Extending the OMOP Standard Vocabulary to Include Botanical Natural Products

Sanya B. Taneja, Mary F. Paine, Sandra L. Kane-Gill, Richard D. Boyce

Background

Botanical and other natural products have become increasingly popular as complementary health approaches. Up to 18% of adults report regular use of natural products[1] and up to 88% of adults report co-consuming natural products and drugs[2], raising concerns for adverse events occurring due to consumption of natural products and natural products with drugs. Spontaneous reporting systems such as the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS), which is widely used for drug safety surveillance, can be utilized for natural product pharmacovigilance with adverse event reports by identifying reports related to natural products and/or dietary supplements. Accordingly, we extended the OMOP vocabulary to include natural products, their synonyms, phytoconstituents, and name variations with the aim to standardize the natural product reports in spontaneous reporting systems.

In related work, Sharma et. al. used ten sources of natural products to identify strings in the FAERS database and identified 185,000 spontaneous reports involving natural products[3]. Sharma et. al. standardized natural product-related reports in the FAERS and Canada Vigilance Adverse Reaction (CVAR) databases by developing a thesaurus of plant names combined with a mapping and normalization approach that accommodated misspellings and variations in natural product names[4]. Results from both studies suggested that the lack of interoperability among natural product data sources, lack of coverage of synonyms, scientific names and common names, and ambiguity in natural product names are major challenges in investigating and monitoring natural products in FAERS[3,4]. The OMOP natural products vocabulary extension presented in this study aims to bridge this gap by creating a standardized resource for botanical natural products to facilitate natural products pharmacovigilance.

Methods

Natural products in adverse event reports may appear as scientific names, common names, market product names, brand names, and chemical (phyto)constituents. Name variations, abbreviations, and misspellings of the natural product names are also common. To add natural products to the OMOP standardized vocabulary, we first compiled a list of more than 700 scientific names (or Latin binomials) of natural products. We then extracted the common names, synonyms, and chemical constituents of the natural products from the Global Substance Registration System (G-SRS) using the public API and relationships in G-SRS[6]. As far as we know, G-SRS contains the largest number of synonyms and variations of natural products among different natural product data sources[3]. We also used the Natural Health Products Ingredients Database and the Licensed Natural Health Products Database (LNHPD) to identify other synonyms[7].

To insert natural products in the OMOP vocabulary tables, we designed SQL queries to create custom concept identifiers, class identifiers, and vocabulary identifier ('NAPDI') in the `concept` table. We then inserted the natural product scientific names, preferred name, synonyms, and constituents in the `concept` table. Each Latin binomial name was assigned a preferred common name either by manual mapping or from data extracted from G-SRS and LNHPD. We created the relationships for preferred common names ('napdi_pt', 'napdi_is_pt_of') and constituents ('napdi_has_const', 'napdi_is_const_of') in the `concept_relationship` table.

Including variations of natural product names in the vocabulary is important because spontaneous

reporting systems such as FAERS collect information using an online entry form where the drug or product name is entered via free text and may contain spelling errors. We sought to address this shortcoming by creating a reference set of natural product name variations for 65 top-selling and top-reported natural products from the FAERS dataset. The publicly available FAERS dataset (Quarter 1 of 2004 to Quarter 2 of 2021) was loaded after de-duplication and drug name standardization as described in Banda et. al.[5] We extracted all strings in the FAERS drug table with natural product scientific names, common names, and synonyms for 65 natural products from our OMOP standardized vocabulary. String matches were based on exact matches and Levenshtein distance matches. Two independent coders then manually reviewed all the FAERS strings to find relevant natural product names. All relevant natural product names from this set were then added to the `concept` table, and two new relationships ('napdi_spell_vr', 'napdi_is_spell_vr_of') for spelling variations were created in the `relationship` and `concept_relationship` tables.

Results

Insertion of natural product scientific names, preferred names, synonyms, and constituents resulted in 2,289 concepts in the `concept` table for 303 unique natural product Latin binomials. The addition of variations from the manually reviewed reference set from the FAERS database resulted in 2,772 name variations for 65 natural products.

Table 1 shows the `concept` table after querying the vocabulary to find the natural product green tea *(Camellia sinensis).* Tables 2 and 3 show the `concept` table for the query demonstrating the 'napdi_pt' and 'napdi_has_const' relationships. Table 4 shows the `concept` table with the first 10 name variations for green tea in the vocabulary from the FAERS reference set. Using the vocabulary extension for the 65 natural products to extract reports from the FAERS database, we were able to extract 47,601 reports matched to natural product names, 60,223 reports matched to natural product names including spelling variations, and 100,522 reports matched to natural product constituents.

Table 1: The concept table with green tea concepts in vocabulary.

concept	concept_name	domain_id	vocabulary_id	concept_class_id
-7000189	Black tea[Camellia sinensis]	NaPDI research	NAPDI	Green tea
-7000190	Green tea[Camellia sinensis]	NaPDI research	NAPDI	Green tea
-7000191	Oolong tea[Camellia sinensis]	NaPDI research	NAPDI	Green tea
-7000192	Tea[Camellia sinensis]	NaPDI research	NAPDI	Green tea
-7000193	White Tea[Camellia sinensis]	NaPDI research	NAPDI	Green tea
-7000293	Camellia sinensis[Camellia sinensis]	NaPDI research	NAPDI	Green tea

Table 2: The `concept` table with green tea preferred name.

concept_name	concept_id	preferred_name	concept_id
Camellia sinensis[Camellia sinensis]	-7001293	Green tea	-7001008

Table 3: The `concept` table with green tea constituents.

concept_name	concept_id	constituent_name	concept_id
Green tea	-7001008	CIANIDANOL	-7001622
Green tea	-7001008	EPICATECHIN	-7001895
Green tea	-7001008	EPICATECHIN GALLATE	-7002175
Green tea	-7001008	EPIGALLOCATECHIN	-7001785
Green tea	-7001008	EPIGALLOCATECHIN GALLATE	-7002248
Green tea	-7001008	GALLOCATECHIN	-7002061
Green tea	-7001008	GALLOCATECHIN GALLATE	-7001793

Table 4: The `concept` table with green tea name variations from the FAERS database in vocabulary.

concept_name	concept_id	name_variation	concept_id
Green tea	-7001008	GUARANA GREEN TEA	-7004112
Green tea	-7001008	CAMELLIA SINENSIS/PANAX GINSENG EXTRACT	-7004069
Green tea	-7001008	APPLE CIDER VINEGAR + GREEN TEA SUPPLEMENT	-7003800
Green tea	-7001008	ACV PLUS WITH GREEN TEA	-7003786
Green tea	-7001008	CINNAMON AND GREEN TEA	-7002953
Green tea	-7001008	VEREGEN GREEN TEA EXTRACT (CAMELLIA SINENSIS)	-7002716
Green tea	-7001008	VEREGEN GREEN TEA EXTRACT (CAMELLA SINENSIS)	-7002715
Green tea	-7001008	UNSPECIFIED GREEN TEA EXTRACT SUPPLEMENT	-7002714
Green tea	-7001008	TEA, GREEN (TEA, GREEN)	-7002713
Green tea	-7001008	TEA (GREEN TEA AND ICED TEA)	-7002712

Conclusion

Computational investigation of natural products requires standardized terminology, both for pharmacovigilance applications[4,8] and potential research with electronic health records[9,10]. Our OMOP vocabulary extension for natural products, including scientific names, preferred names, synonyms, constituents, and name variations enables natural products pharmacovigilance with FAERS and other spontaneous reporting systems (Canada Vigilance, VigiBase) to provide case-based evidence for adverse events related to natural products and natural product-drug interactions. We are currently expanding the vocabulary further to include all natural products, automated machine learning approaches to add variations for all natural products from FAERS, and external references to dietary supplement databases in the vocabulary. The vocabulary tables are publicly available for use by the community. All vocabulary

tables, code, and example queries are available at https://github.com/dbmi-pitt/np-terminology-imports.

Acknowledgments

This study was funded by the National Institutes of Health National Center for Complementary and Integrative Health Grant U54 AT008909.

References/Citations

1. Clarke TC, Black LI, Stussman BJ, Barnes PM, Nahin RL. Trends in the Use of Complementary Health Approaches Among Adults: United States, 2002–2012. Natl Health Stat Rep. 2015 Feb 10;(79):1–16.
2. Agbabiaka TB, Wider B, Watson LK, Goodman C. Concurrent Use of Prescription Drugs and Herbal Medicinal Products in Older Adults: A Systematic Review. Drugs Aging. 2017;34(12):891–905.
3. Sharma V, Sarkar IN. Identifying natural health product and dietary supplement information within adverse event reporting systems. In: Biocomputing 2018 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2018 [cited 2021 Feb 11]. p. 268–79. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789813235533_0025
4. Sharma V, Gelin LFF, Sarkar IN. Identifying Herbal Adverse Events From Spontaneous Reporting Systems Using Taxonomic Name Resolution Approach. Bioinforma Biol Insights. 2020 Jan 1;14:1177932220921350.
5. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data. 2016 Dec;3(1):160026.
6. Peryea T, Southall N, Miller M, Katzel D, Anderson N, Neyra J, et al. Global Substance Registration System: consistent scientific descriptions for substances related to health. Nucleic Acids Res. 2021 Jan 8;49(D1):D1179–85.
7. Canada H. Licensed Natural Health Products Database (LNHPD) [Internet]. 2007 [cited 2022 Jun 15]. Available from: https://www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/applications-submissions/product-licensing/licensed-natural-health-products-database.html
8. Vasilakes JA, Rizvi RF, Zhang J, Adam TJ, Zhang R. Detecting Signals of Dietary Supplement Adverse Events from the CFSAN Adverse Event Reporting System (CAERS). AMIA Summits Transl Sci Proc. 2019 May 6;2019:258–66.
9. Bompelli A, Li J, Xu Y, Wang N, Wang Y, Adam T, et al. Deep Learning Approach to Parse Eligibility Criteria in Dietary Supplements Clinical Trials Following OMOP Common Data Model [Internet]. Health Informatics; 2020 Sep [cited 2022 Apr 20]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.09.16.20196022
10. Fan Y, Zhou S, Li Y, Zhang R. Deep learning approaches for extracting adverse events and indications of dietary supplements from clinical text. J Am Med Inform Assoc. 2021 Mar 1;28(3):569–77.