Acknowledgments: The authors acknowledge the efforts of the Gene Expression Omnibus (GEO) database. The interpretation and reporting of these data are the sole responsibility of the authors.

Disclosure of Interests: None declared

DOI: 10.1136/annrheumdis-2020-eular.1831

FRI0585

HIGH-THROUGHPUT METHODOLOGY FOR EMR-BASED IDENTIFICATION OF CLINICAL SUB-PHENOTYPES IN COMPLEX PATIENT POPULATIONS

M. Maurits¹, T. Huizinga¹, M. Reinders^{2,3}, S. Raychaudhuri^{4,5}, E. Karlson⁵, E. Van den Akker^{2,3}, R. Knevel^{1,4}. ¹Leiden University Medical Center (LUMC), Rheumatology, Leiden, Netherlands; ²Delft University of Technology (TU Delft), Bioinformatics, Delft, Netherlands; ³Leiden University Medical Center (LUMC), Molecular Epidemiology, Leiden, Netherlands; ⁴Brigham and Women's Hospital, Rheumatology, Boston, United States of America; ⁵Brigham and Women's Hospital, Data Sciences, Boston, United States of America

Background: Heterogeneity in disease populations complicates discovery of risk factors. To identify risk factors for subpopulations of diseases, we need analytical methods that can deal with unidentified disease subgroups.

Objectives: Inspired by successful approaches from the Big Data field, we developed a high-throughput approach to identify subpopulations within patients with heterogeneous, complex diseases using the wealth of information available in Electronic Medical Records (EMRs).

Methods: We extracted longitudinal healthcare-interaction records coded by 1,853 PheCodes[1] of the 64,819 patients from the Boston's Partners-Biobank. Through dimensionality reduction using t-SNE[2] we created a 2D embedding of 32,424 of these patients (set A). We then identified distinct clusters post-t-SNE using DBscan[3] and visualized the relative importance of individual PheCodes within them using specialized spectrographs. We replicated this procedure in the remaining 32,395 records (set B).

Results: Summary statistics of both sets were comparable (Table 1). We found 284 clusters in set A and 295 in set B, of which 63.4% from set A could be mapped to a cluster in set B with a median (range) correlation of 0.24 (0.03 – 0.58).

Table 1. Summary statistics of the total Partners Biobank dataset and the 2 partitions.

	Set-A	set-B	Total
Entries	12,200,311	12,177,131	24,377,442
Patients	32,424	32,395	64,819
Patientyears	369,546.33	368,597.92	738,144.2
unique ICD codes	25,056	24,953	26,305
unique Phecodes	1,851	1,853	1,853

Figure 1. Two dimensional representation of Set A generated using dimensionality reduction (tSNE) and clustering (DBScan).

Clusters represented similar yet distinct clinical phenotypes; e.g. patients diagnosed with “other headache syndrome” were separated into four distinct clusters characterized by migraines, neurofibromatosis, epilepsy or brain cancer, all resulting in patients presenting with headaches (Fig. 1 & 2). Though EMR databases tend to be noisy, our method was also able to differentiate misclassification from true cases; SLE patients with RA codes clustered separately from true RA cases.

Figure 2. Phenotype Spectrographs (PheSpecs) of four clusters characterized by “Other headache syndromes”, driven by codes relating to migraine, epilepsy, neurofibromatosis or brain cancer.

Conclusion: We have shown that EMR data can be used to identify and visualize latent structure in patient categorizations, using an approach based on dimension reduction and clustering machine learning techniques. Our method can identify misclassified patients as well as separate patients with similar problems into subsets with different associated medical problems. Our approach adds a new and powerful tool to aid in the discovery of novel risk factors in complex, heterogeneous diseases.

References:

- [1] Denny, J.C. et al. Bioinformatics (2010)
- [2] van der Maaten et al. Journal of Machine Learning Research (2008)
- [3] Ester, M. et al. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. (1996)

Disclosure of Interests: Marc Maurits: None declared, Thomas Huizinga Grant/research support from: Ablynx, Bristol-Myers Squibb, Roche, Sanofi, Consultant of: Ablynx, Bristol-Myers Squibb, Roche, Sanofi, Marcel Reinders: None declared, Soumya Raychaudhuri: None declared, Elizabeth Karlson: None declared, Erik van den Akker: None declared, Rachel Knevel: None declared

DOI: 10.1136/annrheumdis-2020-eular.3489

FRI0586

HOW TO GET FROM THE MULTIDIMENSIONAL HEALTH ASSESSMENT QUESTIONNAIRE TO STANFORD HEALTH ASSESSMENT QUESTIONNAIRE DISABILITY INDEX SCORES IN PATIENTS WITH RHEUMATOID ARTHRITIS, PSORIATIC ARTHRITIS AND AXIAL SPONDYLOARTHRITIS: DEVELOPMENT AND VALIDATION OF A CONVERSION ALGORITHM

E. Svensson¹, K. Løngaard¹, L. Midtboell Ørnbjerg², R. Meincke², J. K. Pedersen³, L. Dreyer⁴, N. Steen Krogh⁵, D. V. Jensen², M. L. Hetland². ¹RKKP - the Danish Clinical Quality Program, Aarhus, Denmark; ²DANBIO, Copenhagen Center for Arthritis Research, Rigshospitalet, Copenhagen, Denmark; ³Svendborg Hospital, Svendborg, Denmark; ⁴Aalborg University Hospital, Aalborg, Denmark; ⁵Zitelab Aps, Frederiksberg, Denmark

Background: In the DANBIO quality registry in Denmark, patients with rheumatoid arthritis (RA) psoriatic arthritis (PsA) and axial spondyloarthritis (axSpA) have reported Patient Reported Outcomes (PROs) including the Stanford Health Assessment Questionnaire Disability Index (HAQ-DI) for nearly twenty years as part of routine care. Patients' feedback have stressed a need for a shorter registration of disability (1). While the shorter Multidimensional Health Assessment Questionnaire (MDHAQ) is preferred by patients, the original HAQ-DI is the preferred tool in observational studies. Thus, a conversion algorithm between the MDHAQ and HAQ-DI scores is warranted.

Objectives: To develop and validate a simple conversion algorithm between MDHAQ and HAQ-DI scores in RA, PsA and axSpA patients.

Methods: Patients registered in DANBIO with a diagnosis of RA, PsA or axSpA who had completed both HAQ-DI and MDHAQ simultaneously at a visit +/- 30 days from start of conventional synthetic (cs)DMARD or biological (b)DMARD were eligible for the analysis, and randomly divided into development and validation cohorts stratified by diagnosis. The conversion algorithm was developed in the RA development cohort using linear regression with HAQ-DI as the dependent variable and MDHAQ as the independent variable. The predicted HAQ (pHAQ) scores were then