Using SNOMED to Automate Clinical Concept Mapping

Shaun Gupta
Predictive Analytics, Real World Solutions, IQVIA
London, United Kingdom

Frederik Dieleman
Predictive Analytics, Real World Solutions, IQVIA
London, United Kingdom

Patrick Long
Predictive Analytics, Real World Solutions, IQVIA
Plymouth Meeting, Pennsylvania
United States

Orla Doyle
Predictive Analytics, Real World Solutions, IQVIA
London, United Kingdom

Nadejda Leavitt
Predictive Analytics, Real World Solutions, IQVIA
Plymouth Meeting, Pennsylvania
United States

ABSTRACT

The International Classification of Disease (ICD) is a widely used diagnostic ontology for the classification of health disorders and a valuable resource for healthcare analytics. However, ICD is an evolving ontology and subject to periodic revisions (e.g. ICD-9-CM to ICD-10-CM) resulting in the absence of complete cross-walks between versions. While clinical experts can create custom mappings across ICD versions, this process is both time-consuming and costly. We propose an automated solution that facilitates interoperability without sacrificing accuracy.

Our solution leverages the SNOMED-CT ontology whereby medical concepts are organised in a directed acyclic graph. We use this to map ICD-9-CM to ICD-10-CM by associating codes to clinical concepts in the SNOMED graph using a nearest neighbors search in combination with natural language processing. To assess the impact of our method, the performance of a gradient boosted tree (XGBoost) developed to classify patients with Exocrine Pancreatic Insufficiency (EPI) disorder, was compared when using features constructed by our solution versus clinically-driven methods. This dataset comprised of 23, 204 EPI patients and 277, 324 non-EPI patients with data spanning from October 2011 to April 2017. Our algorithm generated clinical predictors with comparable stability across the ICD-9-CM to ICD-10-CM transition point when compared to ICD-9-CM/ICD-10-CM mappings generated by clinical experts. Preliminary modeling results showed highly similar performance for models based on the SNOMED mapping vs clinically defined mapping (71% precision at 20% recall for both models). Overall, the framework does not compromise on accuracy at the individual code level or at the model-level while obviating the need for time-consuming manual mapping.

CCS CONCEPTS

• **Computing methodologies** → *Feature selection.*

KEYWORDS

feature engineering, graph algorithms, natural language processing, data interoperability, claims data, ICD9, ICD10, SNOMED

ACM Reference Format:
Shaun Gupta, Frederik Dieleman, Patrick Long, Orla Doyle, and Nadejda Leavitt. 2020. Using SNOMED to Automate Clinical Concept Mapping. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3368555.3384453

1 INTRODUCTION

The International Classification of Disease (ICD) is a widely used diagnostic ontology for the classification of health disorders and the billing of diagnoses and procedure claims [11]. It provides a comprehensive and standardized system that may be leveraged for evidence-based analytics of population health trends and for clinical research. Healthcare data encoded using the ICD ontology are of increasing value to machine learning efforts that use patient medical history to predict future events, such as diagnoses, changes in line of therapy, or responsiveness to available treatments.

Unfortunately, the ICD ontology poses several challenges for longitudinal data capture. First, it is an evolving ontology, which makes it difficult to establish stable representations of disease or other health conditions across the periodic ICD revisions. For example, in October 2015 the United States switched from the use of ICD-9-CM to the much more granular ICD-10-CM system. Throughout the paper we will refer to the ICD-9-CM ontology as ICD-9 and the ICD-10-CM ontology as ICD-10. One-to-one cross-walks do not exist between these revisions. Rather, most ICD-9 and ICD-10 mappings are approximate or one-to-many [1–3] making analytics of medical diagnoses difficult across versions. For instance, changes in the use of certain ICD codes may underlie variation in disease prevalence near the time of the ICD-10 transition [7, 15]. Second, while ICD has a native hierarchy for the grouping of related codes, it is largely used to facilitate medical reimbursement and may not be optimal for healthcare analytics in cases where highly interpretable clinical definitions are desired.

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384453

Current solutions to ICD interoperability and interpretability rely on the manual work of domain experts. Mapping ICD codes across ICD versions may be performed with the aid of general equivalence maps (GEMs); however, the utility of these maps is limited by the elevated granularity and complexity of ICD-10 relative to ICD-9, which resulted in a nearly 5-fold increase in ICD codes between the two ontologies [10]. In fact, only 24% of ICD-9s and 5% of ICD-10s have an exact 1 to 1 match whereas 3% and 1% fail to match, and the remaining codes have one or more approximate or one-to-many relationships [1]. For this reason, the Centers for Medicare and Medicaid Services (CMS) discourages the interpretation of GEMs as simple "cross-walks" but rather as organizing frameworks to guide appropriate code selection [5]. The use of GEMs therefore necessitates expert supervision since they may result in imperfect code pairings [6, 15, 16]. Additionally, ICD codes may be manually aggregated into more interpretable ICD hierarchies, such as ICD level 4, though this too requires expert guidance when aggregating across ICD ontology revisions since hierarchies lack clear relationships. Both of these strategies are time-consuming and may ultimately discourage the use of patient data that span multiple ICD versions.

SNOMED CT is a unified clinical-concept ontology designed to facilitate interoperability between electronic healthcare systems and to support clinical research using electronic medical record data [8]. SNOMED is a Directed, Acyclic, hierarchical, Graph-based ontology (DAG) designed to maximize medical understanding. At its core, SNOMED is comprised of clinical concepts with affiliated clinical information, such as disease descriptions, and relationships, which link concepts together in a clinically meaningful way. For example, within the SNOMED ontology the concept for "bacterial pneumonia" has a child-parent relationship with the concept for "infective pneumonia" which is itself a subtype of the concept for "pneumonia" [9]. The interpretable and clinically-oriented design of SNOMED makes it a promising alternative to ICD-based clinical analysis, which is purposed more for billing and retrospective statistical studies.

Both ICD-9 and ICD-10 may be mapped to SNOMED concepts using mappings from the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM), which provides a unified medical vocabulary for healthcare data interoperability [13, 14]. These ICD to SNOMED mappings were created manually by clinical experts, covering > 99% of both ICD ontologies. In the majority of cases, ICD-9 and ICD-10 codes map to shared SNOMED concepts i.e. concepts that contain both ICD-9s and ICD-10s. However, a subset of ICD-9 and ICD-10 codes fail to map to shared SNOMED codes, which we refer to as "orphan" ICD codes. Orphan codes limit the ICD interoperability that might otherwise be achieved by mapping to SNOMED because unshared SNOMED concepts may misleadingly indicate temporal inconsistencies in disease prevalence near the time of ICD ontology transition.

To resolve these challenges and to identify mappings for orphan codes, we developed a graph-based search algorithm to automate the process of ICD-9 to ICD-10 mapping by leveraging the SNOMED clinical ontology. This algorithm finds shared SNOMED concepts and searches for optimal concept mappings in cases of orphan ICD codes using a nearest neighbors and a natural language processing (NLP)-based strategy. This mapping algorithm improves the interoperability between ICD versions and opens the door to analytics available via the SNOMED ontology. As a proof of concept, we show that the clinical features derived from our SNOMED mapping algorithm achieve stability over the 2015 ICD-10 transition and can be used to train machine learning models for disease detection with equivalent performance to those trained on ICD-9 and ICD-10 features curated manually by clinical experts.

2 METHODS

2.1 Mapping Algorithm

Our solution uses the SNOMED graph-based ontology and preexisting mappings from ICD-9 and 10 to SNOMED to create a map between ICD-9 and 10 codes that accounts for orphan codes. This ICD-9 to ICD-10 mapping is generated via four stages: (1) use the external ICD-SNOMED mappings to map all ICD codes to at least one SNOMED concept, (2) identify orphan ICD codes (ICD codes assigned to SNOMED concepts that have no codes from another version of ICD mapped to it), (3) search for partner codes for these orphan ICD codes, (4) subset identified pairs to those closest on the SNOMED graph and (optionally) use a pre-built Word2Vec model to find the pair with the highest semantic similarity to help identify the optimal partner code.

This Word2Vec model was obtained by training the Continuous Bag Of Words (CBOW) Word2Vec algorithm on ICD code descriptions with the following settings: 5 epochs, vector dimension of 300, window of 8, minimum word occurrence threshold of 3, downsampling threshold of 0.001 and minimum learning rate of 0.0001. The sum of the context vectors was used as the CBOW mean. Negative sampling was applied, with 5 noise words drawn and a negative sampling exponent of 0.75. This parameter choice was based on common Word2Vec configurations. During application of the CBOW algorithm on an ICD description, the Term Frequency–Inverse Document Frequency (TF-IDF) weighted average of word vectors in the description sentence was used to create an embedding vector for the description. For the optional reduction step, embedding vectors of paired ICD-9 and ICD-10 codes were compared, and the pair with the shortest distance between embedding vectors kept. This approach was validated by a clinical expert, whereby candidate pairs were determined to be clinically relevant.

To protect against an overly exhaustive graph traversal that could find pairings far from the original concept, we fixed the maximum search radius for the pairing step to a given number of nodes (e.g. 5) and then searched the graph in all directions within this limit. If a corresponding ICD-10 is not found (for an ICD-9 query), no pairing occurs. This limit is to prevent irrelevant pairings, or pairings that go to the root of the graph.

We define a query concept as a SNOMED concept containing orphan codes that we are trying to map to ICD codes of the opposite type. E.g. a SNOMED concept that only contains ICD-9 codes would ideally be paired with a relevant ICD-10 code/s. Target concepts are all the SNOMED concepts that contain at least one ICD code of the opposite type. This does not need to be exclusively of the opposite type, as we allow mapping of orphan codes to non-orphan codes. The final SNOMED concept that will be used as a feature is called

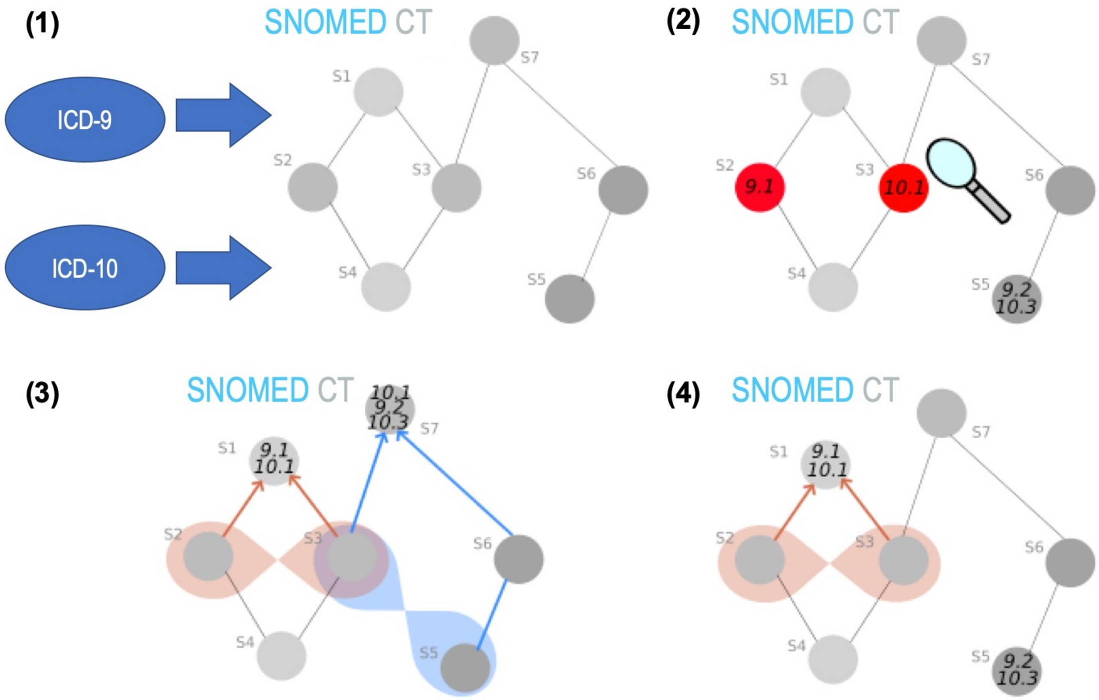


Figure 1: Overview of the complete algorithm, with the SNOMED concepts labelled as SX, ICD-9 codes as 9.X and ICD-10 codes as 10.X. For visual convenience, ancestors are positioned higher on the graph, descendants lower. Only "is-a" type of connections of the SNOMED graph are used in our algorithm. This means that both S3 and S2 are subsets of S1, just like "rheumatic heart disease" is a "cardiovascular disease", or "infectious disease" is a "disease". The four steps applied to this dummy example are: (1) ICD to SNOMED mapping. (2) Orphan identification: S2 and S3 contain orphan codes (9.1 and 10.1 respectively), while S5 contains both an ICD-9 and an ICD-10 code. (3) Pairing: The orphan codes in S2 and S3 can be paired together in their common ancestor S1. The orphan code in S3 can also be paired with the non-orphan codes in S5 into their common ancestor S7. (4) Reduction: Only the pairing of S2 and S3 into S1 is kept as the distance between them is smaller than the distance between S3 and S5.

the pivot node, which will contain the query ICD codes, target ICD codes and the ICD codes already mapped to the pivot node.

For example, to find ICD-10 codes for orphan ICD-9 codes, starting from SNOMED nodes mapped to the ICD-9 query of interest, we traverse the SNOMED DAG until we find a SNOMED concept that has ICD-10 code(s) mapped to it. Then during the reduction step, if multiple codes are equidistant from the ICD-9 query code, the algorithm has the option to look at the clinical descriptions of these target ICD-10 codes and use a pre-built Word2Vec model to find the one with the highest semantic similarity.

The full algorithm is illustrated in Figure 1 and described in detail below:

(1) ICD to SNOMED mapping: Assign all ICD-9 and -10 codes to a SNOMED concept through the externally available mappings.
(2) Orphan identification: Identify all SNOMED concepts that contain only ICD-9 or only ICD-10 codes, i.e. the "orphan concepts".

Perform the following step twice, once with ICD-9 orphans as queries and once with ICD-10 orphans as queries.

(3) Pairing the query concepts:
  (a) Identify all ancestors of the query concepts as defined via 'is-a' SNOMED concept relationships.
  (b) For all targets, identify shared ancestors between the target and the query. Define these shared ancestor nodes as pivot nodes.
  (c) Optional: Perform the search in the reverse direction (i.e. linking to descendant nodes lower in the graph as opposed to ancestor nodes higher on the graph).
  (d) Filter resulting pivot nodes (SNOMED concepts that group together ICD-9 and -10 codes) down to the mappings where the queries and targets are within a certain maximum distance on the graph.
(4) Reduction:
  (a) Remove duplicate pairs, which can arise from pairing in both directions (ICD-9 to -10 and ICD-10 to -9) or from multiple pathways on the graph creating the same query-target-pivot triplet.
  (b) Sub-select pairs closest together on SNOMED graph.

(c) Optional: look at the ICD descriptions of the paired query and target codes, and use a pre-built Word2Vec model to find the pair with the highest semantic similarity. This was applied to the mappings assessed in this paper.

2.2 Application to rare disease detection

The algorithm was applied to a patient cohort diagnosed with Exocrine Pancreatic Insufficiency (EPI) disorder, and a control patient cohort without the disorder. Data were extracted from US prescription and non-adjudicated medical claims. EPI patients were required to have a claim for an EPI ICD-10 diagnosis, or a Pancreatic Enzyme Replacement Therapy (PERT) treatment in conjunction with six treatment claims within a 12-month window between October 2015 and July 2017. Control patients were required to have a claim for at least one of the following diagnoses or specialty visits: cystic fibrosis, chronic pancreatitis, gastroenterology, general surgery, upper gastrointestinal endoscopy or proton pump inhibitors. This was done to ensure that control patients have shared symptoms with patients in the EPI cohort, so that the model focuses on specific aspects of EPI instead of more general characteristics of diseased vs. non-diseases patients. Both sets of patients were required to have 24 months of medical history (i.e. claim activity). The SNOMED algorithm was applied to diagnosis claims within this period, which were subsequently aggregated to counts per diagnosis code/concept for modeling features.

ICD-9 to ICD-10 SNOMED pairings were evaluated on three criteria:

i **Mapping assessment**: Comparison to clinical concepts created manually by a clinical expert.

ii **Temporal stability**: Stability of SNOMED mappings across the ICD transition period (Oct 2015). i.e. Are the SNOMED features continuous before/after October 2015 with respect to the average number of claim occurrences, or is there a sharp discontinuity that would imply the mappings are not adequately capturing the ICD-9/ICD-10 codes for the concept being investigated?

iii **Model performance**: Impact on the precision-recall curve of a gradient boosted tree algorithm (XGBoost [4]) developed to identify patients with EPI.

SNOMED and ICD concepts that occurred for >= 2% of the patient population were used. Model performance was compared to that of models developed using features derived by clinical experts and a naïve data-driven approach using ICD codes with level 4 aggregation. Clinical experts were comprised of pharmacists, nurses, or coding experts with comprehensive expertise in diagnosis and procedure coding. The XGBoost algorithm had the following hyperparameter settings: maximum tree depth of 3, learning rate of 0.1, gamma of 0, and 100 trees.

Two strategies outlined in Figure 2 were used for model evaluation. Strategy 1 – training and holdout are sampled across the transition period. Models were trained using patient data spanning the ICD-10 transition using a 75% to 25% train test split. Strategy 2 – training data sampled prior to the ICD transition period (ICD-9 claims only) and holdout data sampled post the transition period (ICD-10 claims only).

3 RESULTS

The EPI data set comprised $23,204$ EPI patients and $277,324$ control patients. Approximately 14k ICD-9 and 29k ICD-10 codes were grouped into 9k SNOMED concepts, with 737 of these concepts present in $\geq 2\%$ of the EPI or control cohort. In contrast, clinical experts selected 93 concepts relevant to EPI.

3.1 Mapping assessment

For the complete ICD ontology, the algorithm reduced the total number of ICD-10 and ICD-9 orphan codes from $44,774$ to 622 and $7,332$ to 158 respectively, resulting in more than $> 99\%$ of the ICD codes being paired successfully.

For the EPI cohort specifically, if the ICD->SNOMED mappings had been applied without our orphan pairing algorithm, only $\approx$ 7k ($\approx$ 52%) of ICD-9 and $\approx$ 11k ($\approx$ 39%) ICD-10 codes would have been paired (i.e. not orphans). After application of the algorithm, this was improved to $\approx$ 14k ICD-9 and $\approx$ 29k ICD-10 ($>$ 99%).

For the most prevalent 50 features in the EPI cohorts, SNOMED concepts mapped to fewer ICD codes than concepts created by clinical experts ($\approx$ 12 vs. $\approx$ 40, respectively). Our SNOMED mapping approach thus creates more aggregated and easily interpretable features than using the raw ICD codes albeit less aggregated than the concepts created by clinical experts. Despite this, the prevalence rates of the clinical and SNOMED concepts were found to be similar, suggesting that our approach is capturing the most codes with the most coverage. Table 1 shows the three most prevalent SNOMED concepts in our cohort, which show that ICD codes have been mapped to SNOMED concepts in a clinically meaningful manner.

3.2 Temporal stability

In order to assess the temporal stability of paired ICD codes across the transition period, the average number of claims per patient per week were plotted for both the SNOMED and clinical concepts. Discontinuity of the claims across the ICD transition period points to an unsuccessful conceptual mapping of ICD-9 and 10 features if one assumes the underlying medical diagnoses across that period stayed stable. A smooth, continuous transition is thus pivotal for a successful ICD-9 to ICD-10 mapping. Of the 200 most prevalent SNOMED concepts, 30% were considered to have discontinuous transitions (examples shown in Figure 3). The clinical mapping performed by experts (93 clinical concepts) still demonstrated around 10% of these discontinuous transitions. Such discontinuous transitions are an artefact of target ICD codes not being present in the positive/control cohort, resulting in SNOMED concepts linked to mainly/only one ICD code version - something that could be improved by prioritising pairs with high ICD code prevalences from the cohort of interest during the reduction phase of the algorithm.

3.3 Model performance

We evaluated the model performance using precision-recall curves [12], showing precision-recall pairs at different operating points. Precision, or positive predicted value (PPV), is the ratio of true positives over the sum of true positives and false positives, i.e. the fraction of true EPI patients in the group of predicted EPI patients. Recall, or sensitivity, is the fraction of positive patients identified as

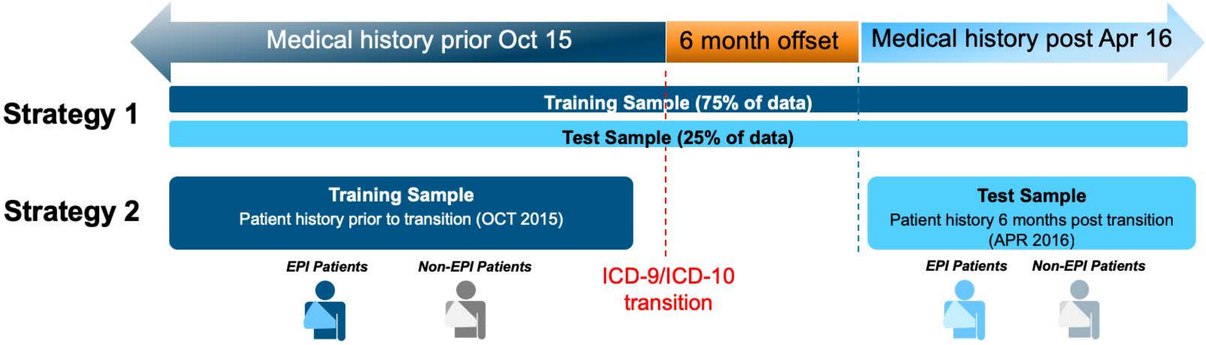
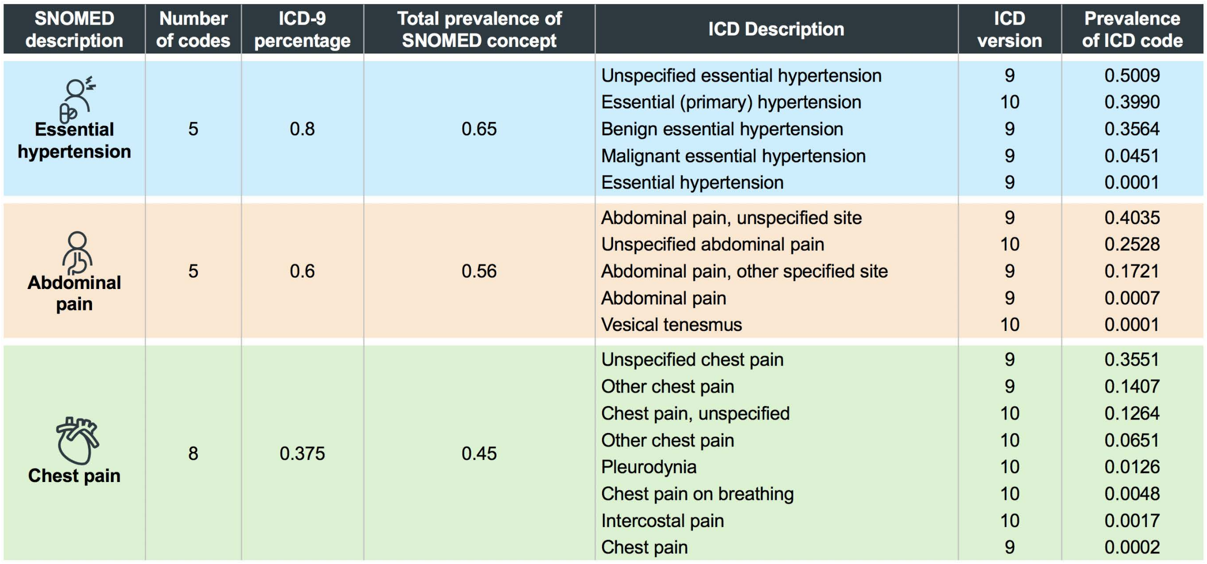

Medical history prior Oct 15
6 month offset
Medical history post Apr 16
Strategy 1
Training Sample (75% of data)
Test Sample (25% of data)
Strategy 2
Training Sample
Patient history prior to transition (OCT 2015)
Test Sample
Patient history 6 months post transition (APR 2016)
EPI Patients
Non-EPI Patients
ICD-9/ICD-10 transition
EPI Patients
Non-EPI Patients

Figure 2: Outline of two sampling strategies used for model evaluation. In Strategy 1, models were trained using patient data spanning the ICD-10 transition using a 75% to 25% train test split. In Strategy 2, models were trained exclusively on patient data prior to the ICD-10 transition and tested on patient data post transition.



SNOMED description	Number of codes	ICD-9 percentage	Total prevalence of SNOMED concept	ICD Description	ICD version	Prevalence of ICD code
Essential hypertension	5	0.8	0.65	Unspecified essential hypertension	9	0.5009
				Essential (primary) hypertension	10	0.3990
				Benign essential hypertension	9	0.3564
				Malignant essential hypertension	9	0.0451
				Essential hypertension	9	0.0001
Abdominal pain	5	0.6	0.56	Abdominal pain, unspecified site	9	0.4035
				Unspecified abdominal pain	10	0.2528
				Abdominal pain, other specified site	9	0.1721
				Abdominal pain	9	0.0007
				Vesical tenesmus	10	0.0001
Chest pain	8	0.375	0.45	Unspecified chest pain	9	0.3551
				Other chest pain	9	0.1407
				Chest pain, unspecified	10	0.1264
				Other chest pain	10	0.0651
				Pleurodynia	10	0.0126
				Chest pain on breathing	10	0.0048
				Intercostal pain	10	0.0017
				Chest pain	9	0.0002

Table 1: ICD-9/ICD-10 to SNOMED mappings for the most prevalent SNOMED features in the EPI cohort. SNOMED mapping aggregated clinically related ICD-9/10 codes to produce features that span the ICD-10 transition.

positive by the model, i.e. of the actual EPI patients, what fraction did the model label as positive?

Figure 4 displays the precision-recall curves for strategies 1 and 2. Models were trained using SNOMED (737), clinical expert (93) and naïve data driven (individual ICD concepts - 715) features. For strategy 1, at 20% recall, precision was 71%, 71% and 68% for SNOMED, clinically-driven and naïve data-driven respectively. For strategy 2, at 20% recall, precision was 51%, 51% and 7% for SNOMED, clinically-driven and naïve data-driven respectively. With strategy 1, the model is trained on patients who have data encoding using both ICD-9 and -10. This allows the model to perform well on the validation data even on the naïve data-driven features as the validation patients also include both ICD versions. This is no longer the case for strategy 2, where the model is validated on patients whose medical history is exclusively encoded using ICD-10. Here, the value of our SNOMED based solution compared to the standard data-driven approach is clear. Successful mapping of ICD-9 and -10 concepts allowed a model trained on ICD-9 information to extrapolate its predictions to patients in the ICD-10 space.

The patterns of patient care captured by modeling using SNOMED mapping and clinically-driven features were similar. The top 5 features contributing to model gain under each feature set in strategy 1 related to conditions of the pancreas or gastrointestinal system. "Disorder of the Pancreas" was the most impactful feature representing 30% of model gain for SNOMED mapped features and 19% of model gain for clinically-driven features and overlapped in 6 of 26 ICD-9/10 codes. Our SNOMED based solution therefore appears capable of identify features that approximate the clinical relevance and predictive impact of features defined by experts, while being considerably more time and cost efficient. In our case, the

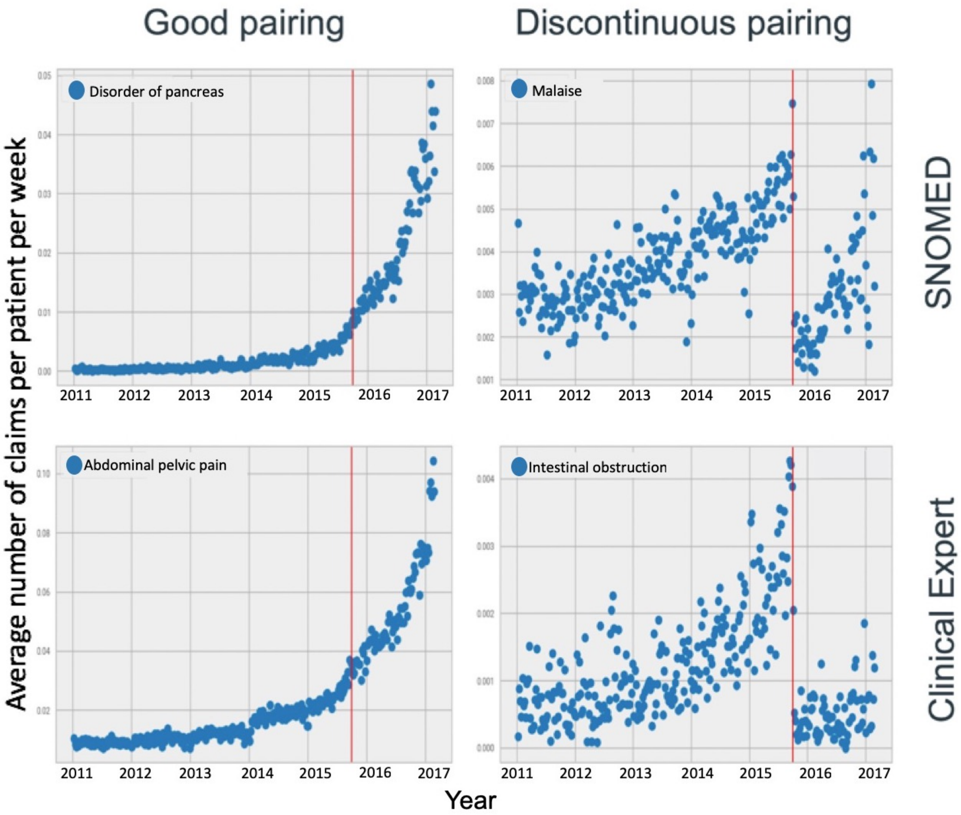


Figure 3: Average number of claims per patient per week for our EPI cohort with examples of good pairings (left) and discontinuous pairings (right), for SNOMED (top) and clinically coded (bottom) features. The ICD-10 transition date is denoted by the red line. Discontinuous pairings are typified by an abrupt shift in average weekly patient claims at the time of ICD-10 transition.

clinical experts took several weeks to define the clinical-expert features, while running our SNOMED-based algorithm took a couple of hours.

4 CONCLUSIONS

Our SNOMED-based mapping algorithm can be used to automatically incorporate clinical domain knowledge into an empirically-orientated approach. The solution does not compromise on accuracy and removes the need for time-consuming manual mapping. It provides two key advantages over existing strategies. First, it automates the ICD ontology interoperability by mapping ICD codes to shared SNOMED concepts. Second, it automates mapping of orphan ICD codes to relevant SNOMED concepts that contain both ICD-9/10 codes, reducing the percentage of ICD-9 and ICD-10 orphans to $< 1\%$, providing better coverage than GEMs (3% ICD-9 and 1% ICD-10 fail to match).

SNOMED mapping generated features with comparable stability to those created by clinical experts when evaluated over the ICD-10 transition. The prevalence of the top features within the EPI patient population were comparable suggesting that each approach captured the most representative clinical definitions. While both approaches generated a subset of unstable features, those defined manually by experts included a greater number of ICD codes. This may have helped minimize variations in feature prevalence since such features may capture broader clinical definitions and therefore be more stable across the ontology transition. By contrast, our algorithm mapped directly to the closest shared SNOMED concept within the SNOMED graph using a limited search radius and selected concepts without regard to SNOMED hierarchy. Potential improvements to our algorithm might therefore be achieved by rolling-up ICD pairings toward higher-level SNOMED concepts. This strategy may minimize feature instability since higher level ICD aggregates may be less affected by imperfect pairings when compared to more granular features. A further benefit of this strategy is that higher-level aggregates may result in more clinically interpretable features for predictive modeling and patient journey profiling.

Several additional modifications to our algorithm are worth exploring in future work. First, a portion of SNOMED mapped features resulted in discontinuous pairings. These patterns of discontinuity point to features that are dominated by a single ICD version. Such discontinuous pairings might be minimized by informing our search algorithm with the prevalence of claim counts within the patient cohort. This strategy would allow the search algorithm to find the

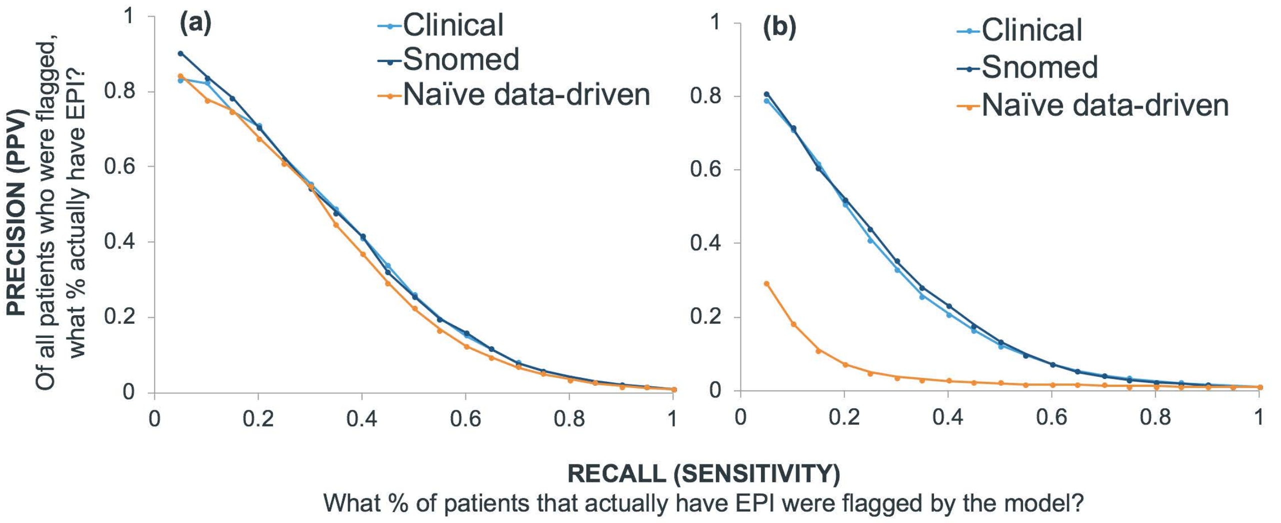


Figure 4: Precision-Recall curve when (a) using Strategy 1, (b) using Strategy 2. Precision and recall were comparable using either SNOMED derived features or clinically defined features in Strategy 2 i.e. when models were trained on patient data prior to the ICD-10 transition and validated on patient data following the transition.

appropriate level of aggregation during the reduction step to create a stable transition between ICD-9 and -10. Second, our algorithm traversed up and down the SNOMED graph hierarchy as opposed to horizontally via interconnected trees. This ensured that all orphan ICDs map to common ancestor or child concepts. However, feature stability and/or interpretability might be improved by mapping only to ancestor concepts via an upward traversal. This might encourage more clinically appropriate ICD pairings since mapped concepts may be more encompassing. Finally, a maximum search radius of 5 was used to resolve orphan ICDs. This threshold was chosen since an overly exhaustive search may result in nonsensical ICD pairings while a conservative search may fail to detect appropriate pairings and/or result in no detectable pairings. A complete analysis of feature stability and interpretability using various search parameters should be explored to determine an optimal search strategy.

Features derived from our SNOMED mapping algorithm showed no improvement over naïve data-driven or expert defined features when models were trained and validated on patient data spanning the 2015 ICD-10 transition. This is likely due to each model being exposed to sufficient patient data from each ontology. Training across the transition may allow XGBoost to learn to avoid naïve features that generalized poorly across ontologies. Additionally, features derived from SNOMED mapping or clinical experts with high discontinuity may suffer from elevated noise and diminished predictive performance and may be similarly avoided. By contrast, SNOMED mapping markedly outperformed the naïve data-driven approach and performed comparably to expert features when models were trained on ICD-9 data and validated on ICD-10 data. This suggests the SNOMED concepts identified by our algorithm generalize across the ICD-10 transition making it a promising tool for automated ICD ontology interoperability and clinical feature engineering for machine learning applications.

A benefit to the use of SNOMED as a common clinical ontology for predictive modeling is that it is compatible with a variety of clinical data sources. For instance, our SNOMED search algorithm should be compatible with future ICD revisions such as ICD-11, which is expected as early as 2022[11]. Automated mapping strategies may become increasingly valuable near the time of ICD transition since domain experts familiar with the newest ICD ontology who are available to support machine learning efforts may be in short supply. Furthermore, other clinical terminologies may be mapped to SNOMED such as lab and procedure claim codes (i.e. LOINC and CPT), and the use of a common ontology may help simplify feature engineering and model interpretation efforts when using diverse clinical data sources.

Automating ICD interoperability would boost the application of machine learning in cases where manual expert mapping is not available or prohibitive. This should relieve barriers to entry for machine learning healthcare practitioners since it does not require the sponsorship of clinical coding experts. Furthermore, in many instances and especially in the realm of rare diseases, predictive performance is hindered by a limited supply of diagnosed patients for model training. Our approach may thus serve to increase the number of eligible patients by promoting the inclusion of patients whose relevant clinical histories predate the ICD-10 transition. Leveraging both ICD ontologies may also increase eligible patient counts for modeling efforts that require a minimal duration of patient medical history. Finally, our approach supports the meaningful use of ICD claims data for predictive modeling, which is a particularly advantageous data source due to its high patient coverage and relative universality. As increased cohort size, longer patient histories, and better clinical data utilization result in richer more representative data sets, our approach supports the development of more statistically robust and performant predictive models.

REFERENCES

[1] American Medical Association. 2012. *Crosswalking Between ICD-9 and ICD-10.* Retrieved September 1, 2019 from https://www.nationalfamilyplanning.org/document.doc?id=780

[2] A.D. Boyd, J.J. Li, M.D. Burton, M. Jonen, V. Garden, I. Anchor, R.Q. Luo, I. Zenku, N. Bahroos, S.B. Brown, T. Vanden Hoek, and Y.A. Lussier. 2013. The discriminatory cost of ICD-10-CM transition between clinical specialties: metrics, case study, and mitigating tools. *J Am Med Inform Assoc.* 20, 4 (2013), 708–717. https://doi.org/10.1136/amiajnl-2012-001358

[3] Donna J. Cartwright. 2013. ICD-9-CM to ICD-10-CM Codes: What? Why? How? *Advances in Wound Care* 2, 10 (2013), 588–592. https://doi.org/10.1089/wound.2013.0478

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).* ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[5] Centers for Medicare and Medicaid Services (CMS). 2014. *2014 ICD-10-CM and GEMs.* Retrieved December 18, 2019 from https://www.cms.gov/Medicare/Coding/ICD10/2014-ICD-10-CM-and-GEMs

[6] K.W. Fung, R. Richesson, M. Smerek, K.C. Pereira, B.B. Green, A. Patkar, M. Clowse, A. Bauck, and O. Bodenreider. 2016. Preparing for the ICD-10-CM Transition: Automated Methods for Translating ICD Codes in Clinical Phenotype Definitions. *EGEMS* 4, 1 (2016), 1211. https://doi.org/10.13063/2327-9214.1211

[7] Kevin Heslin, Pamela Owens, Zeynal Karaca, Marguerite Barrett, Brian Moore, and Anne Elixhauser. 2017. Trends in Opioid-related Inpatient Stays Shifted After the US Transitioned to ICD-10-CM Diagnosis Coding in 2015. *Medical Care* 55, 11 (2017), 918–923. https://doi.org/10.1097/MLR.0000000000000805

[8] SNOMED International. 2019. *SNOMED CT & Other Terminologies, Classifications & Code Systems.* Retrieved September 1, 2019 from https://www.snomed.org/snomed-ct/sct-worldwide

[9] SNOMED International. 2020. *SNOMED Technical Implementation Guide.* Retrieved January 7, 2020 from https://confluence.ihtsdotools.org/display/DOCTIG/3.1.3.+Relationships

[10] Lolita Jones and Stanley Nachimson. 2019. *Use Caution When Entering the Crosswalk: A Warning About Relying on GEMs as Your ICD-10 Solution.* Retrieved December 18, 2019 from http://www.cms.org/uploads/ICDLogicGEMSWhitePaper.pdf

[11] World Health Organization. 2019. *International Classification of Diseases, 11th Revision (ICD-11).* Retrieved September 1, 2019 from https://www.who.int/classifications/icd/en/

[12] Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* 10, 3 (2015). https://doi.org/10.1371/journal.pone.0118432

[13] Observational Health Data Sciences and Informatics. 2019. *ATHENA – OHDSI Vocabularies Repository.* Retrieved September 1, 2019 from http://athena.ohdsi.org

[14] Observational Health Data Sciences and Informatics. 2019. *Standardized Vocabularies of the OMOP Common Data Model.* Retrieved September 1, 2019 from https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:introduction

[15] Christine C. Stewart, Christine Y. Lu, Tae K. Moon, Karen J. Coleman, Phillip M. Crawford, Matthew D. Lakoma, and Gregory E. Simon. 2019. Impact of ICD-10-CM Transition on Mental Health Diagnoses Recording. *EGEMS* 7, 1 (2019), 14. https://doi.org/10.5334/egems.281

[16] N.K. Venepalli, A. Shergill, P. Dorestani, and A.D. Boyd. 2014. Conducting Retrospective Ontological Clinical Trials in ICD-9-CM in the Age of ICD-10-CM. *Cancer Informatics* 13 (2014), 81–88. https://doi.org/10.4137/FCIN.S14032