diagnosis of HTN was finally discarded, and in the remaining 10 (32.2%) diagnosis of HTN was finally ascertained. The high risk category had a specificity of 95% and a PPV of 80% for the diagnosis of HTN.

Abstract USAGE Table 1. Clinical characteristics.

	Number of patients, n (%) N= 217
Female	188 (86.6)
Hypertension	51 (23.5)
T2DM	24 (11.1)
Dyslipidemia	68 (31.3)
Active smoking	13 (6.0)
No rheumatic disease	65 (30.0)
RA	78 (35.9)
SLE	6 (2.8)
Sjögren syndrome	5 (2.3)
OA	29 (13.4)
Other	34 (15.6)
PROOF-BP class:	
Low risk	133 (61.3)
Medium risk	71 (33.6)
High risk	11 (5.1)
New diagnosis of hypertension	14 (38.9)

Abstract USAGE Table 2. Diagnostic performances.

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Low risk	100	55	58	100
High risk	29	95	80	68

PPV: Positive predictive value; NPV: Negative predictive value.

Conclusion: In a cohort of Mexican-mestizo subjects with rheumatic diseases, 38.7% were classified as medium or high-risk for HTN. Only 42.8% of patients that required further evaluation followed medical recommendations. More than 30% of patients in the medium-risk category had a final diagnosis of HTN. Using the PROOF-BP algorithm, none of the patients in the low-risk stratum had a final diagnosis of HTN (sensitivity=100%); the high-risk category had a high specificity (95%) for the diagnosis of HTN.

REFERENCES

- [1] Lewington S, et al. *Lancet*. 2002;360:1903–1913.
- [2] Sheppard JP, et al. *Hypertension*. 2016;67:941–950.

Disclosure of Interests: None declared

DOI: 10.1136/annrheumdis-2019-eular.6238

AB1282

A BIG-DATA APPROACH TO ELECTRONIC HEALTH RECORD DATA – USING DIMENSIONALITY REDUCTION AND CLUSTERING TECHNIQUES TO STUDY LONGITUDINAL RELATIONSHIPS BETWEEN DISEASES

Marc Maurits¹, Thomas Huizinga¹, Soumya Raychaudhuri^{2,3}, Marcel Reinders^{4,5}, Elizabeth Karlson³, Erik van den Akker^{4,5}, Rachel Knevel^{1,3}. ¹LUMC, Rheumatology, Leiden, Netherlands; ²Brigham and Women's Hospital, Centre for Data Sciences, Boston, United States of America; ³Brigham and Women's Hospital, Rheumatology, Boston, United States of America; ⁴LUMC, Leiden Computational Biology Centre, Leiden, Netherlands; ⁵LUMC, Molecular Epidemiology, Leiden, Netherlands

Background: Hypothesis-free, longitudinal collection of patient health data in the form of Electronic Health Records (EHR) offers a wealth of valuable information on complex, slow-developing diseases in regard to aetiology and comorbidities. Conventional analytical methods are ill suited for the highly dimensional, sparse data contained within EHR, highlighting a need for more sophisticated, high-throughput tools. As t-Distributed Stochastic Neighbour Embedding (t-SNE)[1] and Density-Based Spatial Clustering of Applications with Noise (DBScan)[2] are designed to identify patterns in high-dimensional data with possible non-linear relationships, we hypothesized that these methods can aid identification of associations in diseases with multiple aetiologies.

Objectives: Proof of principle showcasing the value of t-SNE and DBScan in detecting longitudinal relationships between diseases in EHR data.

Methods: The Partners HealthCare Biobank from Boston, Massachusetts, includes 64,819 patients with longitudinal visit data from hospitals and general practitioners between June 1987 and June 2017. Each visit and procedure (N = 24,377,442) is coupled to an ICD code (International Classification of Disease) describing a disease or examination. We randomly split the data into two datasets of 32,424 and 32,395 individuals: set 1 to optimise t-SNE and DBScan and set 2 for replication. To trim the overly detailed hierarchy of ICDs, we translated them to Phenotype Codes (PheCodes).[3] t-SNE further reduced dimensionality and indicated groups of patients based on their PheCodes and separated patients based on PheCode patterns rather than singular codes. Subsequently DBScan identified clusters of patients in t-SNE space, by grouping patients based on relative Euclidean distance. Finally transition-probability matrices were constructed for all codes in each cluster, from which probabilistic sequences could be constructed.

We defined replication as an overlap in $\geq 25\%$ of the PheCodes between a cluster of set 1 and 2. Similarity was further assessed by calculating the absolute dissimilarity in transition probabilities for codes shared by matched clusters.

Figure 1. A. t-SNE clustering of EHR data DBScan hullplot (eps = 1.2 minPts = 8). B. Overlap of PheCodes in similarity matched clusters between sets 1 & 2 with 25% similarity threshold. C. Example of longitudinal flow within a single cluster, reproduced in both sets (both probabilities are shown along with averaged days between codes and the associated p-value calculated using a Student's t-test). D. Plot of the absolute dissimilarity in transition probabilities within the example clusters visualised in C.

Figure 1

Results: The average (range) number of codes per individual was 376.3 (1 – 8,419) and 375.9 (1 – 10,315) spread over 4106 (1 – 10,781) and 4153 (1 – 10,746) days for set 1 and 2 respectively. Even though our input data was a sparse, high-dimensional (1,865) matrix of PheCodes, t-SNE and DBScan could clearly separate various unique patient groups with 284 and 295 clusters in set 1 and 2. Clusters consisted of patients with PheCodes of well-defined disease entities such as cardiovascular diseases and neurological disorders with objectively meaningful disease sequences. 34.5% of the clusters identified in set 1 were replicated in set 2 based on our replication criteria. Figure 1 shows the results of each step.

Conclusion: Our proof of principle supports the use of unsupervised techniques such as dimensionality reduction and data clustering to identify longitudinal associations between medical events. These methods could prove useful in our quest to identify medical risk factors for incompletely understood diseases.

REFERENCES

- [1] L.J.P. v.d. Maaten, et al. *Journal of Machine Learning Research* 9 (Nov):2579-2605, 2008.
- [2] M. Ester, et al. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*: 226-231, 1996.
- [3] Denny JC, et al. *Bioinformatics*. 2010;26(9):1205–10

Acknowledgement: Joint last authorship by R. Knevel and E.B. van den Akker.

Disclosure of Interests: Marc Maurits: None declared, Thomas Huizinga Consultant for: Merck, UCB, Bristol Myers Squibb, Biotest AG, Pfizer, GSK, Novartis, Roche, Sanofi-Aventis, Abbott, Crescendo Bioscience Inc., Nycomed, Boehringer, Takeda, Zydus, Epirus, Eli Lilly, Soumya Raychaudhuri: None declared, Marcel Reinders: None declared, Elizabeth Karlson: None declared, Erik van den Akker: None declared, Rachel Knevel: None declared

DOI: 10.1136/annrheumdis-2019-eular.5281