Name:	Matthew Levine
Affiliation:	Department of Biomedical Informatics, Columbia University
Email:	mel2193@cumc.columbia.edu
Presentation type:	Poster

Terminology Information Loss and Gain

Matthew E. Levine, BA[1,2], George Hripcsak, M.D., M.S.[1,2]
[1]Department of Biomedical Informatics, Columbia University, New York, New York, USA
[2]Observational Health Data Sciences and Informatics, New York, New York, USA

Abstract

The power of transforming clinical databases into OMOP CDM format is demonstrated by the many successful studies performed by the OHDSI consortium. Yet, there is an inevitable information loss when translating an existing database into OMOP CDM, which must be weighed against the advantages of OHDSI participation and potential information gain that can occur during translation. We propose an analytical structure for quantifying the potential information loss and gain of a translated clinical database. As an example of the effects of information loss and gain, we will compare results from eMERGE phenotype cohorts in our source database with results using translated cohort definitions on an OMOP CDMv4 version of the database.

Introduction

The OHDSI (Observational Health Data Sciences and Informatics) community has demonstrated the tremendous potential of unifying electronic health records (EHRs) under a common data format and purpose [1,2]. The resources of OHDSI provide significant incentives for stakeholders to translate their clinical databases into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). In addition, many of the OMOP standard terminologies, like SNOMED, can facilitate concept set definitions better than commonly used billing codes, like ICD9. Nevertheless, terminology mapping involves some degree of information loss, which can present a perceived barrier to potential members of the OHDSI community.

OHDSI stakeholders have begun to explore the effects of information loss and gain of terminology mappings, and we aim to formalize and centralize these analyses by developing open-source analytical tools for identifying multiple and missed mappings between standard and non-standard terminologies in the OMOP CDMv5 vocabulary. Moreover, we wish to understand how these transformations affect the information in real patient databases, and how they can manifest as information losses or gains in phenotyping study cohorts, like those from eMERGE [3,4].

Here, we present preliminary work towards this goal that focuses on mappings between ICD9 and SNOMED diagnosis codes, and how their relationships affect phenotype cohort results. To study this, we: 1) quantify single, multiple, and missed mappings between ICD9 and SNOMED diagnosis codes in the OMOP CDMv5 vocabulary, 2) quantify the prevalence of mappings that are not 1-to-1 in our hospital database, and 3) compare resultant cohorts from ICD9-based eMERGE phenotype definitions applied to the source data and SNOMED-translated eMERGE phenotype definitions applied to the OMOP CDMv5 database.

Methods

1. Terminology mapping in the OMOP CDMv5 vocabulary
In order to quantify the potential for information loss when mapping to OMOP CDMv5, we developed SQL queries to the OMOP CDMv5 concept table to answer the following questions:

1. How many ICD9 codes map to exactly n SNOMED codes, for n=0,1,2,3…
2. How many SNOMED codes map to exactly n ICD9 codes, for n=0,1,2,3…
3. How many ICD9 codes have n cousins that also map to its SNOMED mapping, for n=0,1,2,3…

2. Prevalence of non-1-to-1 diagnosis code mappings in a clinical database
In order to characterize the frequencies at which potentially lossy diagnosis codes (non-1-to-1 mappings) appear in a clinical database, we use an OMOP CDMv5 instance of the Columbia University—NewYork-Presbyterian clinical data warehouse, which contains records for over 3 million patients, and answer the following queries:

1. What proportion of condition occurrences have an ICD9 source code that has exacly n SNOMED mappings, for n=0,1,2,3…
2. What proportion of patients have at least one condition occurrence with an ICD9 source code that has exactly n SNOMED mappings, for n=0,1,2,3…

3. Prevalence of non-1-to-1 diagnosis code mappings in eMERGE phenotype concept sets
In order to examine the potential for cohort studies to be affected by lossy diagnosis code mappings, we create OMOP standard mappings of the eMERGE phenotype concept sets hosted by PheKB [5], and query:

1. What proportion of concept sets from eMERGE phenotype definitions on PheKB contain ICD9 codes with exactly n SNOMED mappings, for n=0,1,2,3…

4. Effect of non-1-to-1 diagnosis code mappings on eMERGE phenotype cohorts
Finally, we will examine the similarity of patient cohorts returned using eMERGE phenotype definitions on our source database with patient cohorts returned using OMOP-translated definitions on our OHDSI database. We will compare the cohort sizes the degree of patient overlaps. We also hope to leverage chart reviews previously performed for these definitions on patients in our databse to compare sensitivity and specificity rates for our source cohort and our OHDSI cohort. These analyses, supported by additional chart reviews, will allow us to evaluate both information losses and information gains, depending on the mechanics of the terminology mappings.

Results

We report preliminary results that assess non 1-to-1 mappings between ICD9 and SNOMED codes in the OMOP CDMv5 vocabulary and their prevalence in a clinical database. Results regarding the prevalence of these mappings in eMERGE phenotype concept sets, and their effect on patient cohort definitions are forthcoming and will be presented in the poster.

Table 1 reports the frequencies of loss-prone ICD9-SNOMED mappings in our OMOP CDMv5 database. We note that approximately 2% of ICD codes in the vocabulary have more than 1 SNOMED mapping, and that approximately 4% of patients with ICD9 codes in our database had a condition occurrence source code with more than 1 SNOMED mapping.

N (Number of mappings)	Fraction of ICD codes with N SNOMED mappings	Prevalence of ICD codes with N SNOMED mappings among all ICD codes	Prevalence of patients with at least 1 ICD code with N SNOMED mappings among all patients with ICD codes
0	0.6%	1.0%	0%
1	97.5%	97.8%	95.4%
2	1.8%	1.2%	4.4%
3	0.1%	0.03%	0.1%

Table 1.

Discussion

Potential sources of information loss in non-1-to-1 ICD9 to SNOMED mappings appear not only in the CDM, but are also propagated through patients in our database. These results beget further follow up in order to understand how they affect cohort studies, and we plan to share our preliminary findings on this topic at the OHDSI Symposium

Acknowledgment

This work was funded by National Library of Medicine grant R01 LM006910.

References

[1] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, Stud Health Technol Inform. 216 (2015) 574–578.

[2] G. Hripcsak, P.B. Ryan, J.D. Duke, N.H. Shah, R.W. Park, V. Huser, M.A. Suchard, M.J. Schuemie, F.J. DeFalco, A. Perotte, J.M. Banda, C.G. Reich, L.M. Schilling, M.E. Matheny, D. Meeker, N. Pratt, D. Madigan, Characterizing treatment pathways at scale using the OHDSI network, PNAS. 113 (2016) 7329–7336. doi:10.1073/pnas.1510502113.

[3] C.A. McCarty, R.L. Chisholm, C.G. Chute, I.J. Kullo, G.P. Jarvik, E.B. Larson, R. Li, D.R. Masys, M.D. Ritchie, D.M. Roden, others, The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Medical Genomics. 4 (2011) 13.

[4] J. Pathak, J. Wang, S. Kashyap, M. Basford, R. Li, D.R. Masys, C.G. Chute, Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience, Journal of the American Medical Informatics Association. 18 (2011) 376–386. doi:10.1136/amiajnl-2010-000061.

[5] J.C. Kirby, P. Speltz, L.V. Rasmussen, M. Basford, O. Gottesman, P.L. Peissig, J.A. Pacheco, G. Tromp, J. Pathak, D.S. Carrell, S.B. Ellis, T. Lingren, W.K. Thompson, G. Savova, J. Haines, D.M. Roden, P.A. Harris, J.C. Denny, PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J Am Med Inform Assoc. 23 (2016) 1046–1052. doi:10.1093/jamia/ocv202.