Standardizing Clinical Diagnoses: Evaluating Alternate Terminology Selection

Evanette K. Burrows, MPH[1], Hanieh Razzaghi, MPH[1], Levon Utidjian, MD, MBI[1], L. Charles Bailey, MD, PhD[1]
[1]Children's Hospital of Philadelphia, Philadelphia, PA

Abstract

In most electronic health record (EHR) systems, clinicians record diagnoses using interface terminologies, such as Intelligent Medical Objects (IMO). When extracting data from EHRs for collaborative research, local codes are often transformed to standard terminologies for consistent analyses despite the potential for loss of fidelity. EHR diagnosis codes may be standardized directly during the Extract-Transform-Load (ETL) process to the "Meaningful Use" clinical data exchange standard, SNOMED-CT, or to the International Classification of Diseases (ICD) terminologies commonly used for billing. We examined the performance of ETL standardization via the direct IMO mapping to SNOMED-CT, and via IMO mapping to ICD-9-CM or ICD-10-CM followed by UMLS mapping to SNOMED-CT. We found that for both ICD-9-CM and ICD-10-CM, only 24-27% of diagnosis codes map to the same SNOMED-CT code selected by the direct IMO-SNOMED crosswalk. We identified that differences in mapping lead to loss in the granularity and laterality of the initial diagnosis.

Introduction

In most electronic health record (EHR) systems, clinicians record discrete diagnoses using local or commercial interface terminologies, such as Intelligent Medical Objects (IMO®)[1]. These interface technologies provide internal mappings to various external terminologies needed for clinical or business data exchange, such as the International Classification of Diseases (ICD), 9th Revision, Clinical Modification ICD-9-CM [2], ICD-10-CM[3], and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)[4]. These terminologies are international standards that help to structure information consistently across practices for disease and mortality reporting. The mappings from EHR system terminologies to external standards are proprietary and not available for public use.

Multisite research increasingly makes use of common data models to facilitate data sharing, reuse of analyses across data sources, and distributed analytics. When extracting data for collaborative research from EHRs to common data models, codes are typically be transformed to standard vocabularies to support interoperability and integration[5,6]. Data is often stored in different data structures at different institutions, and even when the same EHR vendor is used, there may be multiple vocabularies in use to represent the same medical condition or clinical outcomes. Further, the Extract-Transform-Load (ETL) process may not be standardized across institutions, which leads to potential information loss that is neither quantified nor represented in the standardized model. This makes analysis of the data challenging and complicated. Transforming to a standard terminology partially addresses these problems to improve interoperability of data from different datasets[7]. In addition, use of standard terminologies may facilitate the research process by allowing for more widely understood and reusable code set generation.

The National Pediatric Learning Health System (PEDSnet) is a PCORnet Clinical Data Research Network (CDRN) [5] comprising eight of the nation's largest children's hospitals [8,9]. The goal of this network is to support a wide range of pediatric research by using a common data model for extensive observational data on over 6 million children. PEDSnet includes patient data such as clinical diagnoses (visit diagnoses and problem list), procedures, medications, labs, vitals, demographics, and visit metadata.

PEDSnet data is standardized to both Observational Health Data Science and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP)[10] and PCORnet[5] common data models. The OHDSI OMOP common data model and standard vocabulary[11] includes precomputed relationships that are established between various diagnosis terminologies (e.g. ICD-9-CM/ICD-10-CM) and standardized SNOMED-CT codes, based on Unified Medical Language System (UMLS) mappings. All diagnosis codes are transformed to SNOMED-CT codes prior to aggregation. SNOMED-CT is a relatively complete ontology created by the College of American Pathologists (CAPS) to support clinical decision making and analytics[12]. However, in this transformation, there is the potential

for loss of fidelity due to semantic differences between the EHR terminology, whether locally-established or commercial, and the target terminology. In some cases, semantic mismatch may arise from differences in terminology size and granularity. For example, the current version of IMO's diagnostic terminology contains over 800,000 terms, while in the OHDSI OMOP standard vocabulary (version v5.3 09-APR-18), there are 297,286 SNOMED-CT codes of a type likely used to describe a diagnosis, 109,389 ICD-10-CM Codes and 18,672 ICD-9-CM Codes. Other causes for semantic mismatch include differences in terminology structure (*e.g.* level of pre-coordination) or semantic focus (*e.g.* disease biology vs resource utilization). The aim of this study was to examine the performance in standardizing local diagnosis codes to SNOMED-CT via IMO's direct mapping to SNOMED-CT, or via the IMO-ICD mapping often done within a health system for claims generation.

Methods

Data Standardization. Figure 1 summarizes the potential ways to standardize source diagnosis codes during ETL. Path A represents the native source system mapping from the IMO code to one or more SNOMED-CT codes. Path B represents the translation from IMO through an ICD code (ICD-9-CM or ICD-10-CM), which was mapped in a second step using OHDSI OMOP/UMLS crosswalk in order to map to a SNOMED-CT Code.

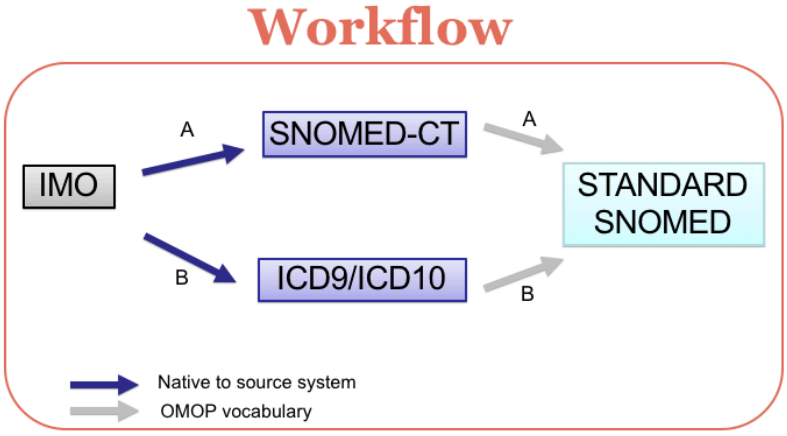


Figure 1. Mappings Path from IMO through SNOMED-CT or ICD9/10-CM to standard SNOMED Code.

Data Extraction. We extracted data for 1,115,772 patients at Children's Hospital of Philadelphia (CHOP) (1) received care at a physical visit and (2) had at least one diagnosis code in their record on or after January 1,2009 using an automated Python based tool. For each diagnosis code entered as an IMO term in the EHR, the IMO-generated mappings to SNOMED-CT, ICD-9-CM, and ICD-10-CM were extracted. We first chose a random subset of patients who had emergency department and outpatient visits between August and November 2015, an interval chosen to span the ICD-9-CM to ICD-10-CM conversion in clinical operations. A second random subset was chosen of patients who had visits between January and April of 2018, to determine whether any phenomena seen in the first time interval were associated specifically with the conversion in clinical operations.

Mapping Assessment. We examined the presence of the IMO-derived codes in the OHDSI vocabulary (version v5.3 09-APR-18). In addition, the rate of mapping to an OHDSI-designated standard SNOMED code through both the native SNOMED mapping (path A) and the translation through an ICD code using the OHDSI/UMLS crosswalk (path B) was assessed using structured query language (SQL) to examine exclusive matches. The difference between ETL paths was summarized as a "mapping identity" proportion, that is, the proportion of final codes in the OHDSI OMOP CDM that were the same via path B as those obtained via path A. We evaluated and termed unmappable codes as "loss" in mapping at various points in each path.

Results

For the earlier sample analysis (August-November 2015), we retrieved 42,454 unique IMO diagnosis terms from 625,204 visits for 251,775 patients (Table 1).

Table 1. Demographics for visit subsets.

	August – November 2015	**January – April 2018**
Number of Patients	251,775	257,026
Number of Visits	625,204	683,166
ED	*34,969*	44,332
Outpatient	*590,235*	593,184
Number of Diagnosis Codes	1,609,988	1,985,993
Unique Diagnoses Codes	42,454	41,446
ICD 9 Codes (%)	44%	N/A
ICD 10 Codes (%)	56%	100%

As shown in Table 2, the internal IMO-SNOMED crosswalk produced the lowest percentage (0.028%) of loss in both inclusion of the source code in the OHDSI OMOP vocabulary and in mapping to a standardized SNOMED-CT code.

Table 2. Comparison of mappings for earlier time interval

IMO Target Terminology	IMO Mapped Codes (N)	IMO Mapped Codes Present in OMOP (N)	Loss in IMO mapping (%)	IMO Mapped Codes with OMOP Standard Code (N)	Loss in OMOP mapping (%)	Cumulative Loss (%)	Mapping identity (%)
SNOMED	42,454	42,444	0.024	42,442	0.0047	0.028	NA
ICD-10-CM	32,108	31,406	2.18	31,389	0.054	2.24	27
ICD-9-CM	29,881	29,746	0.45	29,691	0.18	0.64	24

ICD-10 codes had the highest percentage of loss in the OHDSI OMOP vocabulary (2.24%). This finding was replicated in the later time interval as well (Table 4). To better understand the reasons for mapping failure, we examined the most frequently used mapped and unmapped codes in our data set. Unmapped codes from IMO to any external terminology most often correspond to growth and developmental diagnoses, adverse events, and congenital malformations (Table 3), though the range is wide.

Table 3. Most common mappable and unmappable IMO diagnosis codes.

Mapped IMO Code		**Unmapped to any external terminology**	
Outpatient Visits	n (% visits)	***Outpatient Visits***	n (% visits)
WCC (well child check)	100,128 (16.9)	Other known or suspected fetal abnormality, not elsewhere classified, affecting management of mother, antepartum condition or complication	1,678 (0.2)
Need for prophylactic vaccination and inoculation against influenza	63,307 (10.7)	Deceased-donor kidney transplant recipient	442 (0.07)
Pharyngitis	11,304 (1.9)	Optic glioma	184 (0.03)

URI (upper respiratory infection)	8,848 (1.5)	Glioma of brain	129 (0.02)
Mild intermittent asthma, uncomplicated	7,990 (1.4)	Fetal cardiac anomaly affecting pregnancy, antepartum	98 (0.02)
Emergency Visits		***Emergency Visits***	n (% ED visits)
Fever in pediatric patient	3,217 (9.2)	Neonatal thrush	14 (0.04)
Viral syndrome	2,944 (8.4)	Reflux	13 (0.04)
Fall, initial encounter	2,368 (6.8)	Mononucleosis	8 (0.02)
Viral URI	2,146 (6.1)	Abdominal distention	7 (0.02)
Cough	1,597 (4.6)	CLABSI (central line-associated bloodstream infection)	7 (0.02)

As shown in Table 3, codes that do not map to any external terminology vary and are recorded for 10,704 patients (5%) of the population in the subsets analyzed primarily, and 124,097 patients (14%) of the population as a whole. Overall, for both ICD-9-CM and ICD-10-CM, only 24-27% of diagnosis codes map to the same SNOMED-CT code selected by the direct IMO-SNOMED crosswalk.

For the more recent sample analysis (January-April 2018), we retrieved 41,446 unique IMO diagnosis terms from 683,166 visits for 257,026 patients (Table 1). Because time was well after the ICD conversion, ICD-9-CM codes were not analyzed. However, similar trends were observed, with a marginal decrease in relative mapping fidelity (Table 4). Once again, only 24% of diagnosis codes map to the same SNOMED-CT code by the direct IMO-SNOMED crosswalk and the IMO-ICD10-SNOMED crosswalk.

Table 4. Comparison of Mappings for later subset

IMO Target Terminology	IMO Mapped Codes (N)	IMO Mapped Codes Present in OMOP (N)	Loss in IMO mapping (%)	IMO Mapped Codes with OMOP Standard Code (N)	Loss in OMOP mapping (%)	Cumulative Loss (%)	Mapping identity (%)
SNOMED	41,446	41,349	0.23	41,347	0.0048	0.24	NA
ICD-10-CM	41,446	41,349	0.23	41,008	0.82	1.06	24

To better understand the differences in mapping (Table 5), we randomly selected and manually compared 200 discordant mappings via ICD-9-CM and 200 via ICD-10-CM to the IMO-SNOMED mappings by manual review by one expert (CB). In most cases, the two approaches yield codes at different levels of specificity in the SNOMED-CT hierarchy. For ICD-9-CM, the IMO-SNOMED mapping provides a mapping that preserves the intended diagnosis to the original term 66% of the time, versus 68% when comparing the ICD-10-CM mapping to the IMO-SNOMED mapping.

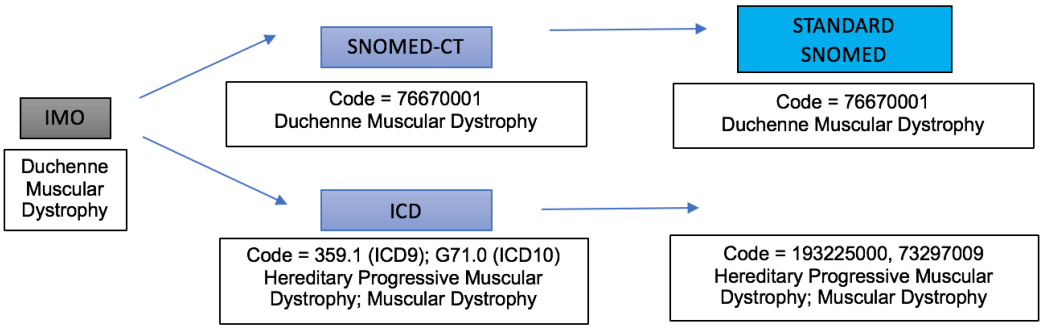
Table 5. Examples of Mapping Differences (IMO-SNOMED, ICD10 and ICD9)

IMO Description	**Direct SNOMED**	**Via ICD**	**Preferred Mapping**
Numbness of Toes	Numbness of toe	Altered Sensation of Skin	Direct SNOMED
Cerebellar ataxia/dyskinesia	Cerebellar Disorder	Cerebellar Ataxia	Via ICD
Choking episode	Choking sensation	Finding of head and neck region	Direct SNOMED

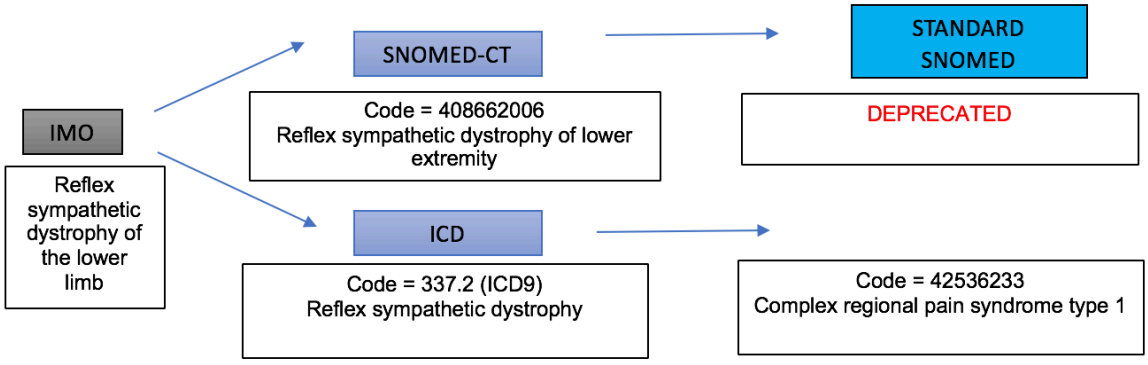
Intestinal malrotation	Congenital malrotation of intestine	Congenital anomaly of fixation of intestine	Equivalent
Genetic disease carrier status testing	Genetic finding	Genetic disorder carrier	Via ICD
Duchenne muscular dystrophy	Duchenne muscular dystrophy	Hereditary progressive muscular dystrophy	Direct SNOMED

During our review of the 400 terms, we observed three different recurring reasons for loss of fidelity: (1) Mappings in the crosswalk that result in a loss of granularity, (2) Deprecated Mappings and (3) Differences in how laterality is preserved. We also observed incorrect mappings presumably due to human error and interpretation that were a result of incorrect mappings in the crosswalk that result in a completely different diagnosis and alternate clinical terms for the same diagnosis. Below are examples of each case:

Loss of granularity: Duchenne Muscular Dystrophy (DMD) is a severe type of muscular dystrophy characterized by changes/mutations in the DMD gene and the absence of dystrophin. SNOMED-CT is the only standard terminology that fully represents the DMD diagnosis. DMD does not have a specific ICD code and therefore transforming to the standard SNOMED code through an ICD crosswalk, be it ICD-9-CM or ICD-10-CM, results in a loss of granularity, leaving an investigator unable to differentiate DMD from other classifications of muscular dystrophy. In this particular scenario, 633 (0.06%) patients are affected.

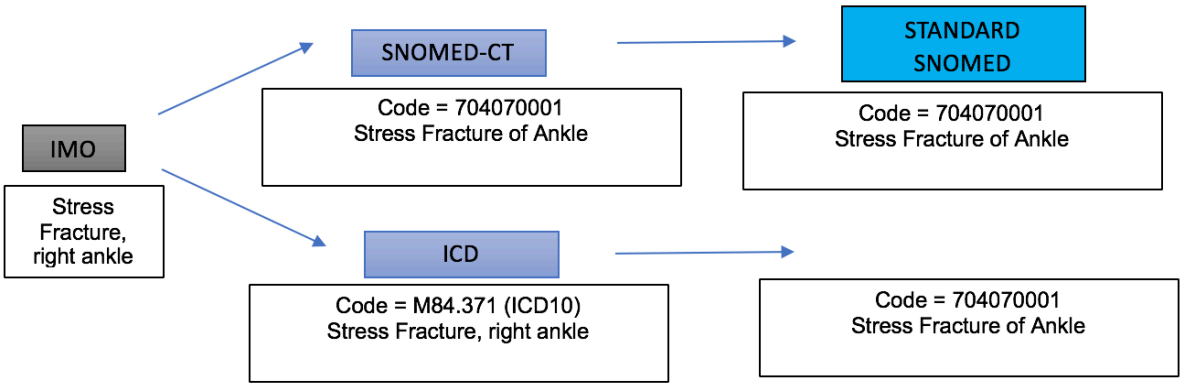


Deprecated or Invalid Mapping: The deprecation or remapping of codes can also lead to loss of fidelity for a diagnosis. In the case of the diagnosis of "reflex sympathetic dystrophy of the lower limb", the IMO-specified mapping is to a retired SNOMED-CT term, for which there is not a live mapping to a current SNOMED-CT term in the OHDSI vocabulary. The IMO-specified mapping to ICD-9-CM loses the anatomic specificity, and is in turn mapped to the SNOMED term "complex regional pain syndrome type 1". In this particular scenario, 886 (0.08%) patients are affected. This is important to highlight because while the current standard SNOMED code is deprecated, at the time of diagnosis the code was valid.

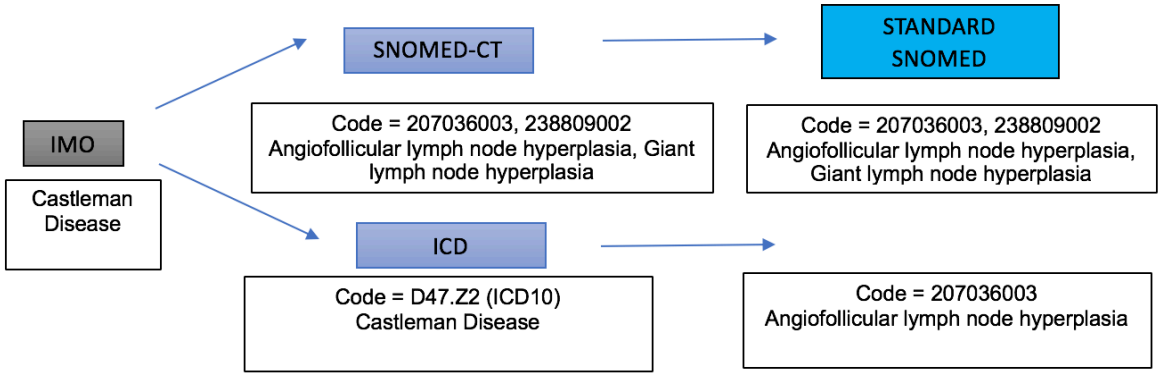


Precoordinated laterality in ICD-10: In some cases, the IMO-specified mapping to SNOMED-CT discards laterality preserved in the ICD-10-CM mapping, choosing a code from the international version rather than a US extension code that parallels available ICD-10-CM specificity. Where the goal is a 1:1 mapping that includes laterality, these

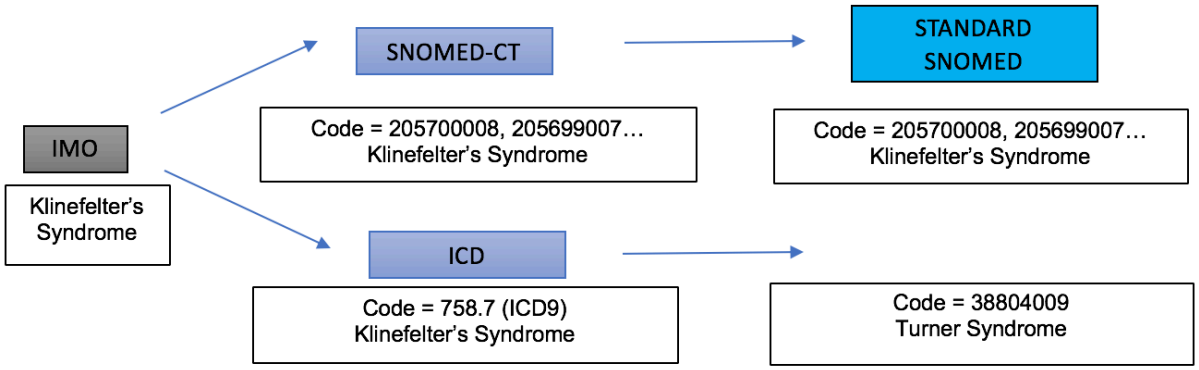
other terminologies are better adapted to the analysis than international SNOMED-CT. For example, for the diagnosis of "stress fracture, right ankle", one can retrieve a single code from IC D-10-CM or the US extension to SNOMED-CT. This is not true for international SNOMED-CT where both the "stress fracture of ankle" and another code must be used to capture the anatomic detail by post-coordination. For this particular term, 4 (0.001%) patients are affected. However, looking at other ICD-10-CM diagnosis terms that include laterality information, such as "left", "right", "upper", "lower", "top" and "bottom", where the mapping to SNOMED does not include this information, 38,471 (15%) patients are affected.



Recent Terms: Castleman Disease, or angiofollicular lymph node hyperplasia, has only recently acquired a specific ICD-10-CM term; editions of ICD-10-CM prior to 2016 as well as ICD-9-CM map to a nonspecific lymphadenopathy term. The up-to-date mapping is shown below. This case affects 15 (0.001%) patients.



Incorrect mappings: Klinefelter Syndrome is a chromosomal condition that affects male physical and cognitive development, resulting from the presence of two or more X chromosomes as well as a Y chromosome. Based on the figure below, Klinefelter Syndrome is appropriately represented in SNOMED-CT and both varieties of ICD. However, an error occurs in the OHDSI mapping of the ICD-9-CM term, where Klinefelter is mapped to Turner Syndrome, a chromosomal condition that affects female physical and cognitive development, wherein a female is partly or completely missing an X chromosome. This case affects 142 (0.01%) patients.



It is important to note for this case in particular that this loss of fidelity was likely the result of human error. However, this also sheds light on the nature of the current crosswalks between various vocabularies. Most of the mappings are not native to the standard vocabulary but are created and supported by various organizations, with varying levels of capacity. The task is very time-consuming and prone to error due to the complexity and abundance of codes.

Conclusion

Terminology standardization across time and across data sources is an important tool in the design of repeatable analyses. However, the path taken from local codes used at the point of care to community standard terminologies can significantly affect the degree to which standardized terms align with the concepts clinicians intended to express. This has potentially significant research implications for accurate identification of cases, diagnosis code-based phenotypes and outcomes of interest in an analysis.

When we examined the standardization of IMO codes to the OHDSI OMOP CDM via the direct IMO mapping to SNOMED-CT and its intermediate mapping to ICD, we found that only 24-27% of IMO codes map to the same SNOMED-CT term via both paths. We found that the internal IMO-SNOMED crosswalk produced the lowest percentage of loss (0.028%) in both inclusion of the source code in the OMOP vocabulary and in mapping to a standardized SNOMED-CT code whereas ICD-10 codes had the highest percentage of loss in the OHDSI vocabulary (2.24%). While we anticipated that there might have been differences between the 2015 and 2018 subsets due to different IMO terms being available around the ICD-9-CM to ICD-10-CM changeover, this was not the case. We observed 24% of codes mapping to the same SNOMED-CT codes from ICD-10-CM in the latter subset. Detailed manual review of a sample of discordant codes revealed a number of potential reasons for the observed differences such as mappings in the crosswalk that result in a loss of granularity, deprecated mappings, differences in how laterality is preserved, recent clinical terms and incorrect mappings.

In related work, other studies demonstrate the loss of fidelity in mapping. Matcho et al. [12], focused on ICD9-CM, SNOMED-CT and MedDRA mapping for medical product surveillance, found that only 46% of diagnosis codes in their study dataset had a 1:1 mapping (mapping fidelity as we described) to SNOMED-CT concepts when assessing completeness of mapping and required some additional grouping to achieve better mapping rates. Hripcsak et al. [13], found that knowledge engineering was required to improve performance of mapping of ICD-9-CM and ICD-10-CM where simple automated methods to generate concept sets had errors up to 10%. Further, Fung et al. [14], in investigating automated methods for translating ICD codes in clinical phenotype definitions found that all methods of mapping require some level of human validation.

When performing retrospective studies, it is important to consider the implications of using terminologies and mappings that are being updated and are dynamic. We thought it important to highlight deprecated and incorrect mappings to relay this. In addition to such, mappings could yield varying levels of specificity. In our manual review, we observed that for ICD-9-CM, the IMO-SNOMED mapping provides a mapping that preserves the intended

diagnosis to the original term 66% of the time, versus 68% when comparing the ICD-10-CM mapping to the IMO-SNOMED mapping.

This study reflects the coding practice at a single institution and may be biased of its clinical practice. However, the high proportion of existing mappings to ICD-9-CM and ICD-10-CM vocabularies implies that standard business practice of constraining clinician input to terms that will yield billable codes would not in itself be expected to alter the results. Further, review of additional discordant codes might reveal new reasons for the difference in mapping, with potentially different impact on analyses. As Rijnbeek, [15] suggests, as we assess common data models we must take into consideration their impact on the study results.

Additional work to assess the quality of both IMO and OMOP terminology mappings used here, as well as other crosswalks, enhanced query tools and mapping services used in the health services research community, will improve our understanding of the potential systematic impact on large-scale clinical research.

References

1. Intelligent Medical Objects, Inc., 2018
2. World Health Organization International Classification of Diseases. 9th edn Geneva: Switzerland, 1977
3. World Health Organization International Classification of Diseases, 10th edn Geneva, Switzerland, 1992
4. International Health Terminology Standards Development Organization SNOMED Clinical Terms (SNOMED CT) Technical Implementation Guide. Copenhagen, Denmark: IHTSDO, 2009
5. Fleurence RL,ea. Launching PCORnet, a national patient-centered clinical research network. Journal of the American Medical Informatics Association 21.4. 2014: p. 578-582.
6. Ohno-Machado Lea. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. Journal of the American Medical Informatics Association 21.4. 2014: p. 621-626.
7. Reich, C., Ryan, P. B., Stang, P. E., & Rocca, M. (2012). Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *Journal of biomedical informatics*, *45*(4), 689-696.
8. Christopher B Forrest, Peter A Margolis, L Charles Bailey, Keith Marsolo, Mark A Del Beccaro, Jonathan A Finkelstein, David E Milov, Veronica J Vieland, Bryan A Wolf, Feliciano B Yu, Michael G Kahn; PEDSnet: a National Pediatric Learning Health System, *Journal of the American Medical Informatics Association*, Volume 21, Issue 4, 1 July 2014, Pages 602–606, https://doi.org/10.1136/amiajnl-2014-002743
9. Ritu Khare, Levon Utidjian, Byron J Ruth, Michael G Kahn, Evanette Burrows, Keith Marsolo, Nandan Patibandla, Hanieh Razzaghi, Ryan Colvin, Daksha Ranade, Melody Kitzmiller, Daniel Eckrich, L Charles Bailey; A longitudinal analysis of data quality in a large pediatric data research network, *Journal of the American Medical Informatics Association*, Volume 24, Issue 6, 1 November 2017, Pages 1072–1079, https://doi.org/10.1093/jamia/ocx033
10. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., & Stang, P. E. (2011). Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, *19*(1), 54-60.
11. Observational Medical Outcomes Partnership Website. http://omop.org/Vocabularies. (version v5.3 09-APR-18)
12. Wasserman, H., & Wang, J. (2003). An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. In *AMIA Annual Symposium Proceedings* (Vol. 2003, p. 699). American Medical Informatics Association.
13. Matcho, A., Ryan, P., Fife, D., & Reich, C. (2014). Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. *Drug safety*, *37*(11), 945-959.
14. Hripcsak, G., Levine, M. E., Shang, N., & Ryan, P. B. (2018). Effect of vocabulary mapping for conditions on phenotype cohorts. *Journal of the American Medical Informatics Association*, *25*(12), 1618-1625.
15. Fung, K. W., Richesson, R., Smerek, M., Pereira, K. C., Green, B. B., Patkar, A., Clowse, M., Bauck, A., Bodenreider, O. (2016). Preparing for the ICD-10-CM Transition: Automated Methods for Translating ICD Codes in Clinical Phenotype Definitions. *EGEMS (Washington, DC)*, *4*(1), 1211. doi:10.13063/2327-9214.1211
16. Rijnbeek, P. R. (2014). Converting to a Common Data Model: What is Lost in Translation?. *Drug safety*, *37*(11), 893-896.