PHOEBE 2.0: selecting the right concept sets for the right patients using lexical, semantic, and data-driven recommendations

Anna Ostropolets, George Hripcsak, Christopher Knoll, Patrick Ryan

Background

Reliable evidence generated from observational data can impact decision-making and improve patient outcomes but requires accurate phenotype algorithms (1). Our ability to create accurate and reliable phenotypes depends on two components: the ability to derive relevant and comprehensive sets of codes and the accuracy of the logic applied to those codes.

The challenge of selecting appropriate codes remains largely unsolved, especially when a study is conducted in a distributed network. As we observed previously, data sources are highly heterogeneous with a large proportion of the codes being uniquely used in one institution (2,3). Phenotype definitions and code sets are not readily transportable to other institutions and are characterized by lower performance when applied to other data sources (4).

Previously, we used the OHDSI Network to create a knowledge base of concepts and their utilization and use it in a concept recommender system – PHOEBE (PHenotype Observed Entity Baseline Endorser). Here, we describe the current achievements and enhancements to the system.

Methods

PHOEBE 1.0

The detailed design of PHOEBE 1.0 is described elsewhere (5). Briefly, 272 billion records from 22 participating data sources were summarized to obtain aggregated frequency estimates for each concept and all its descendants, which were then used to pre-compute a set of recommended terms for all standard concepts in the OMOP Standardized Vocabularies. First, string matching (for concepts and their synonyms) was used to select all lexically similar concepts for each standard concept. Then, semantically similar concepts were added by selecting the most proximal concepts within an ontology and through the crosswalks between the adjacent ontologies. These included both the concepts proximal and distal to the common ancestor as well as those belonging to a different hierarchy sub-trees (no common ancestor).

The resulting data set is used in PHOEBE, available as an R Shiny application (6).

PHOEBE 2.0

Based on the experience with PHOEBE, we identified two main areas of development: a) recommendation enhancement and expansion and b) application performance improvement.

PHOEBE's recommendations are based on a lexical match and ontological proximity. To identify relevant concepts that are not lexically similar, we enhance PHOEBE search by adding a data-driven approach that leverages patient context. Such an approach with modification can be used for all domains, but this abstract will focus on the condition domain.

A proposed approach uses all standard condition codes in participating data sources. To create concept representations, we select the first occurrence of a concept in all patients, so that there is at least 730 days of observation prior to the concept occurrence. For each concept, we then create a set of features: demographics (age and sex on the index date), presentation (condition codes within 30 days prior), treatment (drug and procedure codes within 30 days after the index date) and prognosis (a visit associated with the occurrence and death within 30 days prior).

We then calculate a cosine similarity matrix for all concept pairs based on each concept's feature separately and average the cosine similarity score across all features. This precomputed set will be incorporated into PHOEBE, with the recommendations that have a score of 0.80 or higher labeled "Recommended via patient context." In Results, we illustrate an added benefit of recommendations using an example of three concepts: Type 1 diabetes mellitus, Type 2 diabetes mellitus, and Angina pectoris.

Results

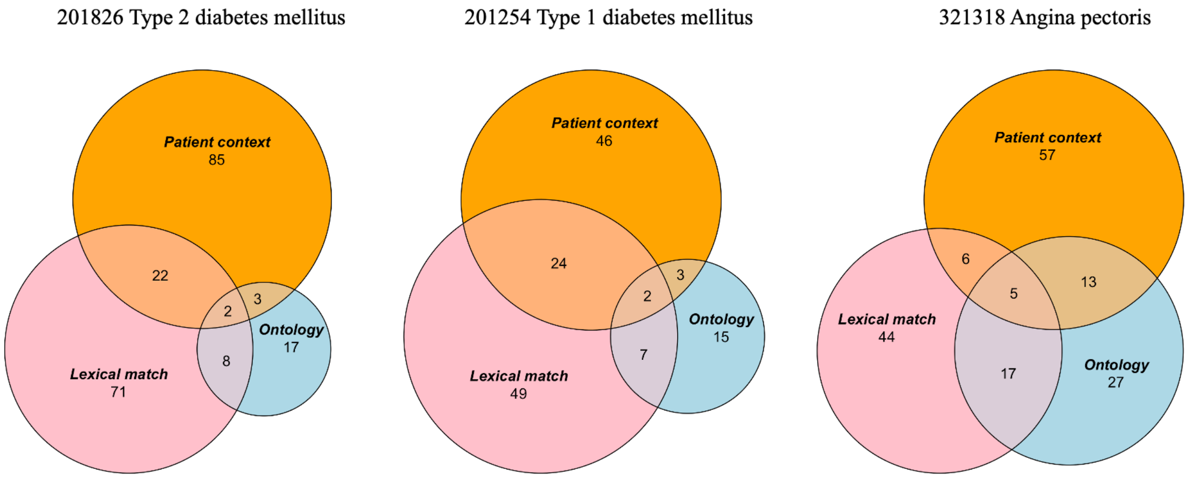
PHOEBE 1.0

Over the past two years, PHOEBE was used in many OHDSI network studies, both clinical (7–11) and methodological (12,13), which were published in high-impact journals like BMJ, Nature Communications, Lancet Rheumatology, and Rheumatology. PHOEBE has been continuously used for new studies such as LEGEND for Type 2 Diabetes (14).

Coupled with the tools for examining patient cohorts (Cohort Diagnostics), PHOEBE enabled systematic phenotype development and refinement across the network. At the same time, high demand highlighted a need for fast processing and retrieval of recommendations for multiple concurrent users.

PHOEBE 2.0

While using PHOEBE helps identify more patients while preserving positive predictive value of an algorithm and capturing patients early on in the course of the disease (5), adding data-driven recommendations may increase its value. Figure 1 illustrates the overlap between the current recommendations (through lexical match and ontology proximity) and proposed data-driven recommendations. We observe a substantial increase in the number of recommendations with adding patient context-generated recommendations (97%, 72%, and 70% respectively).



201826 Type 2 diabetes mellitus

201254 Type 1 diabetes mellitus

321318 Angina pectoris

Figure 1. Recommendations for type 1 diabetes, type 2 diabetes and angina pectoris, stratified by the approach used to derive them: through lexical match, ontology (PHOEBE 1.0) and patient context (PHOEBE 2.0)

Only a few concepts are suggested by all three approaches. For example, for angina all three recommend 'Angina co-occurrent and due to coronary arteriosclerosis' or 'Prinzmetal angina', which are descendants of the original concept, are lexically similar and appear in patients with similar demographics, treatment, and prognosis. Interestingly, the latter concept has a relatively high overall similarity score (Table 1) with high cosine similarity scores for demographics, treatment and prognosis components but a low score for the presentation component, which is clinically plausible (15).

More concepts are found both through the patient context and lexical match, such as 'Coronary artery spasm' or 'Peripheral circulatory disorder due to type 2 diabetes mellitus' as they appear similar in the data and lexically. On contrary, there are relevant concepts that are only captured through patient context, such as 'Acute coronary artery occlusion not resulting in myocardial infarction' and 'Chronic ischemic heart disease' recommended for Angina pectoris. These two concepts alone account for 50,260,576 records in the OHDSI Network and can be potentially missed in concept set expressions.

Aside from adding patient context, the existing lexical search can be improved to enable partial fuzzy matching using both concept names and synonym names.

Conclusion

Together with the OHDSI Network, we developed PHOEBE - a recommender system that facilitates phenotype development standardization and comprehensive concept set creation.
Enhancing PHOEBE by adding data-driven recommendations learnt from patient context provides researchers with additional relevant concepts that otherwise may be missed in the concept set expression.

Table 1. Examples of concepts recommended by PHOEBE 2.0

	Similarity score					Included in PHOBE 1.0
Concept	Overall	Demographic	Presentation	Treatment	Prognosis	
Concepts recommended for the concept 201826 Type 2 diabetes mellitus						
4032787 Hyperosmolarity	0.94	1.00	0.78	0.96	1.00	No
4044391 Neuropathy due to diabetes mellitus	0.89	1.00	0.61	0.95	1.00	No
439233 Poisoning by antidiabetic agent	0.87	0.99	0.67	0.84	0.97	No
40482801 Type II diabetes mellitus uncontrolled	0.96	1.00	0.87	0.96	0.99	Yes
443729 Peripheral circulatory disorder due to type 2 diabetes mellitus	0.92	0.99	0.71	0.97	1.00	Yes
201820 Diabetes mellitus	0.89	0.99	0.66	0.92	1.00	Yes
Concepts recommended for the concept 201254 Type 1 diabetes mellitus						
43021246 Complication associated with insulin pump	0.89	0.95	0.72	0.91	0.99	No
4096804 Drug-induced hypoglycemia without coma	0.88	0.98	0.59	0.96	0.99	No
443735 Coma due to diabetes mellitus	0.87	0.99	0.54	0.98	0.99	No
37016348 Hyperglycemia due to type 1 diabetes mellitus	0.98	0.99	0.94	0.98	1.00	Yes
439770 Ketoacidosis due to type 1 diabetes mellitus	0.93	0.99	0.84	0.91	0.99	Yes
4227210 Retinopathy due to type 1 diabetes mellitus	0.89	0.99	0.66	0.92	1.00	Yes
Concepts recommended for the concept 321318 Angina pectoris						
315286 Chronic ischemic heart disease	0.89	0.99	0.62	0.95	1.00	No
44784623 Acute coronary artery occlusion not resulting in myocardial infarction	0.88	0.99	0.64	0.89	1.00	No
36712779 Chronic total occlusion of coronary artery	0.87	0.94	0.7	0.86	0.99	No
36712983 Angina co-occurrent and due to coronary	0.91	0.98	0.7	0.97	1.00	Yes
4127089 Coronary artery spasm	0.89	1.00	0.61	0.96	1.00	Yes
315830 Prinzmetal angina	0.89	0.97	0.62	0.98	1.00	Yes

References

1. Dreyer NA, Tunis SR, Berger M, Ollendorf D, Mattox P, Gliklich R. Why Observational Studies Should Be Among The Tools Used In Comparative Effectiveness Research. Health Affairs. 2010 Oct;29(10):1818–25.
2. Ostropolets A. Concept Heterogeneity in the OHDSI Network [Internet]. 2019; OHDSI Symposium. Available from: https://www.ohdsi.org/2019-us-symposium-showcase-19/
3. Ostropolets A. Characterizing database granularity using SNOMED-CT hierarchy. In: AMIA 2020 Proceedings. 2020.
4. Chen Y. Opportunities and Challenges in Data-Driven Healthcare Research [Internet]. MEDICINE & PHARMACOLOGY; 2018 Jun [cited 2021 Dec 3]. Available from: http://www.preprints.org/manuscript/201806.0137/v1
5. Ostropolets A. Phenotype observed entity baseline endorsements (PHOEBE) - Recommender system for concept selection in phenotype algorithm development [Internet]. American Medical Informatics Association Annual Symposium 2021; 2021 Oct 30. Available from: https://www.proceedings.com/content/061/061315webtoc.pdf
6. PHOEBE [Internet]. [cited 2022 Feb 6]. Available from: https://data.ohdsi.org/PHOEBE/
7. Prieto-Alhambra D, Kostka K, Duarte-Salles T, Prats-Uribe A, Sena A, Pistillo A, et al. Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. 2021;
8. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MTF, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nature communications. 2020;11(1):1–11.
9. Reyes C, Pistillo A, Fernández-Bertolín S, Recalde M, Roel E, Puente D, et al. Characteristics and outcomes of patients with COVID-19 with and without prevalent hypertension: a multinational cohort study. BMJ Open. 2021 Dec;11(12):e057632.
10. Tan EH, Sena AG, Prats-Uribe A, You SC, Ahmed WUR, Kostka K, et al. COVID-19 in patients with autoimmune diseases: characteristics and outcomes in a multinational network of cohorts across three countries. Rheumatology. 2021 Oct 9;60(SI):SI37–50.
11. Morales DR, Ostropolets A, Lai L, Sena A, Duvall S, Suchard M, et al. Characteristics and outcomes of COVID-19 patients with and without asthma from the United States, South Korea, and Europe. Journal of Asthma. 2022 Feb 11;1–11.
12. Reps J, Kim C, Williams R, Markus A, Yang C, Salles TD, et al. Can we trust the prediction model? Demonstrating the importance of external validation by investigating the COVID-19 Vulnerability (C-19) Index across an international network of observational healthcare datasets. 2020;
13. Ostropolets A, Li X, Makadia R, Rao G, Rijnbeek PR, Duarte-Salles T, et al. Empirical evaluation of the sensitivity of background incidence rate characterization for adverse events across an international observational data network [Internet]. Health Informatics; 2021 Jul [cited 2021 Jul 15]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.06.27.21258701
14. Khera R, Schuemie MJ, Lu Y, Ostropolets A, Chen R, Hripcsak G, et al. Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies. BMJ Open. 2022 Jun;12(6):e057977.
15. Yeh BK, Rogers CM. Prinzmetal Angina. Chest. 1970 Oct;58(4):396–8.