strain 2700 UNT/ML Injectable Solution [Varivax] (43774093)	Varicella-zoster virus vaccine live (Oka-Merck) strain 2700 UNT/ML Injection [Varivax] (46275141)	0.5 ML Haemophilus influenzae type b strain 1482, capsular polysaccharide inactivated tetanus toxoid conjugate vaccine 0.068 MG/ML Injectable Solution [ActHIB] (42800273)	Haemophilus influenzae type b strain 1482, capsular polysaccharide inactivated tetanus toxoid conjugate vaccine 0.068 MG/ML Injection [ActHIB] (46275194)	3(1)
Rabies antigen 2.5 UNT/ML Injectable Solution (40119951)	Rabies virus vaccine flury-lep strain 2.5 UNT/ML Injectable Solution (43822935)	Rabies antigen 2.5 UNT/ML [RabAvert] (40119954)	Rabies virus vaccine flury-lep strain 2.5 UNT/ML [RabAvert] (19133403)	3(1)
DTaP containing vaccines (45949555)	Pertussis vaccine (40213196)	DTaP containing vaccines (45949555)	Pertussis vaccine (19033193)	3(2)
Product containing diphtheria antitoxin (3206241)	Diphtheria antitoxin (40213279)	Product containing diphtheria antitoxin (3206241)	Diphtheria antitoxin (19031041)	3(2)
Influenza virus live attenuated (21271897)	Influenza B virus antigen (4044621)	Influenza virus live attenuated (21271897)	Influenza B virus antigen (46275999)	3(3)

were standard concepts and 4114 were non-standard concepts. We generated 4764 mapped vaccine-pairs and derived 2798 unique TDPs from these mapped vaccine-pairs. Out of these TDPs, 509 were observed among unmapped vaccine-pairs leading to 2087 potential mapping inconsistencies (see Supplementary data). Table 1 shows the vocabularies of the non-standard and standard concepts involved in the mapped vaccine-pairs of potential mapping inconsistencies, as well as the number of unique concepts involved in each vocabulary. Note that for a certain vocabulary (e.g., SNOMED CT), it may appear as a vocabulary for both non-standard concepts and standard concepts, since in some mapped pairs the non-standard concept comes from this vocabulary, while in some other mapped pairs the standard concept comes from this vocabulary.

To evaluate the effectiveness of the TDP method and further understand the underlying mapping issues, we randomly selected 200 potential mapping inconsistencies for domain experts' manual review (see Supplementary data). Domain experts confirmed quality issues exist in 95 cases. We classified these mapping issues into 3 categories (see Table 2). Note that the "Other issues" in Table 2 have been further categorized into 3 subcategories.

Table 3 shows 10 examples of valid mapping issues confirmed by domain experts.

3.1. Erroneous existing mapping between mapped vaccine-pair (issue type 1)

There were 46 cases where the mapping between the mapped vaccine-pair was found to be erroneous. Out of these cases, 9 were absolutely wrong. For example, "*diphtheria vaccine*" was mapped to "*diphtheria antitoxin*"; and "*influenza vaccine*" was mapped to "*Haemophilus influenzae b vaccine*". The other 37 cases were not completely wrong, but had more or less information change between the non-standard and standard concepts. For example, "*Hepatitis b 20 microgram/ml Vaccination*" (45740521) was mapped to "*hepatitis B virus*" (43532406). The dose information was lost in the mapped standard concept. Ideally, it can be mapped to the RxNorm concept "*Hepatitis B Vaccine 0.02 MG/ML*" (501524), but the RxNorm concept is not a standard OMOP concept. The closest standard concept could be "*hepatitis B surface antigen vaccine 0.02 MG/ML*" (528324).

3.2. Missing mapping between unmapped vaccine-pair (issue type 2)

There were 6 cases where the unmapped vaccine-pair actually should be mapped (i.e., missing mappings). For instance, as shown in Fig. 2, mapped vaccine-pair "*Rubella Vaccine*" (45742246) and "*rubella virus vaccine*" (40213223) derives the same TDP as the unmapped vaccine-pair "*Varicella vaccine*" (21235079) and "*varicella*

virus vaccine" (40213251). Further examination of concept "*Varicella vaccine*" (21235079) reveals that it was mapped to "*varicella-zoster virus vaccine live (Oka-Merck) strain*" (42800027), which is a more granular concept than the source concept. In this instance the concept "*varicella virus vaccine*" (40213251) is a more appropriate concept to map to. In another case, "*Diphtheria/tetanus/pertussis dtpw vaccination*" (37876857) was only mapped to "*pertussis vaccine*" (19033193), and diphtheria and tetanus ingredients were missed. Based on the mapping of "*Diphtheria/tetanus/pertussis dtpw vaccination*" (37898539) to three individual ingredients, we can add two additional mappings for the concept 37876857.

3.3. Other issues

We found 43 samples to denote other issues surrounding the mapped vaccine-pair and/or the unmapped vaccine-pair. These are categorized as follows.

3.3.1. RxNorm upgrade - Issue type 3(1)

In this category, an RxNorm code was mapped to another RxNorm code according to the RxNorm internal concept replacement process. Some RxNorm codes are deprecated occasionally by its developers, and because OMOP vocabulary does not delete any historical concepts, these codes were set to non-standard, and were mapped to other standard concepts, mostly RxNorm codes provided by the source. Since RxNorm codes are unique, the mapped RxNorm codes would not contain exactly the same information and thus the mappings have more or less information loss. For example, "*0.5 ML Haemophilus influenzae type b strain 1482, capsular polysaccharide inactivated tetanus toxoid conjugate vaccine 0.068 MG/ML Injectable Solution [ActHIB]*" (42800273) was remapped to "*Haemophilus influenzae type b strain 1482, capsular polysaccharide inactivated tetanus toxoid conjugate vaccine 0.068 MG/ML Injection [ActHIB]*" (46275194), where the total volume "0.5 ML" is missing in the standard concept (46275194). In total 18 samples belonged to this category.

3.3.2. Duplicated concepts in two vocabularies - Issue type 3(2)

There were 17 cases of standard concepts from different vocabularies, which were all in the Drug domain and had the same concept name, e.g., both RxNorm ingredient concept 19033193 and CVX concept 40213196 are "*pertussis vaccine*". The majority of the vaccine related standard concepts are RxNorm codes. CVX codes were introduced as standard concepts to accommodate vaccination procedure codes in claims data and clinical notes in EHR. But there are a few duplicated concepts between CVX and RxNorm codes. Although both are in the Drug domain, a high-quality vocabulary should be concise and have those duplicated concepts destandardized and mapped over to each other.

3.3.3. Standard vaccine concept in the observation domain - Issue type 3(3)

There were 8 concepts mapped to an RxNorm ingredient concept, "*influenza B virus antigen*" (46275999). However, there is a standard SNOMED substance concept with the exact same concept name (4044621) identified by TDP. But it is assigned to the Observation domain. Based on the OMOP CDM guideline, vaccines should be mapped to the Drug domain. The standard vaccine concepts in the Observation domain may cause vaccine records mapped to the Observation domain instead of the Drug domain. When a user follows the OMOP guideline and queries vaccine records in the Drug domain only, they would miss those records in the Observation domain. In addition, 5 of the source concepts should not be mapped to "*influenza B virus antigen*" (4044621), e.g., "*Influenza virus live attenuated*" (21271897), "*influenza, inactivated, split virus or surface antigen; systemic*" (21601335), "*influenza, live attenuated; systemic*" (21601336), "*Influenza virus surface antigens*" (21178596), and "*Influenza virus surface antigens virosome*" (21197412), because these concepts do not specify *B* type of influenza.

3.4. False positives

The remaining 105 cases were false positives indicating that the mapping between the mapped vaccine-pair is accurate, the unmapped vaccine-pair does not form a valid mapping, and no other issues were discovered surrounding the mapped and unmapped vaccine-pairs. It should be noted that all these cases were NDC to RxNorm mappings. Standard OMOP concepts in the Drug domain are mostly RxNorm codes, and the mapping from vaccine NDC codes to RxNorm codes were directly retrieved from the Unified Medical Language System (UMLS) [28]. These mappings were manually curated by the UMLS community based on the product details, which usually were not included in the concept name. For example, the concept "*influenza virus vaccine 15ug/.5mL INTRAMUSCULAR SUSPENSION [flulaval quadrivalent 2015/2016]*" (46366728) was mapped to "*influenza A virus A/California/7/2009 (H1N1) antigen 0.03 MG/ML/influenza A virus A/Switzerland/9715293/2013 (H3N2) antigen 0.03 MG/ML/influenza B virus B/Brisbane/60/2008 antigen 0.03 MG/ML/influenza B virus B/Phuket/3073/2013 antigen 0.03 MG/ML Injectable Suspension [FluLaval Quadrivalent 2015-2016]*" (46276189). Although the source concept name does not specify the individual influenza strains and dose, we can get this information based on the NDC code. The NDC codes are unique identifiers for drug products in the United States, and provide information of manufacturer, drug name, dosage, strength, formulation and package size of specific drug products.

Our TDP method focuses on the concept names and assumes concept names contain the concept information with full transparency, so it cannot properly assess the mappings inferred from information other than concept names. Nevertheless, even though it is difficult to assess the mappings from NDC codes to RxNorm codes, we found NDC codes were mapped to different types of RxNorm codes that contain different levels of granularity. Ideally, we can get all the product details from NDC codes and map to the most granular level of RxNorm codes, i.e., Quant Branded Drug or Quant Clinical Drug. Mapping to higher levels of RxNorm codes indicates potential information loss, for example, Branded Drug or Clinical Drug loses the drug quantity information. Inconsistent rules for utilizing information from NDC codes when mapping to RxNorm may be one reason for the inconsistent mapping. Another possible reason is due to the RxNorm upgrade, some of the corresponding Quant Branded or Clinical Drug RxNorm codes were deprecated, and the mapping has to compromise.

4. Discussion

We have developed a lexical-based approach to identify mapping inconsistencies by comparing term-difference pairs of mapped and unmapped pairs of concepts. Focusing on vaccine-related concepts, we successfully demonstrated the effectiveness of our method. The method can greatly increase the efficiency in identifying and prioritizing possible mapping inconsistencies and reduce the time and burden from manual review by human experts. Our lexical-based approach is generally applicable for auditing concept mappings across different terminologies.

4.1. Experiment with non-vaccine concepts

To demonstrate the generalizability of the approach, we applied our TDP method to all the concepts in the Condition domain of the OMOP vocabulary, which contains 579,400 concepts with 603,778 "Maps to" relations. This resulted in 139,203 potential mapping inconsistencies. To validate the obtained potential inconsistencies, we leveraged the Unified Medical Language System (UMLS) [29] that provides a mapping structure among different terminologies by grouping concepts conveying the same meaning under the same UMLS concept. Given a potential mapping inconsistency, which involves a mapped pair of

concepts and an unmapped pair of concepts sharing the same term-difference pair, if the unmapped concepts are grouped under the same UMLS concept, then we consider it as a valid inconsistency.

Out of 139,203 potential mapping inconsistencies, 2406 of them had unmapped concept-pairs that could be mapped to UMLS concepts. Among these, 94 were grouped under the same UMLS concept and considered as valid cases. Out of these 94 valid cases, 8 represent situations with duplicated standard concepts, while the remaining 86 represent missing mappings between unmapped concept-pairs. For instance, the non-standard concept "*Pain in joints of left hand*" (37200702) should be mapped to the standard concept "*Joint pain in left hand*" (759906). In addition, the standard concepts "*Hypoglycemia*" (42600315) and "*Hypoglycemia*" (24609) are duplicates.

4.2. Comparison with related work

To our knowledge, this is the first work in quality assurance of the concept mappings in the OMOP vocabulary. Previously, we used a similar lexical-based approach for a different application to audit hierarchical relations in the Gene Ontology [30], where the concept-pairs chosen were limited to having the same number of words, at least one word in common and n different words (n = 1, 2, 3, 4, or 5). In this work, we did not have such restrictions in picking concept-pairs. Another major distinction is that the pairs of concepts in this work belong to different vocabularies, while the concept-pairs in our previous work are within the same vocabulary.

4.3. Limitations and future directions

Our lexical-based method has a couple of limitations. First, the method cannot identify inconsistent cases when there does not exist any unmapped vaccine-pair that shares the same TDP with a mapped vaccine-pair. For example, "*Influenza virus vaccine, trivalent (IIV3), split virus, preservative free, 0.25 mL dosage, for intramuscular use*" (2213437) was mapped to "*Influenza, seasonal, injectable, preservative free*" (40213154), where information of trivalent and dosage were dropped. This mapping inconsistency was not captured by our TDP method. In our previous work on manual identification of mapping issues [26], we found 104 mapping inconsistencies by manual review, but only 5 of these issues showed up in the mapping inconsistencies identified in this work. On the other hand, however, this work covered a vast number of issues that were not captured by the previous manual approach. For instance, 93 valid mapping inconsistencies identified from a random sample of 200 potential mapping inconsistencies in this work were not captured by the previous work. It is worth noting that quality assurance work is discovery oriented and is not expected to identify all existing mapping issues by a single method. Therefore, more investigation is needed to develop additional methods to systematically uncover other potential cases.

Secondly, our method relies on the assumption that the mapping is fully based on the information in the concept names. However, concepts in some coding systems have much more information beyond the concept names. For example, an NDC code represents a unique drug product in the United States, and additional information can be retrieved using the NDC code from NDC Directory [6], and its published package insert, e.g., manufacturer, detailed ingredients, virus or bacteria strain type, administration route, dosage and quantity of the package, eligible population, etc. We observed many mappings from NDC codes to RxNorm codes that used that information beyond concept names. Therefore, our method is more likely to produce false positives (or invalid mapping inconsistencies) when extra information in addition to concept names is used for mapping. In the future, we would like to explore whether infusing the method with other internal and external information would have a positive impact on addressing such issues.

Note that our TDP method in this work was used for quality assurance of the existing mappings rather than introducing a new approach for performing ontology mapping. However, it would be interesting to apply the TDP method to compare the quality of mapping results obtained by different ontology mapping approaches (e.g., based on the number of identified inconsistencies).

It should also be noted that this approach can be generalized to audit relations within a terminology such as identifying missing or erroneous relations. Since a relation in a terminology always involves a pair of concepts, our term-difference pair approach may be applied to the related concept pairs and unrelated concept pairs to detect potential inconsistencies, which may reveal missing relations and erroneous relations by human expert review.

5. Conclusion

We have developed a new semi-automated method for identifying vaccine concept mapping inconsistencies in the OMOP vocabulary. Domain expert evaluation showed promising performance of our method. The mapping issues we identified highlight the need for semi-automated or fully-automated quality control methods for concept mappings since these issues may affect downstream applications and analyses. More research is needed to assess the applicability and effectiveness of our TDP-based method to other clinical domains for quality assurance of concept mappings.

CRediT authorship contribution statement

Rashmie Abeysinghe: Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Adam Black:** Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Denys Kaduk:** Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Yupeng Li:** Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Christian Reich:** Data curation, Writing – review & editing. **Alexander Davydov:** Data curation, Writing – review & editing. **Lixia Yao:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Licong Cui:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lixia Yao and Yupeng Li are employees and stockholders of Merck & Co., Inc., Rahway, NJ.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF), USA through grant 2047001, National Institutes of Health (NIH), USA through grants R01LM013335 and R01NS116287 received by LC. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

Appendix A. Supplementary data

Spreadsheets contain 2087 potential mapping inconsistencies and results of 200 manually evaluated random samples.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104162.

References

[1] J.R. Verani, A.H. Baqui, C.V. Broome, T. Cherian, C. Cohen, J.L. Farrar, D.R. Feikin, M.J. Groome, R.A. Hajjeh, H.L. Johnson, et al., Case-control vaccine effectiveness studies: Preparation, design, and enrollment of cases and controls, Vaccine 35 (25) (2017) 3295–3302.

[2] J. Baggs, J. Gee, E. Lewis, G. Fowler, P. Benson, T. Lieu, A. Naleway, N.P. Klein, R. Baxter, E. Belongia, et al., The vaccine safety datalink: a model for monitoring immunization safety, Pediatrics 127 (Supplement_1) (2011) S45–S53.

[3] P. Ray, S. Black, H. Shinefield, A. Dillon, D. Carpenter, E. Lewis, P. Ross, R.T. Chen, N.P. Klein, R. Baxter, et al., Risk of rheumatoid arthritis following vaccination with tetanus, influenza and hepatitis B vaccines among persons 15–59 years of age, Vaccine 29 (38) (2011) 6592–6597.

[4] R. Borrow, M.-K. Taha, M.M. Giuliani, M. Pizza, A. Banzhoff, R. Bekkat-Berkani, Methods to evaluate serogroup B meningococcal vaccines: from predictions to real-world evidence, J Infect. 81 (6) (2020) 862–872.

[5] C. Isitt, C.A. Cosgrove, M.E. Ramsay, S.N. Ladhani, Success of 4CMenB in preventing meningococcal disease: evidence from real-world experience, Arch. Dis. Child. 105 (8) (2020) 784–790.

[6] The United States Food and Drug Administration, National drug code directory, 2021, (accessed 13 December 2021) https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory.

[7] The Centers for Medicare and Medicaid Services (CMS), HCPCS - general information, 2021, (accessed 13 December 2021) https://www.cms.gov/medicare/coding/medhcpcsgeninfo.

[8] P. Dotson, CPT® codes: What are they, why are they necessary, and how are they developed? Adv. Wound Care (New Rochelle) 2 (10) (2013) 583–587.

[9] The Centers for Medicare and Medicaid Services (CMS), 2021 ICD-10-PCS, 2021, (accessed 13 December 2021) https://www.cms.gov/medicare/icd-10/2021-icd-10-pcs.

[10] J. Hirsch, G. Nicola, G. McGinty, R. Liu, R. Barr, M. Chittle, L. Manchikanti, ICD-10: history and context, AJNR Am. J. Neuroradiol. 37 (4) (2016) 596–599.

[11] C.E. Lipscomb, Medical subject headings (MeSH), Bull. Med. Libr. Assoc. 88 (3) (2000) 265–266.

[12] Center for Disease Control and Prevention, International classification of diseases, ninth revision, clinical modification (ICD-9-CM), 2022, (accessed 13 July 2022) https://www.cdc.gov/nchs/icd/icd9cm.htm#:~:text=ICD%2D9%2DCM%20is%20the,10%20for%20mortality%20coding%20started.

[13] K. Donnelly, et al., SNOMED-CT: The advanced terminology and coding system for ehealth, Stud. Health Technol. Inf. 121 (2006) 279–290.

[14] S. Liu, W. Ma, R. Moore, V. Ganesan, S. Nelson, RxNorm: prescription for electronic drug information exchange, IT Prof. 7 (5) (2005) 17–23.

[15] C.J. McDonald, S.M. Huff, J.G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, et al., LOINC, a universal standard for identifying laboratory observations: a 5-year update, Clin. Chem. 49 (4) (2003) 624–633.

[16] J.A. Hirsch, T.M. Leslie-Mazwi, G.N. Nicola, R.M. Barr, J.A. Bello, W.D. Donovan, R. Tu, M.D. Alson, L. Manchikanti, Current procedural terminology; a primer, J. Neurointerv. Surg. 7 (4) (2015) 309–312.

[17] Observational Health Data Sciences and Informatics (OHDSI), OMOP common data model version 5.0, 2021, (accessed 13 December 2021) https://www.ohdsi.org/data-standardization/the-common-data-model/.

[18] Observational Health Data Sciences and Informatics (OHDSI), Mapping of concepts, 2022, (accessed 8 April 2022) https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:mapping.

[19] K.R. Stratton, C.J. Howe, R.B. Johnston, et al., Adverse Events Associated with Childhood Vaccines: Evidence Bearing on Causality, Natl Academy Pr, 1994.

[20] N. Choi, I.-Y. Song, H. Han, A survey on ontology mapping, ACM Sigmod. Record 35 (3) (2006) 34–41.

[21] S.M. Falconer, N.F. Noy, M.-A.D. Storey, Ontology mapping-A user survey, in: Proceedings of the Workshop on Ontology Matching (OM2007) At ISWC/ASWC2007, Busan, South Korea, 2007, pp. 113–125.

[22] Y.K. Hooi, M.F. Hassan, A.M. Shariff, A survey on ontology mapping techniques, Adv. Comput. Sci. Appl. 279 (2014) 829–836.

[23] P. Ochieng, S. Kyanda, Large-scale ontology matching: State-of-the-art analysis, ACM Comput. Surv. 51 (4) (2019) 1–35.

[24] D. Oliveira, C. Pesquita, Improving the interoperability of biomedical ontologies with compound alignments, J. Biomed. Semant. 9 (1) (2018) 1–13.

[25] Y. Li, A. Black, G.A. Baltus, C. Cho, V. Chandrasekaran, E. Dasbach, L. Yao, Quality assessment of vaccine concepts in OMOP common data model, 2021, (accessed 11 May 2022) https://www.ohdsi.org/2020-global-symposium-showcase-26/.

[26] D. Kaduk, A. Black, Y. Li, L. Yao, Evaluation of vaccine concept mappings in OMOP vocabulary: a real-world database study, 2022, (accessed 11 May 2022) https://www.ohdsi.org/2021-global-symposium-showcase-11/.

[27] C. Reich, A. Ostropolets, Chapter 5 standardized vocabularies, 2022, (accessed 8 April 2022) https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html.

[28] National Library of Medicine, Rxnorm technical documentation, 2022, (accessed 15 February 2022) https://www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.

[29] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucl. Acids Res. 32 (Supplement_1) (2004) D267–D270.

[30] R. Abeysinghe, F. Zheng, E.W. Hinderer, H.N. Moseley, L. Cui, A lexical approach to identifying subtype inconsistencies in biomedical terminologies, in: Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 1982–1989.