Learning Contextual Hierarchical Structure of Medical Concepts with Poincairé Embeddings to Clarify Phenotypes

Brett K. Beaulieu-Jones, Isaac S. Kohane and Andrew L. Beam†

Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

†*E-mail: Andrew_Beam@hms.harvard.edu*

dbmi.hms.harvard.edu

Biomedical association studies are increasingly done using clinical concepts, and in particular diagnostic codes from clinical data repositories as phenotypes. Clinical concepts can be represented in a meaningful, vector space using word embedding models. These embeddings allow for comparison between clinical concepts or for straightforward input to machine learning models. Using traditional approaches, good representations require high dimensionality, making downstream tasks such as visualization more difficult. We applied Poincaré embeddings in a 2-dimensional hyperbolic space to a large-scale administrative claims database and show performance comparable to 100-dimensional embeddings in a euclidean space. We then examine disease relationships under different disease contexts to better understand potential phenotypes.

Keywords: Clinical Concept Embeddings, Poincaré, Contextual Disease Relationships, Context-dependent Phenotypes, Deep Learning.

1. Introduction

Word embeddings[1] are a popular way to represent natural language and have seen wide use in machine learning applied to document classification,[?,?] machine translation,[?,?] sentiment analysis,[2] and question answering.[3,4] Clinical concept embeddings extend this approach to model healthcare events,[5–8] and have been particularly useful modeling longitudinal clinical data.[?,9–11] Traditional approaches such as word2vec[1] and GloVe[12] embed entities within a Euclidean space.

However, recent work by Nickel and Kiela on *Poincaré embeddings*[13] claims to provide better embedding representations of hierarchically structured data using a hyperbolic embedding space within the Poincaré ball. This n-dimensional hyperbolic space has a significantly higher capacity than the Euclidean space, which allows it to effectively embed structured trees while preserving distance relationships.[14–17] Moreover, this space allows for embedding of hierarchical, tree-like structures, as Nickel and Kiela[13] observed high fidelity embeddings of ontologies. This has an obvious relevance to medical concepts, given many have an inherent tree structure (e.g. disease nosology) that should be recapitulated in the embedding space.

When clinicians consider a disease, they examine the disease in the context of the patient's

overall environment.[18] For example, renal failure caused by poor blood flow to the kidneys as a result of long-term hypertension would be considered differently from renal failure as the result of a specific infection or immune system disorder like Lupus.[19] Accurate and precise phenotyping is critical to modern clinical studies using the electronic healthcare record (EHR) and other '-omic' associations studies (e.g. genomic, transcriptomic, metabolomic). Misclassified phenotypes have a severe effect on tests of association and require increased sample sizes to maintain constant power.[20–22] Increases in genetic testing and the availability of clinical data repositories (Electronic Health Record, Administrative Claims, large-scale Cohort) have enabled PheWAS association studies to be performed without the need to target and recruit specific populations for each individual study.[23–25] It is important to develop methods that enable researchers to consider a specific disease or phenotype in the context of the overall patient and environment.

We applied Poincaré embeddings to a large-scale administrative claims database to examine how the relationships of different conditions changed in distinct contexts. Our hypothesis was that the increased representational capacity offered by Poincaré embeddings and their ability to naturally model hierarchical data would result in improved embeddings for clinical concepts. We first demonstrate this by showing they can accurately reconstruct the ICD-9 hierarchy on synthetic data. Next we show that they find an improved representation on real data relative to traditional embedding approaches at the same number of dimensions. We conclude with a disease-specific embedding hierarchy within an obese population. Our results could provide a better representation of disease and allow for more accurate machine learning models as well as the fine-tuning of targeted phenotypes for association studies.

2. Methods

To examine the effectiveness of Poincaré embeddings for clinical concept embedding, we: 1.) trained Poincaré embeddings on the ICD-9 hierarchy as validation of parent-child tuples, 2a.) selected and preprocessed chronological member sequences of each diagnosis experienced for a specified cohort (e.g. obese vs. no metabolic disorders diagnosed), 2b.) Learned distributed vector representations for the real data by training a Poincaré embedding model in a two-dimensional space. 3.) Visualized the Poincaré embeddings in a two dimensional space. 4a.) Constructed a distance matrix within the hyperbolic space. 4b.) Analyzed the distance matrix to measure how effectively the embeddings represent clinical groupings (e.g. ICD9 Chapter, Sub-chapter and major codes).

2.1. *Source Code*

The source code used for the analyses in this work are freely available on Github (https://github.com/brettbj/poincareembeddings) under a permissive open source license. The optimized C++ Poincare Embedding implementation by Tatsuya Shirakawa is available under the MIT license (https://github.com/TatsuyaShirakawa/poincare-embedding).

2.2. *Data Source*

These analyses were performed using de-identified insurance administration data including diagnostic billing codes from January 1, 2008 until February 29, 2016 for more than 63 million members. The database does not include any socioeconomic, race or ethnicity data. The Institutional Review Board at Harvard Medical School waived the requirement for approval as it deemed analyses of the de-identified dataset to be non-human subjects research.

The data to rebuild the reference ICD9 hierarchy tree is available in the GitHub repository (https:/github.com/brettbj/poincareembeddings/data/icd9.tsv).

2.3. *Data Selection and Preprocessing*

2.3.1. *Reference ICD9 Example*

We first benchmarked against a known hierarchy, the ICD9 2015-Clinical Modification code ontology. To do this we extracted the ICD9 codes into four levels: 1.) Chapters (e.g. codes 390-459: Diseases of the circulatory system), 2.) Sub-chapters (e.g. codes 401-405: Hypertensive disease), 3.) Major Codes (e.g. code 401: Essential hypertension), and 4.) Detail level codes (e.g. code 401.0: Hypertension, malignant). We assigned relationships between each detail level code and the chapter, sub-chapter and major code it belonged to, each major code to the appropriate sub-chapter and chapter, and each sub-chapter to the chapter it belonged to.

2.3.2. *Real Member Analyses*

We performed cohort analyses by defining different study groups. First we included ten million randomly selected members (without replacement) who were enrolled for at least two years from the database of 63 million members. Next we separated two groups based on obesity diagnoses: 1.) ten million members who do not have a diagnosis for metabolic disorders with ICD9 codes between 270 and 279 2.) 3.38 million members who were diagnosed with obesity ICD9 codes (278.00 and 278.01).

Poincaré embeddings learn distributed vector representations from hierarchical data (e.g. a directed graph or tree). The input to the model is a list of tuples of the form $< A, B >$, which indicates that A and B have some form of unspecified relationship (e.g. *parent of*, *co-occurs with*, etc). In our case, the list of relationships specify that two diagnoses occurred sequentially, within a one year period, and had to occur more than ten total times and in more than 2% of all diagnoses.

2.4. *Poincaré Embeddings*

The key way in which Poincaré embeddings differ from traditional approaches is the distance metric which is used to compare the embeddings for two concepts. This distance metric is given in equation 1:

$$dist((x_1, y_1), (x_2, y_2)) = arccosh(1 + \frac{(x_2 - x_1)^2 + (y_2 - y_1)^2}{2y_1y_2}) \quad (1)$$

Equation 1 shows the distance between two points in the Poincaré ball hyperbolic space.

Training a Poincaré embedding model occurs by maximizing the distance (Equation 1) between unconnected nodes or diagnoses while minimizing the distance between highly connected nodes. This is done using a stochastic Riemannian optimization method, specifically stochastic gradient descent on riemmanian manifolds as seen in Bonnabel.[15]

2.5. *Processing and Evaluating Embeddings*

Once each concept is embedded into a two dimensional space, it is possible to calculate the pair-wise distance between all concepts using Equation 1. To assess how well the embeddings captured the ICD hierarchy on real data, we compared the average distances between concepts in the same ICD9 major code, sub-chapter and chapter against the distances of all other concepts. We then compared the capacity of a two-dimensional Poincairé space with varying size euclidean spaces. To do this, we repeated distance calculations with the clinical concept embeddings trained in a euclidean space on more than 63 million members in 2, 10 and 100 dimensions from Beam et al.[5] To normalize the distance comparisons between hyperbolic and euclidean spaces, we compared the ratio of distances between ICD codes within the same major, sub-chapter and chapter and the other ICD codes outside of the major, sub-chapter, and chapter.

3. Results

3.1. *ICD9 Hierarchy Evaluation*

To evaluate the method with a known ground truth, we embedded the ICD9 hierarchy and then reconstructed it as a tree. Because there are no counts included, stochasticity for all relationships at the same level (Chapter, Sub-chapter, Major, Detail) was expected. Figure 1 shows the reconstructed tree of the predefined ICD9 tree. This served as evidence that Poincairé embeddings can effectively embed a clean ICD9 hierarchy.

3.2. *Poincaré Embeddings on 10 Million Members*

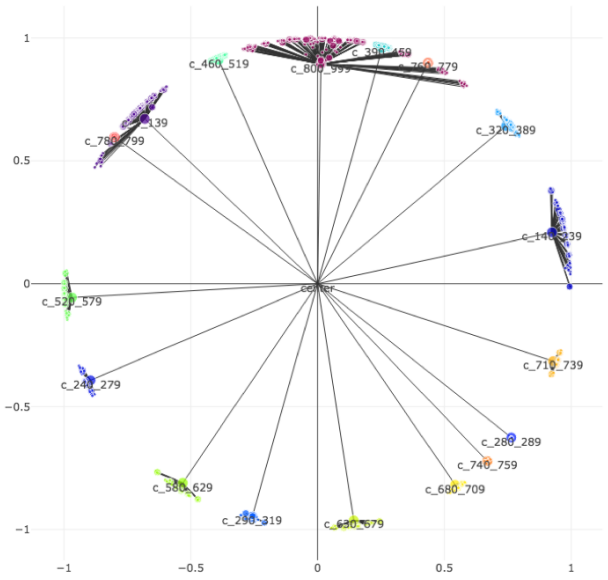
We then trained Poincaré embeddings in a two-dimensional space for 10 million randomly selected members (Table 1).

Table 1 Member Demographics of the Training Data

Demographics	
Male	40.4%
Female	59.6%
Age (2016)	48.66 (22.68)
ICD9 Diagnoses	22.38 (28.70)

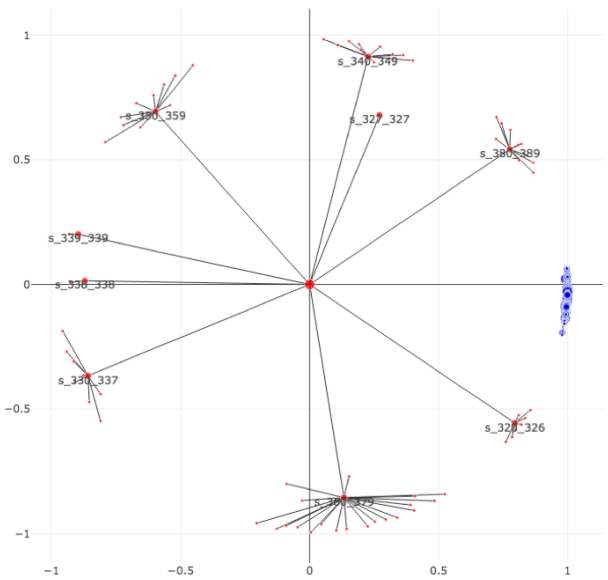
Figure 2A shows the ICD9 concepts (labeled by chapter) embedded in a two-dimensional space. While there were over 223 million total diagnoses, the majority of concepts had less than 100 distinct relations (Figure 2B) and the number of distinct relations was correlated with the distance from the origin ($R^2 = 0.61$) (Figure 2C).

A.



ICD Chapter	Code Range
Infectious and Parasitic Diseases	1-139
Neoplasms	140-239
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	240-279
Diseases Of The Blood And Blood-Forming Organs	280-289
Mental Disorders	290-319
Diseases Of The Nervous System And Sense Organs	320-389
Diseases Of The Circulatory System	390-459
Diseases Of The Respiratory System	460-519
Diseases Of The Digestive System	520-579
Diseases Of The Genitourinary System	580-629
Complications Of Pregnancy, Childbirth, And The Puerperium	630-679
Diseases Of The Skin And Subcutaneous Tissue	680-709
Diseases Of The Musculoskeletal System And Connective Tissue	710-739
Congenital Anomalies	740-759
Certain Conditions Originating In The Perinatal Period	760-779
Symptoms, Signs, And Ill-Defined Conditions	780-799
Injury And Poisoning	800-999

B.



ICD Chapter	Code Range
Diseases of the Nervous System and Sense Organs	1-139
Other	140-239

Fig. 1. ICD Example All codes

Figure 2 shows that the ICD hierarchy is correctly reconstructed using by the Poincaré embeddings in two dimensions. The distances between ICD codes in the same major, sub-chapter and chapter are smaller than the distances across different major codes, sub-chapters and chapters (Table 2). This shows that Poincaré embeddings are representing the data in a way that has similarities with the human-defined ICD9 hierarchy.

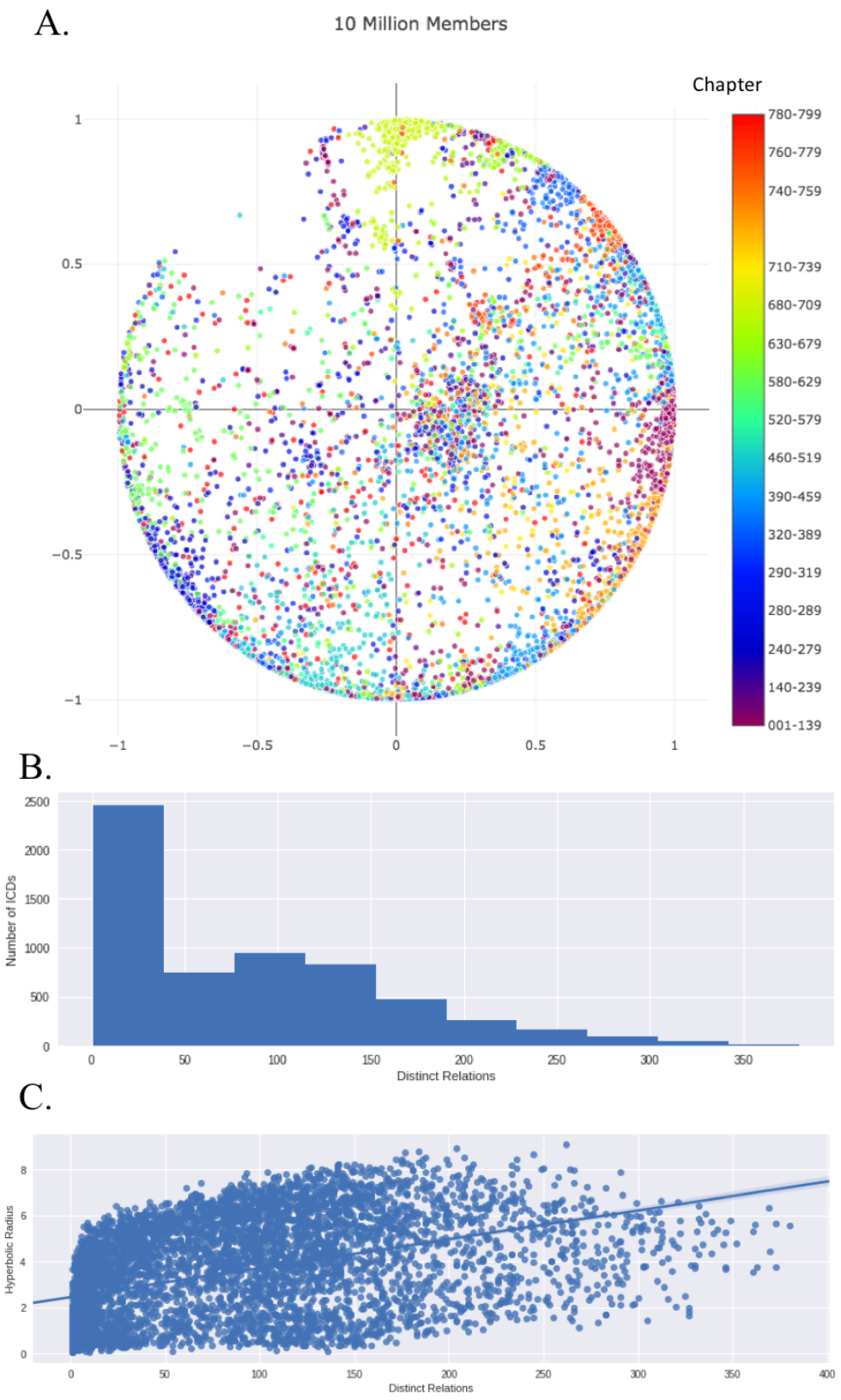


Fig. 2. A.) ICD9 Diagnoses Codes Embedded in a two-dimensional space. B.) Examination of the number of distinct relations for each ICD9 code. C.) Examination of the Correlation between the number of distinct relations and hyperbolic distance.

Table 2. Hyperbolic Distance comparison within Major, Sub-chapter and Chapter

Category	In Category	Outside of Category
Major	3.87 (1.71)	5.89 (1.92)
Sub-chapter	4.47 (1.73)	5.89 (1.92)
Chapter	4.91 (1.81)	5.91 (1.94)

3.3. *Comparison with Euclidean Embeddings*

To evaluate Poincaré embeddings against traditional euclidean embeddings, we compared the 2-dimensional Poincaré embeddings with 2, 10 and 100 dimension embeddings. The Poincaré embeddings were trained on 10 million randomly selected members. Running the preprocessing pipeline required 42 minutes on 16 cores but training the embeddings required only 49 seconds on 16 cores. All euclidean embeddings were trained on more than 63 million members. Table 3 shows the ratios of the mean distances of ICD codes in the same category over ICD codes in all other categories. We show the ratio to allow for comparison between Poincaré and Euclidean distances. As the dimensionality of the euclidean embeddings increased, the ratio of distance in-group vs. out of group decreased, indicating that the higher capacity enabled a better representation. The 2-dimensional Poincaré embeddings compared most closely to the 100-dimensional euclidean embeddings.

Table 3 Distance (ratio) comparison between Poincaré (2-dimensional) and Euclidean (2, 10, & 100-dimensional) within Major, Sub-chapter and Chapter.

Category	Poincaire (2d)	Euclidean (2d)	Euclidean (10d)	Euclidean (100d)
Major	0.657	0.758	0.668	0.649
Sub-chapter	0.759	0.863	0.794	0.774
Chapter	0.831	0.894	0.856	0.830

3.4. *Cohort Specific Embeddings*

Finally, we trained two separate Poincaré embeddings on patients with either: 1.) no prior diagnoses from the sub-chapter of metabolic disorders between ICD code 270 and 279 (N=10,000,000) and 2.) members diagnosed with obesity (ICD codes 278.00, 278.01, N=3,377,267) to first visualize the differences in the context of type 2 diabetes mellitus (Figure 3). Because the Poincaré embedding model was trained in 2-dimensions this was done without any further dimensionality reduction step.

We then examined the diseases in the closest quartile of either cohort to determine which showed the greatest movement from type 2 diabetes (Table 4). Of note, 22 of the top 50 were pain related and there are numerous links in the literature between both obesity (particularly joint and fibromyalgia[26,27]) and type 2 diabetes (particularly neuropathy[28]) with pain.

4. Discussion and Conclusion

Machine learning has great potential to improve the delivery of healthcare to patients, but many methodological challenges remain before this potential can be realized.[29,30] In this work,

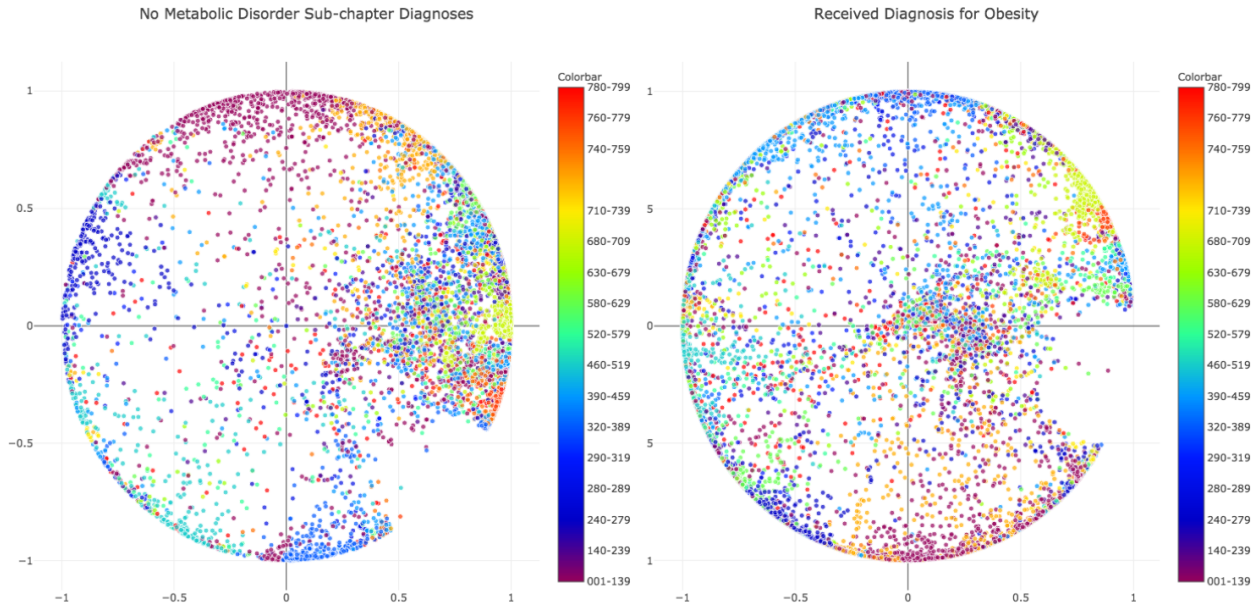


Fig. 3. A.) Poincaré Embeddings trained on 10M members with no metabolic disorder diagnoses (centered on type 2 diabetes). B.) Poincaré Embeddings trained on 3.38M members diagnosed with obesity (centered on type 2 diabetes).

Table 4. ICD9 Codes with the largest changes in distance from Type 2 Diabetes (250.00).

	ICD	Description
1	553.21	Incisional hernia
2	786.09	Other Respiratory Abnormalities
3	599.0	Urinary tract infection
4	285.9	Anemia
5	571	Chronic Liver Disease
6	583.6	Nephritis
7	724.5	Backache, unspecified
8	710.5	Eosinophilia myalgia syndrome
9	796.2	Elevated blood pressure w/o hypertension
10	719.46	Pain in Leg

we showed the increased capacity and hierarchical positioning of Poincaré embedding models can be useful to learn representations of disease diagnosis codes. Two-dimensional Poincaré embeddings were on par with 100-dimension euclidean embeddings when compared to the human-defined ICD hierarchy. Importantly the extra capacity of Poincaré embeddings may directly allow for visualization in a two-dimensional space, while traditional euclidean embedding techniques require an additional dimensionality reduction step (PCA, t-SNE, UMAP). Many of these techniques are non-deterministic and may not preserve global structure.

An important limitation of our current method is that the pre-processing step constructs binary relations between concepts whenever they occur with a specified threshold (more than

10 occurrences and 2% of cases). It is likely that additional information could be learned by encoding the actual frequency between concepts. In addition, it could be useful to evaluate additional distance matrices that have worked well for hierarchical problems in other domains, such as pg-gram and Edit distance.[31]

There are significant opportunities to expand on and apply these techniques to biomedical domains in order to examine and consider phenotypic context when performing associations. We are especially interested in the ability to contextualize a phenotype for association studies by considering the way ICD code relationships change given comorbidities. For example, start by measuring the way Poincaré embeddings change given a comorbidity (e.g. type 2 diabetes given metabolic disorder). If there are significant changes, it may be helpful to design association studies to separate endpoints, for example diabetes with no prior metabolic disorders and diabetes with prior metabolic disorders. In this case, the disease etiology may be distinct, and therefore we would expect the potential for different genetic drivers.

5. Acknowledgments

The authors thank Tatsuya Shirakawa for developing and open-sourcing an efficient implementation of the Poincaré Embedding Model. This work was supported in part by NLM grant 4 T15 LM007092-25.

References

1. B. T. Mikolov, K. Chen, G. Corrado and J. Dean, *arXiv:1301.3781* (2013).
2. C. R. Association for Computational Linguistics. Meeting (45th : 2007 : Prague, R. E. Association for Computational Linguistics., P. T. Pham, D. Huang, A. Y. Ng and C. Potts, *ACL 2007 : proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic.* (Association for Computational Linguistics, 2007).
3. J. Zhou and O. G. Troyanskaya, *Nature Methods* **12**, 931 (2015).
4. A. Bordes, J. Weston and N. Usunier, Open Question Answering with Weakly Supervised Embedding Models (Springer, Berlin, Heidelberg, 2014) pp. 165–180.
5. A. L. Beam, B. Kompa, I. Fried, N. P. Palmer, X. Shi, T. Cai and I. S. Kohane, *arXiv preprint arXiv:1804.01486* (2018).
6. Y. Choi, C. Y.-I. Chiu and D. Sontag, *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* **2016**, 41 (2016).
7. T. Ching, D. Himmelstein, B. Beaulieu-Jones, A. Kalinin, B. Do, G. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. Hoffman, W. Xie, G. Rosen, B. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. Carpenter, A. Shrikumar, J. Xu, E. Cofer, C. Lavender, S. Turaga, A. Alexandari, Z. Lu, D. Harris, D. Decaprio, Y. Qi, A. Kundaje, Y. Peng, L. Wiley, M. Segler, S. Boca, S. Swamidass, A. Huang, A. Gitter and C. Greene, *Journal of the Royal Society Interface* **15** (2018).
8. B. Beaulieu-Jones, *Machine learning for structured clinical data* 2018.
9. E. Choi, M. Taha Bahadori, A. Schuetz and W. F. Stewart, *Doctor AI: Predicting Clinical Events via Recurrent Neural Networks*, tech. rep.
10. Z. C. Lipton, D. C. Kale, C. Elkan and R. Wetzel (11 2015).
11. A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum,

K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado and J. Dean, *npj Digital Medicine* **1**, p. 18 (12 2018).
12. J. Pennington, R. Socher, C. M. P. o. t. 2014 and u. 2014, *aclweb.orgSign in* .
13. M. Nickel and D. Kiela, *Poincaré Embeddings for Learning Hierarchical Representations*, tech. rep.
14. M. Gromov., Hyperbolic groups., in *Essays in group theory, Springer.*, 1987 p. pages 75–263.
15. S. Bonnabel, *Stochastic gradient descent on Riemannian manifolds*, tech. rep.
16. A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston and O. Yakhnenko, *Translating Embeddings for Modeling Multi-relational Data*, tech. rep.
17. D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat and M. Boguna (6 2010).
18. B. K. B. Beaulieu-Jones and C. S. Greene, *Journal of Biomedical Informatics* **64**, 168 (2016).
19. M. M. Salem, *Seminars in nephrology* **22**, 17 (1 2002).
20. S. Smith, E. H. Hay, N. Farhat and R. Rekaya, *BMC genetics* **14**, p. 124 (12 2013).
21. S. Buyske, G. Yang, T. C. Matise and D. Gordon, *Human Heredity* **67**, 287 (2009).
22. R. Rekaya, S. Smith, E. H. Hay, N. Farhat and S. E. Aggrey, *The application of clinical genetics* **9**, 169 (2016).
23. A. Verma, A. Lucas, S. S. Verma, Y. Zhang, N. Josyula, A. Khan, D. N. Hartzel, D. R. Lavage, J. Leader, M. D. Ritchie and S. A. Pendergrass, *American journal of human genetics* **102**, 592 (4 2018).
24. J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden and D. C. Crawford, *Bioinformatics* **26**, 1205 (2010).
25. S. A. Pendergrass, K. Brown-Gentry, S. Dudek, A. Frase, E. S. Torstenson, R. Goodloe, J. L. Ambite, C. L. Avery, S. Buyske, P. Bůžková, E. Deelman, M. D. Fesinmeyer, C. A. Haiman, G. Heiss, L. A. Hindorff, C. N. Hsu, R. D. Jackson, C. Kooperberg, L. Le Marchand, Y. Lin, T. C. Matise, K. R. Monroe, L. Moreland, S. L. Park, A. Reiner, R. Wallace, L. R. Wilkens, D. C. Crawford and M. D. Ritchie, *PLoS Genetics* **9** (2013).
26. A. Okifuji and B. D. Hare, *Journal of pain research* **8**, 399 (2015).
27. D. S. McVinnie, *British journal of pain* **7**, 163 (11 2013).
28. M. J. Young, A. J. M. Boulton, A. F. Macleod, D. R. R. Williams and P. H. Sonksen, *Diabetologia* **36**, 150 (2 1993).
29. A. L. Beam and I. S. Kohane, *JAMA* **319**, p. 1317 (4 2018).
30. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam and R. Ranganath (6 2018).
31. D. Hassan, U. Aickelin and C. Wagner, *Comparison of Distance Metrics for Hierarchical Data in Medical Databases*, tech. rep.