

## Characterizing database granularity using SNOMED-CT hierarchy

Anna Ostropelets, MD<sup>1</sup>, Christian Reich, MD, PhD<sup>2</sup>, Patrick Ryan, PhD<sup>1,3</sup>, Chunhua Weng, PhD<sup>1</sup>, Anthony Molinaro<sup>3</sup>, Frank DeFalco<sup>3</sup>, Jitendra Jonnagaddala, MIS, PhD<sup>4</sup>, Siaw-Teng Liaw, PhD, FACMI<sup>4</sup>, Hokyun Jeon<sup>5</sup>, Rae Woong Park, MD, PhD<sup>5</sup>, Matthew E. Spotnitz, MD, MPH<sup>1</sup>, Karthik Natarajan, PhD<sup>1</sup>, George Argyriou<sup>2</sup>, Kristin Kostka, MPH<sup>2</sup>, Robert Miller<sup>6</sup>, Andrew Williams<sup>6</sup>, Evan Minty<sup>7</sup>, Jose Posada<sup>8</sup>, George Hripcsak<sup>1,9</sup>

<sup>1</sup>Columbia University, New York, NY, USA; <sup>2</sup>IQVIA, Cambridge, MA, USA; <sup>3</sup>Janssen Epidemiology Analytics, Janssen Research & Development, Titusville, NJ, USA; <sup>4</sup>University of New South Wales Sydney, Australia; <sup>5</sup>Ajou University Graduate School of Medicine, Suwon, Gyeonggi-do, Republic of Korea; <sup>6</sup>Tufts Medical Center, Institute for Clinical Research and Health Policy Studies, Boston, MA, USA; <sup>7</sup>O'Brien Centre for Population Health, Faculty of Medicine, University of Calgary, Canada; <sup>8</sup>Stanford Center for Biomedical Informatics Research, Stanford, CA, USA; <sup>9</sup>Medical Informatics Services, New York-Presbyterian Hospital, New York, NY, USA

### Abstract

*Multi-center observational studies require recognition and reconciliation of differences in patient representations arising from underlying populations, disparate coding practices and specifics of data capture. This leads to different granularity or detail of concepts representing the clinical facts. For researchers studying certain populations of interest, it is important to ensure that concepts at the right level are used for the definition of these populations. We studied the granularity of concepts within 22 data sources in the OHDSI network and calculated a composite granularity score for each dataset. Three alternative SNOMED-based approaches for such score showed consistency in classifying data sources into three levels of granularity (low, moderate and high), which correlated with the provenance of data and country of origin. However, they performed unsatisfactorily in ordering data sources within these groups and showed inconsistency for small data sources. Further studies on examining approaches to data source granularity are needed.*

### Introduction

Over the past years, there has been an increasing need for studies on real-world patient scenarios (1). Such observational studies use electronic health records (EHR) and reimbursement claims data, but are criticized for potential residual confounding and bias (2). To address that, best practices were established to ensure internal validity. Some examples include data quality assurance (3), propensity score adjustment (4), negative and positive controls (5). External validity and generalization of findings obtained in observational studies can, in turn, be ensured by conducting multi-center studies, which also present multiple challenges to study design and execution(6). Modern common data models such as Informatics for Integrating Biology & the Bedside (i2b2) (7), Clinical Data Interchange Standards Consortium Study Data Tabulation Model (CDISC SDTM) (8) and Observational Health Data Sciences and Informatics (OHDSI) (9) combine observational data across different sites into a network, which enables research at large scale and improves study validity (10,11). Participating data partners can be from multiple countries and institutions, requiring harmonization of their disparate coding schemas and practices (12), different disorder definitions and underlying populations (13). Even when formal semantic interoperability is achieved by standardizing data formats and applying common terminologies, there can be substantial data heterogeneity across sites. Therefore, multi-center studies require familiarity with local patterns of clinical concept use to ensure that all concepts of interest are captured.

Researchers define patient cohorts of interest based on the clinical concepts available at their data sources. The problem is they cannot assess the availability of these concepts at other sites, and it is unclear to what extent concept utilization differs across data sources. There is little knowledge about how granular and heterogeneous concepts are

in different data sites. For example, patients with chronic kidney disease can be identified based on presence of chronic kidney disorder codes (ICD9-CM 585 or ICD10-CM N18 “Chronic kidney disease”) (14). But codes with less explicit content (ICD9-CM 586, ICD10-CM N18.9 ‘Renal failure, unspecified’ or N19 ‘Kidney failure’) are also used in the data to represent such patients (15,16). In this case, knowing data source granularity is essential for the appropriate disorder definition that can be used across different sites.

To our knowledge, there is no established practice on how to estimate granularity of data sources. This study aims to fill this knowledge gap by investigating heterogeneity, diversity, and granularity of clinical concept across different data sources within the OHDSI network.

## Methods

For the purpose of this study, we focused on representation of Conditions, which are defined as diagnoses, symptoms and signs.

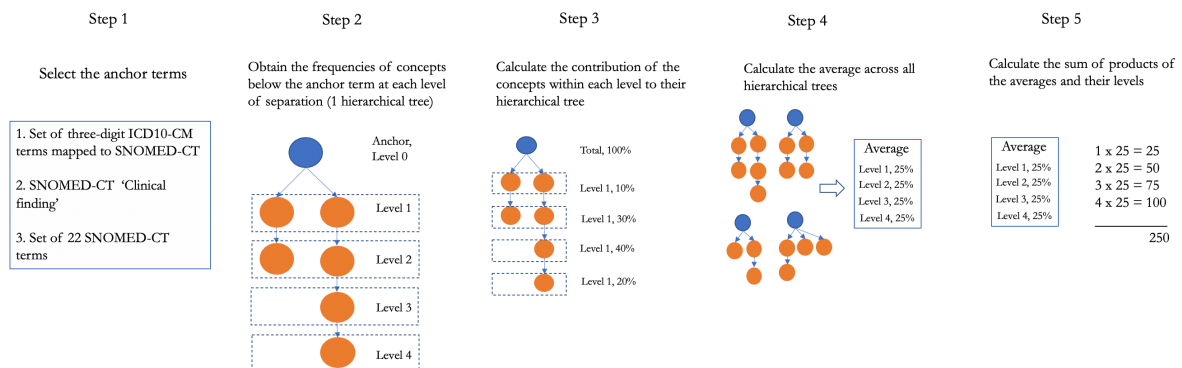
### 1. Data collection

We conducted a study within the OHDSI network with participating data sites having standardized their data to OHDSI’s Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) version 5 (9). Within the CDM, OMOP Standardized Vocabularies provide the comprehensive crosswalks from source terminologies to a standard terminology, which is then used to populate CDM tables (9). For example, our study included several source vocabularies (ICD10, ICD10-CM, ICD9-CM, ICDO3, KCD7) as well as free text entries, all of which were mapped to the target vocabulary - SNOMED-CT (17). For each SNOMED-CT concept, we counted the number of records in each data source. Hereon, we use the term ‘concept’ to refer to SNOMED-CT concepts and ‘frequency’ to refer to the number of records in a data source.

### 2. Data Analysis

#### 2.1. Data source granularity score

Here, we introduce the term ‘granularity score’, which refers to the overall level of granularity of conditions in a data source and can be used as a relative metric to compare different data instances. We calculated the granularity score for each data source using three alternative approaches described below. In each approach, we calculated the minimal number of steps (‘Is a’ relationships) within the SNOMED-CT hierarchy needed to get from a concept A found in the data to a generic anchor concept B. These steps or levels of separation were used as a proxy for granularity, assuming that concepts within one level of separation have similar semantic distance. First of all, these three approaches (Figure 1, Step 1) differ in the anchor concepts from which to measure the granularity score. The reasoning behind using anchors (ancestor terms) was to have a consistent metrics for concepts and, therefore, for different concepts to be comparable.



**Figure 1.** Overall study design for comparing concept granularity across different data sources. DS – data source.

We tested the following anchor concepts:

**Approach 1.** SNOMED-CT concepts mapped from three-character ICD10CM codes, excluding chapters 18-21 (signs and symptoms, injuries, external causes of morbidity and factors influencing health status). The purpose of this approach is to adjust for the fact that SNOMED-CT may have different levels of granularity in different parts of the hierarchy, whereas the ICD10CM three-character codes may be less variable.

**Approach 2.** Broadest SNOMED-CT term ‘Clinical Finding’. This assumes that in different parts of the hierarchy, the same degree of detail is encoded at about the same level down from ‘Clinical Finding’ for different diseases.

**Approach 3.** A set of 22 hand-selected SNOMED-CT terms that represent groups of conditions central to medicine (Table 1). In this way, we could manually ensure that the concepts were at a similar level of granularity.

**Table 1.** SNOMED concepts used as ancestors in calculating granularity score

SNOMED code	SNOMED category name	SNOMED code	SNOMED category name
55342001	Neoplastic disease	53619000	Disorder of digestive system
362971004	Disorder of lymphatic system	80659006	Disorder of skin and/or subcutaneous tissue
111590001	Disorder of lymphoid system	928000	Disorder of musculoskeletal system
362970003	Disorder of hemostatic system	42030000	Disorder of the genitourinary system
299691001	Finding of blood, lymphatics and immune system	362972006	Disorder of labor / delivery
362969004	Disorder of endocrine system	173300003	Disorder of pregnancy
74732009	Mental disorder	362973001	Disorder of puerperium
118940003	Disorder of nervous system	414025005	Disorder of fetus or newborn
128127008	Visual system disorder	66091009	Congenital disease
362966006	Disorder of auditory system	49601007	Disorder of cardiovascular system
271983002	Disorder of cardiac pacemaker system	50043002	Disorder of respiratory system

For each of these anchor concepts, we obtained the frequencies of all the descendant concepts according to the SNOMED-CT hierarchy at each level (Figure 1, Step 2) and calculated the distribution of concepts across different levels (Figure 1, Step 3). We then calculated the average distribution across all anchors (Figure 1, Step 4) and multiplied it with the corresponding levels of separation to arrive at a weighted distribution. Finally, granularity score was defined as the sum of weighted distribution obtained at Step 4 (Figure 1, Step 5).

Full process is described as:

$$\sum_{L \in 0, \dots, \max \text{ levels}} \left( \frac{\sum_{a \in A} \left( \frac{\sum_{l: c_l \text{ in } L \text{ levels from } c_a}{\sum_{l: c_l \text{ is descendant of } c_a} c_l} \right)}{N_A} \times L \right)$$

where C is the frequency of a concept, level is the level of separation, A is the set of ancestors (anchor terms) and  $N_A$  – number of ancestors.

## 2.2. Vocabulary granularity

Separately, we examined the granularity of vocabularies used in participating data sources (rather than the data sources themselves) to distinguish influence of different source coding schemas driving the granularity score as

opposed to preferences in the capture process To achieve that, we calculated the weighted distribution of target SNOMED-CT concepts across different levels of separation, where the levels were computed from three separate anchor terms described above.

### 2.3. Granularity applied to the real-world phenotyping tasks

Finally, to illustrate how the idea of granularity can be used to analyze data sources for specific disorders, we examined the granularity of databases for chronic kidney disorder. We used the most common definition of chronic kidney disorder (all concept in groups ICD9-CM 585 or ICD10-CM N18 ‘Chronic kidney disease’) (14), mapped it to SNOMED-CT (709044004 ‘Chronic kidney disease’) and calculated total frequency of concepts within its hierarchical tree at each level.

## Results

We collected data from seven data partners and twenty-two data sources: 14 US and 8 non-US. Their description, total number of condition records and unique condition concept codes per data source can be found on the GitHub page of the study (<https://github.com/ohdsi-studies/ConceptPrevalence/wiki/Participating-data-sources>).

The data originated mainly from administrative claims (8), hospital charge data (3) and electronic health records collected in large teaching hospitals (3) or primary and secondary practices (5). The size of the datasets varied greatly, with the average number of 644 million (51 million – 3 billion) condition records and 15.8 thousand (6.3–16.5 thousand) unique condition concepts per data source.

### *Data source granularity*

We analyzed data source granularity using the three approaches and established 5 empirical granularity levels based on the distribution of granularities of the data sources: high, high/moderate, moderate, moderate/low and low. In most cases, all three approaches agreed (Table 2). For high/moderate and moderate/low data sources two approaches showed moderate granularity and one – high or low respectively.

**Table 2.** Granularity scores for 22 participating data sources.

Database	Approach 1	Approach 2	Approach 3	Empirical level of granularity if agreed across approaches
AU-ePBRN	157	512	344	High granularity
Ajou	117	516	347	High granularity
CUMC	114	519	355	High granularity
MDCR	114	519	357	High granularity
NHIS/NSC Korea	111	510	336	Moderate/high granularity
STaRR	113	509	345	Moderate/high granularity
HCUP	125	498	346	Moderate granularity
PanTher	107	503	324	Moderate granularity
PREMIER	110	496	332	Moderate granularity
MDCD	111	490	333	Moderate granularity
Hospital CDM	111	503	335	Moderate granularity
CCAE	110	500	340	Moderate granularity
OpenClaims	110	505	342	Moderate granularity
Optum DOD	110	506	342	Moderate granularity
Optum SES	110	506	342	Moderate granularity
AmbEMR	114	490	314	Moderate/low granularity

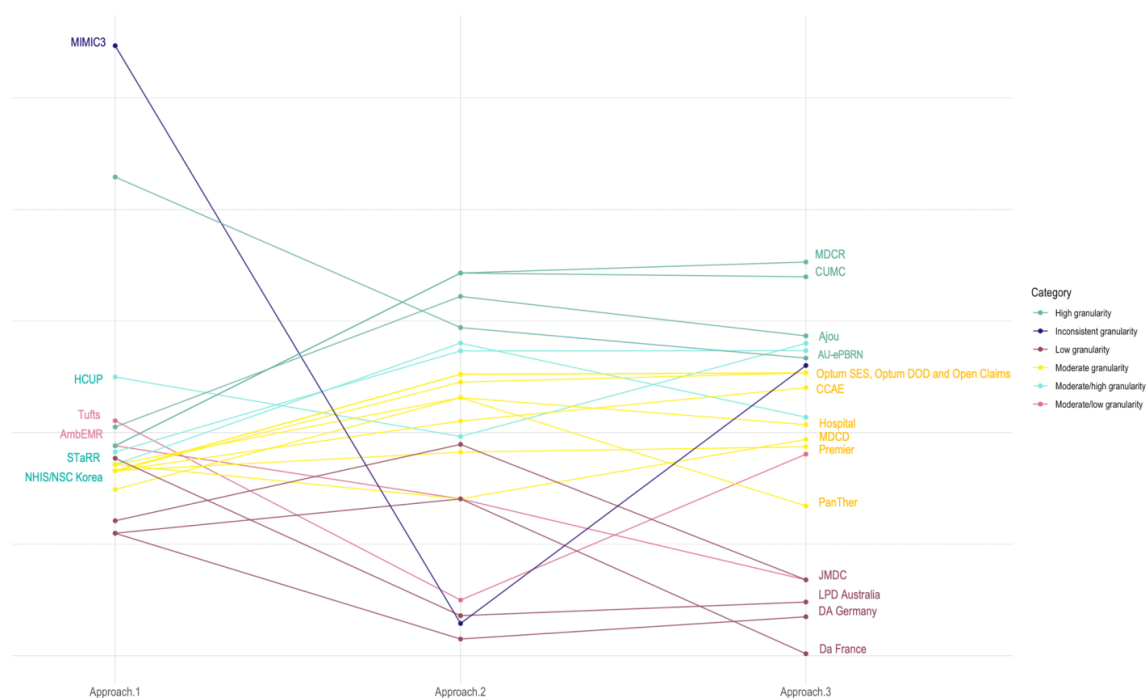
Tufts	118	477	331	Moderate/low granularity
DA France	100	490	304	Low granularity
DA Germany	100	472	309	Low granularity
JMDC	102	497	314	Low granularity
LPD Australia	112	475	311	Low granularity
MIMIC3	178	474	343	Inconsistent granularity

Regardless of the approach, most of the data sources had moderate granularity (Figure 2). This group included mainly administrative claims (MDCD, CCAE, OpenClaims, OptumDOD, and OptumSES) and hospital charge data (Hospital, HCUP, and Premier) along with only one EHR source (PanTher).

We identified four data sources with high granularity: AU-ePBRN, MDCR, CUMC and Ajou University database, which remained relatively granular regardless of the method used. STaRR and NHIS/NSC Korea appeared to be highly granular or moderately granular depending on the approach.

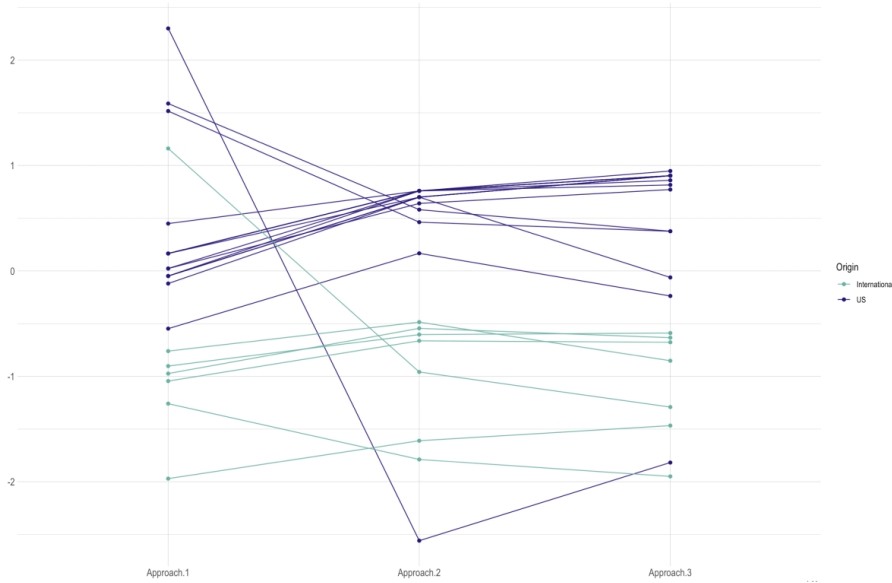
The low granularity group was the most homogeneous group, consisting of international data sources, which were primarily EHR-derived (LPD Australia, DA France and DA Germany), accompanied by one claims-derived source (JMDC). Another EHR source, AmbEMR, appeared as a low or moderate granularity data source.

We found one data source with noticeable inconsistency across approaches: MIMIC3, characterized by the limited number of unique concepts, was highly granular in approaches 1 and 3 and the least granular compared to other sources in approach 2.



**Figure 2.** Granularity score for participating data sources, grouped by level of granularity.

We also found some patterns in data granularity related to the provenance of the data. Overall, EHR data sources originated from primary and secondary care practices appear to be less granular, while administrative claims data, hospital charge data and EHR data originating from large tertiary care hospitals were more granular. International data sources were on average less granular with only three out of eight non-US sources being moderately or highly granular (Figure 3).



**Figure 3.** Vocabulary granularity for participating data sources, US (blue) and international (green) data sources.

Administrative claims data and hospital charge data showed similar patterns of granularity, but relative granularity within this group differed depending on the approach. MDCR had the highest granularity among other claims data, Optum DOD, Optum SES and OpenClaims had similarly moderate granularity, and MDCD with Premier had consistently low granularity in the group.

### *Granularity for specific disorders*

When analyzing the distribution of concepts for chronic kidney disorder, we found that on average 59% percent of records were as granular as the concept ‘Chronic kidney failure stage 3’ (Table 3). Some of sources comprised broader terms. For example, less precise concept ‘Renal impairment’ accounted for 23% of all concepts related to chronic kidney disorder in LPD Australia. Given its prevalence, this concept should be placed under scrutiny to determine if it should be used to find patients with chronic kidney disorder in this data source.

**Table 3.** Selected datasets for assessing granularity of chronic kidney disorder.

Levels of separation	DA France	JMDC	LPD Australia	AmbEMR	CUMC	MDCD	Average*
0;Renal impairment			23%	4%	0.01%		2.2%
1; Chronic kidney disease	94%	90%	32%	17%	25%	13%	24.7%
2; Chronic kidney disease stage 3	5%	9%	45%	69%	64%	68%	59.4%
3;Chronic kidney disease stage 3 due to hypertension	0.04%	1%		10%	9%	19%	13.1%
4;Malignant hypertensive chronic kidney disease stage 3	0.04%	0.1%		0.2%	1%	1%	0.5%
5;Malignant hypertensive end stage renal disease on dialysis				0.01%			0.0001%

\*Average frequency of concepts at a level across all databases

### **Discussion**

In this study, we explored concept granularity across disparate data sources with different provenance of data, country of origin and various coding methods.

Acknowledging data source granularity is a necessary step in observational studies run on multiple data sources. As similar patients can be coded with various granularity in different data sources, it is important to be aware of the overall data source granularity to make informed decisions about phenotyping algorithms. When using data sources with low granularity (as LPD Australia in this study), using less precise broad concepts is needed in order not to lose patients of interest. For example, when identifying patients with chronic kidney failure, researchers may opt for looking at broader concepts such as renal impairment. The latter accounts for nearly a quarter of all kidney disorder-related records in LPD Australia. Given that such a broad concept is not likely to initially be included in the phenotyping algorithm, it is important to recognize the fact that a large fraction of patients has this code.

We will first discuss advantages and disadvantages of scoring approaches, followed by discussing our observations. When examining granularity, we have to account for three groups of factors:

1. Vocabulary
2. Underlying population
3. Granularity of the data capture

### *Vocabulary*

SNOMED-CT is the most comprehensive reference terminology available and is a mandatory standard vocabulary for conditions in the OHDSI network. SNOMED-CT supports polyhierarchy, where a concept may have multiple ancestors and inherits their meaning. Such polyhierarchies coexist in SNOMED-CT equally, so that a single main hierarchical path cannot be identified. A concept can appear in multiple hierarchical trees at different levels, which obstructs assessing its complexity level when multiple anchoring terms are used. For example, 51292008 ‘Hepatorenal syndrome’ appears in two hierarchical trees: 42030000 ‘Disorder of the genitourinary system’ (5 levels of separation) and 53619000 ‘Disorder of digestive system’ (2 levels of separation). While it poses challenges to establishing hierarchy-based granularity of an individual concept or an individual data source, such ambiguity is leveled out when seeking relative comparison.

In this work, we used different approaches that vary in anchoring terms. Using ‘Clinical Finding’ as a single ancestor term prevented duplication of terms across different hierarchical trees and allowed us to analyze the whole set of condition concepts in the data sources and obtain a more comprehensive picture than in the other approaches. A disadvantage of such approach is participation of all concepts, even those that carry insignificant clinical meaning. For example, ICD9-CM concept 780.99 ‘Other general symptoms’ frequently occurred in some of the data sources and, being mapped to SNOMED-CT 365860008 ‘General clinical state finding’, conveyed little clinical meaning. Even if such a concept is present in a data source, it cannot be acted upon: it communicates too little clinical meaning to be used to define any disorder of interest.

The ICD10-CM term-based approach was motivated by selecting patients in observational research, which is typically performed by selecting appropriate ICD10CM codes to define disease or state. Such study design can be inefficient when international data sources or data sources with unstructured data processing are involved. Indeed, source vocabularies in non-US data sources were less granular. This was expected as ICD10-CM, used in the US, is more granular than ICD10 used internationally. If a feasibility study is performed on a highly granular data source, too specific concepts may be selected for phenotyping, which will lead to the patient loss. The ICD10-CM-based approach allows exploring data granularity, which will be extremely relevant to studies performed on the US data sources or driven by ICD10-CM concept selection. Moreover, the granularity score is computed based on clinically meaningful concepts, which potentially have higher practical impact.

On the other hand, this approach neglects concepts broader than the selected ICD10-CM counterparts, which can be of a particular interest in low granular data sources. The SNOMED-CT concept set approach (approach 3) overcomes this shortcoming by querying broad disorder groups.

In these two approaches duplication of concepts across different trees was offset by averaging those trees. Nevertheless, if a concept space within a dataset is limited (like in MIMIC3, which only contains 749 unique SNOMED-CT concepts), this approach will be sensitive to concept selection. Although we tried to minimize this effect by excluding groups of disorders that have high overlap, duplicates can still be found and can potentially bias the granularity score for small data sources.

### ***Underlying population***

Granularity can reflect the features of the population that had given rise to a data source. Unbalanced data sources with a focus on a specific population may be biased towards higher granularity for this population but remain otherwise non-granular. For example, 85% of MDCR patients are elderly, who tend to have more co-morbidities compared to young healthy patients (18). Co-morbidities, in turn, are coded as granular complex concepts that reflect associations between disorders, e.g. 422166005 ‘Peripheral circulatory disorder associated with type 2 diabetes mellitus’ or 19034001 ‘Hyperparathyroidism due to renal insufficiency’. Such high granularity is attributable rather to characteristics of the population (patients) than to characteristics of processes (data collection, coding or transformation). If a certain level of granularity belongs only to a specific portion of the data source, we need to disentangle this effect to be able to assess the baseline level of granularity. The latter will then reflect the granularity for the other groups of patients in a source, which can also be used for research.

We proposed to offset the influence of a particular patient group on data source granularity by stratifying concepts by disorder group (approach 1 and 3). In particular, it resulted in a reduced difference in the granularity of MDCD and MDCR, which was more extreme in approach 2.

### ***Granularity of data capture***

The data can be generated to address different needs: electronic health records facilitate clinical records storage and retrieval and administrative claims data is used in the reimbursement process. Clinical documents within electronic health records and administrative claims may capture similar patients differently. Electronic health records may tend to be less granular due to the nature of clinical workflow, while claims data can be more granular to maximize reimbursement.

It is supported by our observations that administrative claims data and hospital charge data were on average more granular than EHR data, especially if a data source originated from primary or specialty practices. Large hospitals EHR data appeared to be highly granular, which may suggest similar coding procedures for these sources.

We previously discussed granularity should be adjusted if a subset of patients influences granularity. Patient characteristics can also be viewed as a feature of data source granularity if the patient population is homogeneous. In this way, granularity has the potential to remain stable regardless of a selected fraction of patients.

Coding methods applied to unstructured data can also contribute to concept heterogeneity. Extracting data from clinical notes is a tedious and complicated process, which may decrease concept granularity as free text, especially if in large volume, may be converted to broad and imprecise structured data (19).

### **Limitations**

We did not perform targeted SNOMED-CT auditing to identify hierarchy inconsistencies, incomplete modeling or other issues described elsewhere (20,21). As SNOMED-CT is the most comprehensive and continuously updated reference terminology, we assumed that such issues will not be detrimental to assessing granularity or will influence all data sources equally. In this study, we only analyzed conditions as a comprehensive hierarchy for procedures or measurements is lacking; including other domains in granularity score may be included in future work. This work does not focus on the implications of different granularity on patient selection, and we did not evaluate performance of computable algorithms with different granularity.

### **Future work**

SNOMED-CT defines its concepts not only with hierarchical links, but also with ‘has-a’ relationships, which can potentially be used to assess granularity. While attribute-based granularity inference is complicated by inconsistencies in assigning attributes and high volume of relationship types (20), future work may include comparing hierarchy-based approaches to attribute-based approaches.

Future work may also include further characterization of factors that contribute to data source granularity, for example disentangling coding variances and data transformation aspects.

### **Conclusion**



Multi-center observational studies require recognition and reconciliation of differences in patient representations arising from underlying populations, disparate coding practices and specifics of data capture. Granularity of data sources should be evaluated to ensure comprehensiveness, yet appropriateness of concepts selected to represent a condition of interest. When examining granularity, researcher should account for three main components: vocabulary, underlying population and granularity of data source itself. Here, we presented three approaches to calculating data source granularity based on SNOMED-CT hierarchy. They showed consistency in classifying data sources into three levels of granularity (low, moderate and high), which correlated with the provenance of data and country of origin. However, they performed unsatisfactorily in ordering data sources within these groups and showed inconsistency for small data sources. Further studies on examining approaches to data source granularity are needed.

## Acknowledgments

This project was supported by The National Library of Medicine grants R01LM006910 and R01LM009886.

## 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. Grimes DA, Schulz KF. Bias and causal associations in observational research. *The Lancet*. 2002 Jan;359(9302):248–52.
3. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2016 Sep 11;4(1):18.
4. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*. 1984 Sep;79(387):516–24.
5. J. Schuemie M, Soledad Cepede M, A. Suchard M, Yang J, Tian Alejandro Schuler Y, B. Ryan P, et al. How Confident Are We About Observational Findings in Health Care: A Benchmark Study. *Harvard Data Science Review* [Internet]. 2020 Jan 31 [cited 2020 Feb 29]; Available from: <https://hdr.mitpress.mit.edu/pub/fxz7kr65>
6. Sprague S, Matta JM, Bhandari M. Multicenter Collaboration in Observational Research: Improving Generalizability and Efficiency: *The Journal of Bone and Joint Surgery-American Volume*. 2009 May;91(Suppl 3):80–6.
7. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*. 2010 Mar 1;17(2):124–30.
8. Wood F, Guinter T. Evolution and implementation of the CDISC study data tabulation model (SDTM). *Pharmaceutical Programming*. 2008;1(1):20–27.
9. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574–8.
10. Garza M, Del Fiore G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. *Journal of Biomedical Informatics*. 2016 Dec;64:333–41.
11. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*. 2012 Jan;19(1):54–60.
12. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. *Journal of Public Health*. 2012 Mar 1;34(1):138–48.
13. Lix L, Yogendran M, Shaw S, Burchill C, Metge C, Bond R. Population-based data sources for chronic disease surveillance. *Chronic Dis Can*. 2008;29(1):31–38.
14. Vlasschaert MEO, Bejaimal SAD, Hackam DG, Quinn R, Cuerden MS, Oliver MJ, et al. Validity of Administrative Database Coding for Kidney Disease: A Systematic Review. *American Journal of Kidney Diseases*. 2011 Jan;57(1):29–43.
15. Gentile G, Postorino M, Mooring RD, De Angelis L, Manfreda VM, Ruffini F, et al. Estimated GFR reporting is not sufficient to allow detection of chronic kidney disease in an Italian regional hospital. *BMC Nephrology*

- [Internet]. 2009 Dec [cited 2020 Mar 8];10(1). Available from:  
<http://bmcnephrol.biomedcentral.com/articles/10.1186/1471-2369-10-24>
16. Kern EFO, Maney M, Miller DR, Tseng C-L, Tiwari A, Rajan M, et al. Failure of ICD-9-CM Codes to Identify Patients with Comorbid Chronic Kidney Disease in Diabetes. *Health Services Research*. 2006 Apr;41(2):564–80.
  17. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. *Journal of the American Medical Informatics Association*. 2018 Dec 1;25(12):1618–25.
  18. Rowland D, Lyons B. Medicare, Medicaid, and the elderly poor. *Health Care Financ Rev*. 1996;18(2):61–85.
  19. Hruschka DJ, Schwartz D, St.John DC, Picone-Decaro E, Jenkins RA, Carey JW. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods*. 2004 Aug;16(3):307–31.
  20. Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *Journal of the American Medical Informatics Association*. 2011 Jul;18(4):432–40.
  21. Cui L, Bodenreider O, Shi J, Zhang G-Q. Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs. *Journal of Biomedical Informatics*. 2018 Feb;78:177–84.