Writing habits and telltale neighbors: analyzing clinical concept usage patterns with sublanguage embeddings

Denis Newman-Griffis[a,b] and **Eric Fosler-Lussier** [a]
[a]Dept of Computer Science and Engineering, The Ohio State University, Columbus, OH
[b]Rehabilitation Medicine Dept, Clinical Center, National Institutes of Health, Bethesda, MD
{newman-griffis.1, fosler-lussier.1} @ osu.edu

Abstract

Natural language processing techniques are being applied to increasingly diverse types of electronic health records, and can benefit from in-depth understanding of the distinguishing characteristics of medical document types. We present a method for characterizing the usage patterns of clinical concepts among different document types, in order to capture semantic differences beyond the lexical level. By training concept embeddings on clinical documents of different types and measuring the differences in their nearest neighborhood structures, we are able to measure divergences in concept usage while correcting for noise in embedding learning. Experiments on the MIMIC-III corpus demonstrate that our approach captures clinically-relevant differences in concept usage and provides an intuitive way to explore semantic characteristics of clinical document collections.

1 Introduction

Sublanguage analysis has played a pivotal role in natural language processing of health data, from highlighting the clear linguistic differences between biomedical literature and clinical text (Friedman et al., 2002) to supporting adaptation to multiple languages (Laippala et al., 2009). Recent studies of clinical sublanguage have extended sublanguage study to the document type level, in order to improve our understanding of the syntactic and lexical differences between highly distinct document types used in modern EHR systems (Feldman et al., 2016; Grön et al., 2019).

However, one key axis of sublanguage characterization that has not yet been explored is how domain-specific clinical *concepts* differ in their usage patterns among different document types. Established biomedical concepts may have multiple, often non-compositional surface forms (e.g., "ALS" and "Lou Gehrig's disease"), making them difficult to analyze using lexical occurrence alone. Understanding how these concepts differ between document types can not only augment recent methods for sublanguage-based text categorization (Feldman et al., 2016), but also inform the perennial challenge of medical concept normalization (Luo et al., 2019): "depression" is much easier to disambiguate if its occurrence is known to be in a social work note or an abdominal exam.

Inspired by recent technological advances in modeling diachronic language change (Hamilton et al., 2016; Vashisth et al., 2019), we characterize concept usage differences within clinical sublanguages using nearest neighborhood structures of clinical concept embeddings. We show that overlap in nearest neighborhoods can reliably distinguish between document types while controlling for noise in the embedding process. Qualitative analysis of these nearest neighborhoods demonstrates that these distinctions are semantically relevant, highlighting sublanguage-sensitive relationships between specific concepts and between concepts and related surface forms. Our findings suggest that the structure of concept embedding spaces not only captures domain-specific semantic relationships, but can also identify a "fingerprint" of concept usage patterns within a clinical document type to inform language understanding.

2 Related Work

Sublanguage analysis historically focused on describing the characteristic grammatical structures of a particular domain (Friedman, 1986; Grishman, 2001; Friedman et al., 2002). As methods for automated analysis of large-scale data sets have improved, more studies have investigated lexical and semantic characteristics, such as usage patterns of different verbs and semantic categories

Type	Docs	Lines	Tokens	Matches	Concepts	High Confidence	
						Concepts	Consistency (%)
Case Management	967	20,106	165,608	45,306	557	111	75
Consult	98	15,514	96,515	26,109	812	0	–
Discharge Summary	59,652	14,480,154	104,027,364	30,840,589	6,381	1,599	67
ECG	209,051	1,022,023	7,307,381	2,163,682	540	14	56
Echo	45,794	2,892,069	19,752,879	6,070,772	1,233	157	65
General	8,301	307,330	2,191,618	552,789	2,559	0	–
Nursing	223,586	9,839,274	73,426,426	18,903,892	4,912	2	58
Nursing/Other	822,497	10,839,123	140,164,545	31,135,584	5,049	83	60
Nutrition	9,418	868,102	3,843,963	1,147,918	1,911	198	73
Pharmacy	103	4,887	39,163	8,935	376	0	–
Physician	141,624	26,659,749	148,306,543	39,239,425	5,538	122	57
Radiology	522,279	17,811,429	211,901,548	34,433,338	4,126	599	63
Rehab Services	5,431	585,779	2,936,022	869,485	2,239	9	62
Respiratory	31,739	1,323,495	6,358,924	2,255,725	1,039	5	63
Social Work	2,670	100,124	930,674	195,417	1,282	0	–

Table 1: Document type subcorpora in MIMIC-III. Tokenization was performed with SpaCy; Matches and Concepts refer to number of terminology string match instances and number of unique concepts embedded, respectively, using SNOMED-CT and LOINC vocabularies from UMLS 2017AB release. The number of high-confidence concepts identified for each document type is given with their mean consistency.

(Denecke, 2014), as well as more structural information such as document section patterns and syntactic features (Zeng et al., 2011; Temnikova et al., 2014). The use of terminologies to assess conceptual features of a sublanguage corpus was proposed by Walker and Amsler (1986), and Drouin (2004); Grön et al. (2019) used sublanguage features to expand existing terminologies, but large-scale characterization of concept usage in sublanguage has remained a challenging question.

Word embedding techniques have been utilized to describe diachronic language change in a number of recent studies, from evaluating broad changes over decades (Hamilton et al., 2016; Vashisth et al., 2019) to detecting fine-grained shifts in conceptualizations of psychological concepts (Vylomova et al., 2019). Embedding techniques have also been used as a mirror to analyze social biases in language data (Garg et al., 2018). Similar to our work, Ye and Fabbri (2018) investigate document type-specific embeddings from clinical data as a tool for medical language analysis. However, our approach has two significant differences: Ye and Fabbri (2018) used word embeddings only, while we utilize concept embeddings to capture concepts across multiple surface forms; more importantly, their work investigated multiple document types as a way to *control* for specific usage patterns within sublanguages in order to capture more general term similarity patterns, while our study aims to *capture* these sublanguage-specific usage patterns in order to analyze the representative differences in language use between different expert communities.

3 Data and preprocessing

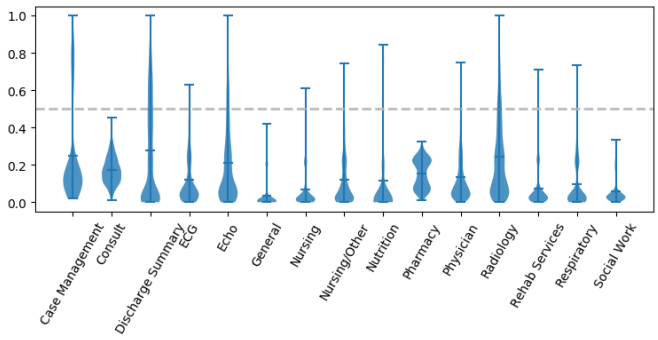
We use free text notes from the MIMIC-III critical care database (Johnson et al., 2016) for our analysis. This includes approximately 2 million text records from hospital admissions of almost 50 thousand patients to the critical care units of Beth Israel Deaconess Medical Center over a 12-year period. Each document belongs to one of 15 document types, listed in Table 1.

As sentence segmentation of clinical text is often optimized for specific document types (Griffis et al., 2016), we segmented our documents at linebreaks and tokenized using SpaCy (version 2.1.6; Honnibal and Montani 2017). All tokens were lowercased, but punctuation and deidentifier strings were retained, and no stopwords were removed.

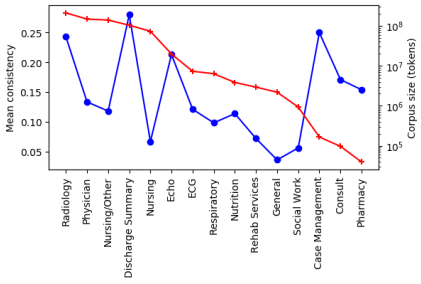
4 Experiments

Methods for learning clinical concept representations have proliferated in recent years (Choi et al., 2016; Mencia et al., 2016; Phan et al., 2019), but often require annotations in forms such as billing codes or disambiguated concept mentions. These annotations may be supplied by human experts such as medical coders, or by adapting medical NLP tools such as MetaMap (Aronson and Lang, 2010) or cTAKES (Savova et al., 2010) to perform concept recognition (De Vine et al., 2014).

For investigating potentially divergent usage patterns of clinical concepts, these strategies face



(a) Self consistency by document type; line at 50% threshold



(b) Self consistency compared to corpus size (log scale), with document types sorted by decreasing corpus size.

Figure 1: Distribution of self-consistency rates (i.e., overlap in nearest neighbors between replicate embeddings of the same concept) among MIMIC document types.

serious limitations: the full diversity of MIMIC data has not been annotated for concept identifiers, and the statistical biases of trained NLP tools may suppress underlying differences in automatically-recognized concepts. We therefore take a distant supervision approach, using JET (Newman-Griffis et al., 2018). JET uses a sliding context window to jointly train embedding models for words, surface forms, and concepts, using a log-bilinear objective with negative sampling and shared embeddings for context words. It leverages known surface forms from a terminology as a source of distant supervision: each occurrence of any string in the terminology is treated as a weighted training instance for each of the concepts that string can represent. As terminologies are typically many-to-many maps between surface forms and concepts, this generally leads to a unique set of contexts being used to train the embedding of each concept, though any individual context window may be used as a sample for training multiple concepts. We constrain the scope of our analysis to only concepts and strings from SNOMED-CT and LOINC,[1] two popular high-coverage clinical vocabularies.

4.1 Identifying concepts for comparison

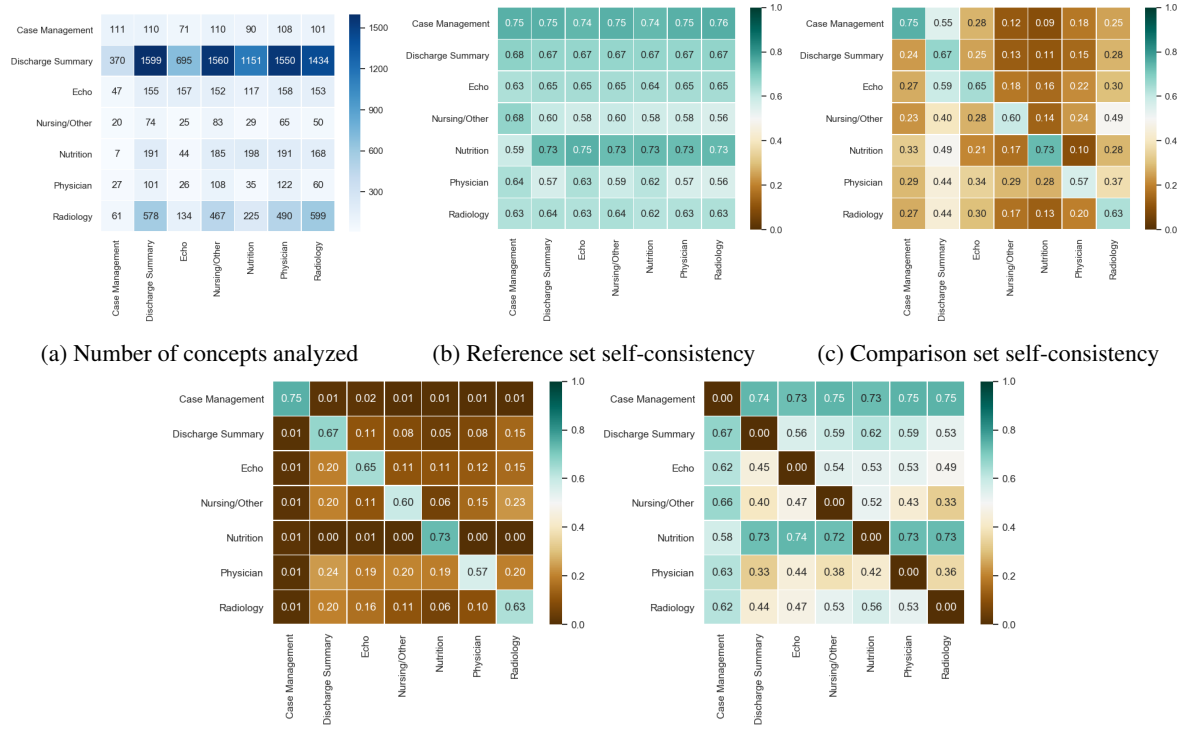
For each document type, we concatenate all of its documents (maintaining linebreaks), identify all occurrences of SNOMED-CT and LOINC strings in each line, and use these occurrences to train word, term, and concept embeddings with JET. Due to the size of our subcorpora, we used a window size of 5, minimum frequency of 5, embedding dimensionality of 100, initial learning rate of 0.05, and 10 iterations over each corpus.

Prior research has noted instability of nearest neighborhoods in multiple embedding methods (Wendlandt et al., 2018). We therefore train 10 sets of embeddings from each of our subcorpora, each using the same hyperparameter settings but a different random seed. We then use all 10 replicates from each subcorpus in our analyses, in order to control for variation in nearest neighborhoods introduced by random initialization and negative sampling. To evaluate the baseline reliability of concept embedding neighborhoods from each subcorpus, we calculated per-concept consistency by measuring, over all pairs of embedding sets within the 10 replicates, the average set membership overlap between the top 5 nearest neighbors by cosine similarity for each concept embedding.[2] As shown in Figure 1a, these consistency scores vary widely both within and between document types, with some document types producing no concept embeddings with consistency over 40%. Interestingly, as illustrated in Figure 1b, there is no linear relationship between log corpus size and mean concept consistency ($R^2 \approx 0.011$), suggesting that low consistency is not solely due to limited training data.

To mitigate concerns about the reliability of embeddings for comparison, a set of **high-confidence concepts** is identified for each document type by

[1] We used the versions distributed in the 2017AB release of the UMLS (Bodenreider, 2004).

[2] We chose five nearest neighbors for our analyses based on qualitative review of neighborhoods for concepts within different document types. We found nearest neighborhoods for concept embeddings to vary more than for word embeddings, often introducing noise beyond the top five nearest neighbors; we therefore set a conservative baseline for reliability by focusing on the closest and most stable neighbors. However, using 10 neighbors, as Wendlandt et al. (2018) did, or more could yield different qualitative patterns in document type comparisons and bears exploration.



(a) Number of concepts analyzed (b) Reference set self-consistency (c) Comparison set self-consistency

(d) Cross-type consistency (e) Consistency deltas

Figure 2: Comparison of concept neighborhood consistency statistics across document types, using high-confidence concepts from the reference type. Figure 2a provides the number of concepts shared between the high-confidence reference set and the comparison set. All values are the mean of the consistency distribution calculated over all concepts analyzed for the document type pair.

retaining only those with a self-consistency of at least 50%; Table 1 includes the number of high-confidence concepts identified and the mean consistency among this subset.[3] These embeddings capture reliable concept usage information for each document type, and form the basis of our comparative analysis.

4.2 Cross-corpus analysis

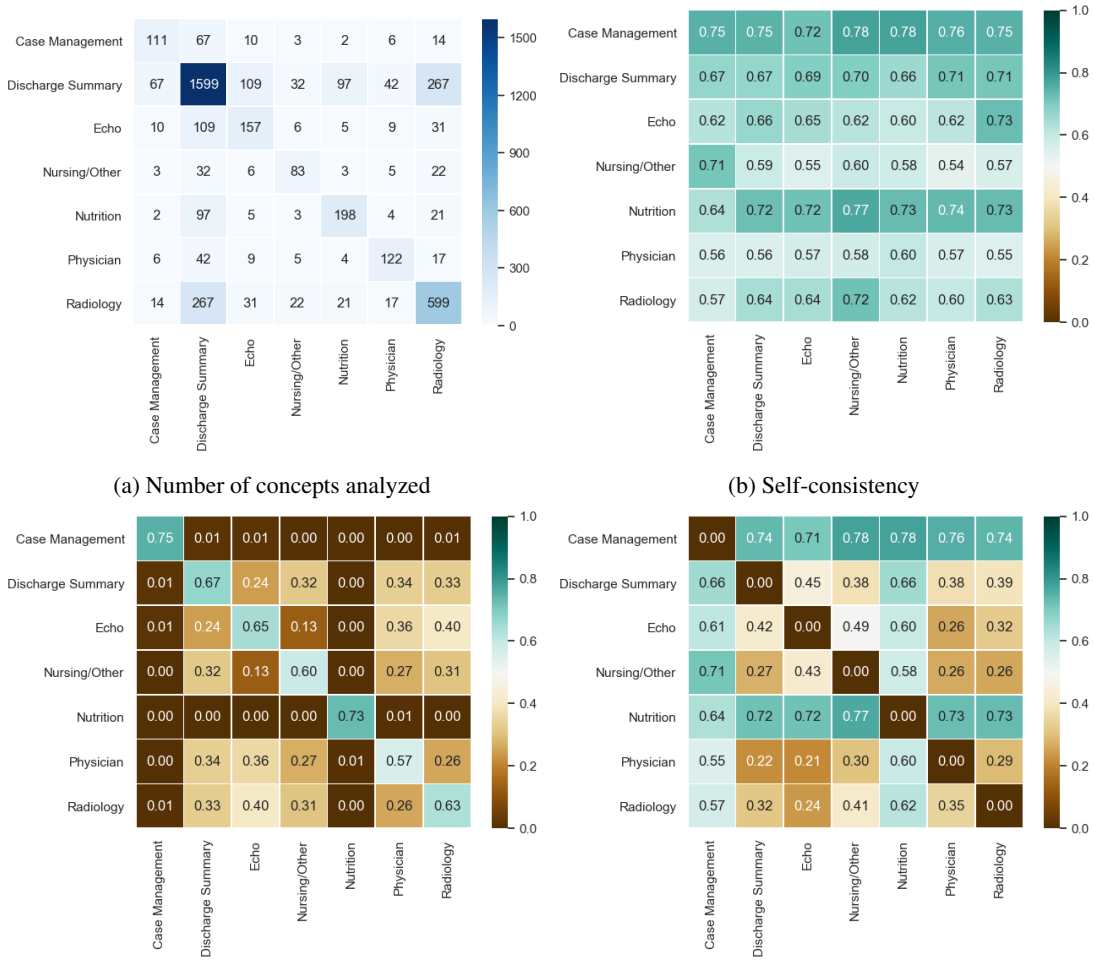
Our key question is what concept embeddings reveal about clinical concept usage *between* document types. To maintain a sufficient sample size, we restrict our comparison to the 7 document types with at least 50 high confidence concepts: *Case Management*, *Discharge Summary*, *Echo*, *Nursing/Other*, *Nutrition*, *Physician*, and *Radiology*. *Physician*, *ECG*, and *Nursing* were also used by Feldman et al. (2016) for their lexicosyntactic analysis, although they combined *Nursing* documents (longer narratives) and *Nursing/Other* (which tend to be much shorter) into a single set, while we retain the distinction. Interestingly, the fourth type they analyzed, *ECG*, produced only 14 high-confidence concepts in our analysis, suggesting high semantic variability despite the large number of documents.

As learned concept sets differ between document types, the first step for comparing a document type pair is to identify the set of concepts embedded for both. For reference type A and comparison type B, we identify high-confidence concepts from A that are also present in B, and calculate four distributions using this shared set:

Reference consistency: self-consistency across each of the shared concepts, using only other shared concepts to identify nearest neighborhoods in embeddings for the reference set.

Comparison consistency: self-consistency of each shared concept in embeddings for the comparison document type, again using only shared concepts for neighbors. As the shared set is based on high-confidence concepts from the reference

[3] We found in our analysis that most concept consistency numbers clustered roughly bimodally, between 0-30% or 60-90%; this is reflected at a coarse level in the overall distributions in Figure 1a. Varying the threshold outside of these ranges did not have a significant impact on the number of concepts retained; the 50% threshold was chosen for simplicity. With larger corpora, yielding higher concept coverage, a higher threshold could be chosen for a stricter analysis.



(a) Number of concepts analyzed

(b) Self-consistency

(c) Cross-type consistency

(d) Consistency deltas

Figure 3: Concept neighborhood consistency statistics, restricted to concepts that are high-confidence in both reference and comparison sets. In this case, reference self-consistency and target self-consistency are symmetric, so only reference self-consistency is presented in Figure 3b.

set, this is not symmetric with reference consistency (as the high-confidence sets may differ).

Cross-type consistency: average consistency for each shared concept calculated over all pairs of replicates (i.e., comparing the nearest neighbors of all 10 reference embedding sets to the nearest neighbors in all 10 comparison embedding sets).

Consistency deltas: the difference, for each shared concept, between its reference self-consistency and its cross-type consistency. This provides a direct evaluation of how distinct the concept usage is between two document types, where a high delta indicates highly distinct usage.

Mean values for these distributions are provided for each pair of our 7 document types in Figure 2. Comparing Figures 2b and 2c, it is clear that high-confidence concepts for one document type are typically not high-confidence for another. Most document type pairs show fairly strong divergence, with deltas ranging from 30-60%. *Physician* notes have comparatively high cross-set consistency of around 20% for their high-confidence concepts, likely reflecting the all-purpose nature of these documents, which include patient history, medications, vitals, and detailed examination notes. Interestingly, *Case Management* and *Nutrition* are starkly divergent from other document types, with near-zero cross-set consistency and comparatively high self-consistency of over 70% in the compared concept sets, despite a relatively high overlap between their high-confidence sets and concepts learned for other document types.

In order to control for the low overlap between high-confidence sets in different document types, we also re-ran our consistency analysis restricted to only concepts that are high-confidence in *both* the reference and comparison sets. As shown in Figure 3, this yields considerably smaller concept

Query	Discharge Summary	Nursing/Other	Radiology
Diabetes Mellitus (C0011849)	Diabetes (C0011847)	Gestational Diabetes (C0085207)	Poorly controlled (C3853134)
	Type 2 (C0441730)	A2 immunologic symbol (C1443036)	Insulin (C0021641)
	Type 1 (C0441729)	Diabetes Mellitus, Insulin-Dependent (C0011854)	Diabetes Mellitus, Insulin-Dependent (C0011854)
	Gestational Diabetes (C0085207)	Factor V (C0015498)	Diabetes Mellitus, Non-Insulin-Dependent (C0011860)
	Diabetes Mellitus, Insulin-Dependent (C0011854)	A1 immunologic symbol (C1443035)	Stage level 5 (C0441777)
	Discharge Summary	Echo	Radiology
Mental state (C0278060)†	Coherent (C4068804)	Donor:Type:Point in time:^Patient:Nominal (C3263710)	Mental status changes (C0856054)
	Confusion (C0009676)	Donor person (C0013018)	Abnormal mental state (C0278061)
	Respiratory status:-:Point in time:^Patient:- (C2598168)	Respiratory arrest (C0162297)	Level of consciousness (C0234425)
	Respiratory status (C1998827)	Organ donor:Type:Point in time:^Donor:Nominal (C1716004)	Level of consciousness:Find:Pt:^Patient:Ord (C4050479)
	Abnormal mental state (C0278061)	Swallowing G-code (C4281783)	Mississippi (state) (C0026221)

Table 2: 5 nearest neighbor concepts to *Diabetes Mellitus* and *Mental state* from 3 high-confidence document types, averaging cosine similarities across all replicate embedding sets within each document type. †The two nearest neighbors to *Mental state* for all three document types were two LOINC codes using the same "mental status" string; they are omitted here for brevity.

sets for comparison, with single-digit overlap for 18/42 non-self pairings. Cross-set consistency increases somewhat, most significantly for pairings involving *Physician* or *Radiology*; however, no consistency delta falls below 20% for any non-self pair, indicating that concept neighborhoods remain distinct even within high-confidence sets.

4.3 Qualitative neighborhood analysis

Analysis of neighborhood consistency enables measuring divergence in the contextual usage patterns of clinical concepts; however, this divergence could be due to spurious or semantically uninformative correlations instead of clinically-relevant distinctions in concept similarities. To confirm that our methodology captures informative distinctions in concept usage, we qualitatively review example neighborhoods. To mitigate variability of nearest neighborhoods in embedding spaces, we identify a concept's *qualitative* nearest neighbors for a given document type by calculating its pairwise cosine distance vectors for all 10 replicates in that document type and taking the k neighbors with lowest average distance.

As with our consistency analyses, we focus on the neighborhoods of high-confidence concepts, although we do not filter the neighborhoods themselves. Of all high-confidence concepts identified in our embeddings, only two were high-confidence in 5 different document types, and these were highly generic concepts: *Interventional procedure* (C0184661) and a corresponding LOINC code (C0945766). Seven concepts were high-confidence for 4 document types; of these, two were generic procedure concepts, two were concepts for the broad gastrointestinal category, and three were versions of body weight. For a diversity of concepts, we therefore turned to the 75 concepts that were high-confidence within 3 document types. We reviewed each of these concepts, and describe our findings for three of the most broadly clinically-relevant below.

Diabetes Mellitus (C0011849) *Diabetes Mellitus* (search strings: "diabetes mellitus" and "diabetes mellitus dm") was high-confidence in *Discharge Summary*, *Nursing/Other*, and *Radiology* document types; Table 2 gives the top 5 neighbors from each type. These neighbors are semantically consistent across document types: more specific diabetes-related concepts, related biological fac-

tors; continuing down the nearest neighbors list yields related symptoms and comorbidities such as *Irritable Bowel Syndrome* (C0022104) and *Gastroesophageal reflux disease* (C0017168).

Memory loss (C0751295) *Memory loss* (search string: "memory loss") was also high-confidence in *Discharge Summary*, *Nursing/Other*, and *Radiology* documents. For brevity, its nearest neighbors are omitted from Table 2, as there is little variation among the top 5. However, the next neighbors (at only slightly greater cosine distance) vary considerably across document types, while remaining highly consistent within each individual type. In *Discharge Summary*, more high-level concepts related to overall function emerge, such as *Functional status* (C0598463), *Relationships* (C0439849), and *Rambling* (C4068735). *Radiology* yields more symptomatically-related neighbors: *Aphagia* (C0221470) is present in both, but *Radiology* includes *Disorientation* (C0233407), *Delusions* (C0011253), and *Gait, Unsteady* (C0231686). Finally, *Nursing/Other* finds concepts more related to daily life, such as *Cigars* (C0678446) and *Multifocals* (C3843228), though at a greater cosine distance than the other document types (Figure 4).

Mental state (C0278060) *Mental state* (search strings: "mental status", "mental state") was high-confidence in *Discharge Summary*, *Echo*, and *Radiology*, and highlighted an unexpected consequence of relying on the Distributional Hypothesis (Harris, 1954) for semantic characterization in sublanguage-specific corpora. The top 5 nearest neighbors (excluding two trivial LOINC codes for the same concept, also using the "mental status" search string) are given in Table 2. In *Discharge Summary* documents, "mental status" is typically referred to in detailed patient narratives, medication lists, and the like, and this yields semantically-reasonable nearest neighbors such as *Confusion* (C0009676) and *Coherent* (C4068804).

In *Echo* documents, however, "mental status" occurs most frequently within an "Indication" field of the "PATIENT/TEST INFORMATION" section. Two common patterns emerge in "Indication" texts: references to altered or reduced mental status, or patients who are vegetative and being evaluated for organ donor eligibility. Though "mental status" and "organ donor" do not co-occur, their consistent occurrence in the same contextual structures leads to extremely similar embeddings (see Figure 4). A similar issue occurs in *Radiology* notes, where the "MEDICAL CONDITION" section includes several instances of elderly patients presenting with either hypothermia or altered mental status; as a result, two hypothermia concepts (C1963170 and C0020672) are in the 10 nearest neighbors to *Mental state*.

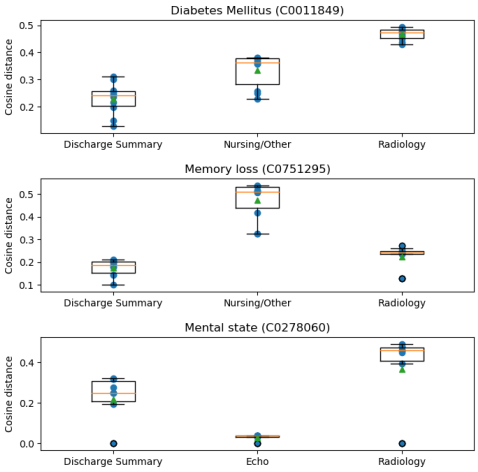


Figure 4: Cosine distance distribution of three concepts to their 10 nearest neighbors, averaged across document type replicate embeddings.

Results from *Radiology* also highlight one limitation of distant supervision for learning concept embeddings: as the word "state" is polysemous, including a geopolitical entity, geographical concepts such as *Mississippi* (C0026221) end up with similar embeddings to *Mental state*. A similar issue occurs in the neighbors for *Memory loss*; due to string polysemy, the concept *CIGAR string - sequence alignment* (C4255278) ends up with a similar embedding to *Cigars* (C0678446).

4.4 Nearest surface form embeddings

As JET learns embeddings of concepts and their surface forms jointly in a single vector space, we also analyzed the surface forms embeddings nearest to different concepts. This enabled us both to evaluate the semantic congruence of surface form and concept embeddings, and to further delve into corpus-specific contextual patterns that emerge in the vector space. As with our concept neighborhood analysis, for each of our 10 replicate embeddings in each document type, we calculated the cosine distance vector from each high-confidence concept to all of the term embeddings in the same replicate, and then averaged these distance vectors to identify neighbors robust to embedding noise. Table 3 presents surface form neighbors identified

Query	Discharge Summary	Nutrition	Case Management
Community (C0009462)	Community	Dilute	Substance
	Health center	Social work	Monitoring
	Acquired	Surgical site	Somewhat
	Residence	In situ	Hearing
	Nursing facility	Nephritis	Speech
	Discharge Summary	Echo	ECG
ECG (C0013798)	ECG	ECG	ECG
	EKG	Exercise	Physician
	Sinus tachycardia	Stress	Last
	Sinus bradycardia	Fair	No change
	Right bundle branch block	Specific	Abnormal
	Discharge Summary	Echo	Radiology
Blood pressure (C0005823)	Blood pressure	Blood pressure	Blood pressure
	Heart rate	Heart rate	Heart rate
	Pressure	Rate	Rate
	Systolic blood pressure	Exercise	Method
	Rate	Stress	Exercise

Table 3: 5 nearest neighbor surface forms to three frequent clinical concepts, across document types for which they are high-confidence.

for three high-confidence clinical concepts chosen for clinical relevance and wide usage; these concepts are discussed in the following paragraphs.

Blood pressure (C0005823) *Blood pressure* is high-confidence in *Discharge Summary*, *Echo*, and *Radiology* documents. It is a key concept that is measured frequently in various settings; intuitively, it is a sufficiently core concept that it should exhibit little variance. Its neighbor surface forms indeed indicate fairly consistent use across the three document types, referencing both related measurements ("heart rate") and related concepts ("exercise" and "stress").

Echocardiogram (C0013798) *Echocardiogram* is high-confidence in *Discharge Summary*, *Echo* (detailed summaries and interpretation written after the ECG), and *ECG* (technical notes taken during the procedure) documents. ECGs are common, and are performed for various purposes and discussed in varying detail. Interestingly, neighbor surface forms in *Discharge Summary* embeddings reflect specific pathologies, potentially capturing details determined post diagnosis and treatment. In *Echo* embeddings, the neighbors are more general surface forms evaluating the findings ("fair") and relevant history/symptoms that led to the ECG ("exercise", "stress"). *ECG* embeddings reflect their more technical nature, with surface forms such as "no change" and "abnormal" yielding high similarity.

Community (C0009462) *Community* is a very broad concept and a common word, and is discussed primarily in documents concerned with whole-person health; it is high confidence in *Discharge Summary*, *Nutrition*, and *Case Management* documents. Each of these document types reflects different usage patterns. The nearest surface forms in *Discharge Summary* embeddings reflect a focus on living conditions, referring to "health center", "residence", and "nursing facility". In *Nutrition* documents, *Community* is discussed primarily in terms of "community-acquired pneumonia", likely leading to more treatment-oriented neighbor surface forms. Finally, in *Case Management* embeddings, nearby surface forms reflect discussion of specific risk factors or resources ("substance", "monitoring") to consider in maintaining the patient's health and responding to their specific needs (e.g., "hearing", "speech").

5 Discussion

We have shown that concept embeddings learned from different clinical document type corpora reveal characteristics of how clinical concepts are used in different settings. This suggests that sublanguage-specific embeddings can help profile distinctive usage patterns for text categorization, offering greater specificity than latent topic distributions while not relying on potentially brittle lexical features. In addition, such profiles could also assist with concept normalization by providing more-informed prior probability distributions for medical vocabulary senses that are conditioned on the document or section type that they occur in.

A few limitations of our study are important to note. The embedding method we chose offers flexibility to work with arbitrary corpora and vocabularies, but its use of distant supervision

introduces some undesirable noise. The example given in Section 4.3 of the similar embeddings learned for the concept *cigars* and the concept of the CIGAR string in genomic sequence editing illustrates the downside of not leveraging disambiguation techniques to filter out noisy matches. On the other hand, our restriction to strings from SNOMED-CT and LOINC provided a high-quality set of strings intended for clinical use, but also removed many potentially helpful strings from consideration. For example, the UMLS also includes the non-SNOMED/LOINC strings "diabetes" and "diabete mellitus" [*sic*] for *Diabetes Mellitus* (C0011849), both of which occur frequently in MIMIC data. Misspellings are also common in clinical data; leveraging well-developed technologies for clinical spelling correction would likely increase the coverage and confidence of sublanguage concept embeddings.

At the same time, the low volume of data analyzed in many document types introduces its own challenges for the learning process. First, though JET can in principle learn embeddings for every concept in a given terminology, this is predicated on the relevant surface forms appearing with sufficient frequency. For a small document sample, many such surface forms that would otherwise be present in a larger sample will either be missing entirely or insufficiently frequent, leading to effectively "missed" concepts. While we are not aware of another concept embedding method compatible with arbitrary unannotated corpora that could help avoid these issues, some strategies could be used to reduce the potential impact of both training noise and low sample sizes. One approach, which might also help improve concept consistency in the document types that yielded few or no high-confidence concepts, would be pretraining a shared base embedding on a large corpus such as PubMed abstracts, which could then be tuned on each document type-specific subcorpus. While this could introduce its own noise in terms of the differences between biomedical literature language and clinical language (Friedman et al., 2002), it could help control for some degree of sampling error and provide a linguistically-motivated initialization for the concept embedding models.

Finally, as we observed with *Mental state* (C0278060), relying on similarity in contextual patterns can lead to capturing more corpus-specific features with embeddigns, as opposed to (sub)language-specific features, as target corpora become smaller and more homogeneous. If a particular concept or set of concepts are always used within the same section of a document, or in the same set phrasing, the "similarity" captured by organization of an embedding space will be more informed by this writing habit endemic to the specific corpus than by clinically-informed semantic patterns that can generalize to other corpora.

6 Conclusion

Analyzing nearest neighborhoods in embedding spaces has become a powerful tool in studying diachronic language change. We have described how the same principles can be applied to sublanguage analysis, and demonstrated that the structure of concept embedding spaces captures distinctive and relevant semantic characteristics of different clinical document types. This offers a valuable tool for sublanguage characterization, and a promising avenue for developing document type "fingerprints" for text categorization and knowledge-based concept normalization.

Acknowledgments

The authors gratefully thank Guy Divita for helpful feedback on early versions of the manuscript. This research was supported by the Intramural Research Program of the National Institutes of Health and the US Social Security Administration.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–36.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):D267–D270.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1495–1504, San Francisco, California, USA. ACM.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical semantic similarity with a neural language model. In

Proceedings of the 23rd ACM International Conference on Information and Knowledge Management - CIKM '14, CIKM '14, pages 1819–1822, Shanghai, China. ACM.

Kerstin Denecke. 2014. Sublanguage Analysis of Medical Weblogs. *Studies in Health Technology and Informatics*, 205:565–569.

Patrick Drouin. 2004. Detection of Domain Specific Terminology Using Corpora Comparison. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

K Feldman, N Hazekamp, and N V Chawla. 2016. Mining the Clinical Narrative: All Text are Not Equal. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 271–280.

Carol Friedman. 1986. Automatic structuring of sublanguage information. In Ralph Grishman and Richard Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 85–102. Lawrence Erlbaum Associates.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: A description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635—-E3644.

Denis Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M Lai. 2016. A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain. In *AMIA Summits on Translational Science Proceedings 2016*, pages 88–97. American Medical Informatics Association.

Ralph Grishman. 2001. Adaptive information extraction and sublanguage analysis. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining, Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, pages 1–4, Seattle, Washington, USA.

Leonie Grön, Ann Bertels, and Kris Heylen. 2019. Leveraging Sublanguage Features for the Semantic Categorization of Clinical Terms. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 211–216, Florence, Italy. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Veronika Laippala, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2009. Towards automated processing of clinical Finnish: Sublanguage analysis and a rule-based parser. *International Journal of Medical Informatics*, 78(12):e7 – e12.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. MCN: A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92:103132.

Eneldo Loza Mencia, Gerard de Melo, and Jinseok Nam. 2016. Medical Concept Embeddings via Labeled Background Corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4629–4636. European Language Resources Association (ELRA).

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2018. Jointly Embedding Entities and Text with Distant Supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206. Association for Computational Linguistics.

Minh C Phan, Aixin Sun, and Yi Tay. 2019. Robust Representation Learning of Biomedical Names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285, Florence, Italy. Association for Computational Linguistics.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Irina P Temnikova, William A Baumgartner, Negacy D Hailu, Ivelina Nikolova, Tony McEnery, Adam Kilgarriff, Galia Angelova, and K Bretonnel Cohen. 2014. Sublanguage Corpus Analysis Toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 2014:1714–1718.

Gaurav Vashisth, Jan-Niklas Voigt-Antons, Michael Mikhailov, and Roland Roller. 2019. Exploring Diachronic Changes of Biomedical Knowledge using Distributed Concept Representations. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 348–358, Florence, Italy. Association for Computational Linguistics.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of Semantic Change of Harm-Related Concepts in Psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Donald E Walker and Robert A Amsler. 1986. The use of machine-readable dictionaries in sublanguage analysis. In Ralph Grishman and Richard Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 69–83. Lawrence Erlbaum Associates, Hillsdale, NJ.

Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102. Association for Computational Linguistics.

Cheng Ye and Daniel Fabbri. 2018. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *Journal of Biomedical Informatics*, 83:63–72.

Qing T Zeng, Doug Redd, Guy Divita, Samah Jarad, Cynthia Brandt, and Jonathan R Nebeker. 2011. Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. *Journal of Health & Medical Informatics*, S3.