|  |  |
| --- | --- |
| **eMERGE Network: Proposal for Analysis**  Project/Manuscript Concept Sheet | |
| **Reference Number** | NT286 |
| **Submission Date** | 5/10/2018 |
| **Project Title** | Using genetic data in a Bayesian precision medicine framework to prioritize rheumatic disease diagnoses |
| **Tentative Lead Investigator** *(first author)* | Rachel Knevel |
| **Tentative Senior Author**  *(last author)* | Soumya Raychaudhuri |
| **All Other Authors** | Elizabeth Karlson |
| **Sites Involved** | A network-wide study (all sites invited to participate). |
| **Background / Significance** | In the current era of precision medicine and increasing application of next-generation sequencing, there is a rapidly increasing number of patients who have whole-genome genotype data available.1 If these data can be integrated with medical records, they can be used in clinical care, and have valuable information that can be exploited in advance of a clinical visit.One of the important challenges at the patient’s first visit is to differentiate between the possible diseases that might be causing symptom. This is particularly valuable for diseases where distinguishing symptoms might evolve over time and watchful waiting might be essential before a final diagnosis can be achieved. In certain instances, genetic information might be useful at prioritizing likely diagnoses for patients with similar early-disease-stage symptoms.2,3  To explore the opportunities of genetics for disease differentiation, we used the rheumatology outpatient clinic as an example. Here, the majority of the patients present with synovitis as the first symptom and early disease identification is crucial for optimal care.4,5 Rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), spondyloarthropathy (SpA), psoriatic arthritis (PsA), and gout are synovitis causing diseases that encompass approximately 80% of the cases in the outpatient rheumatology clinic.6,7 These rheumatic diseases are particularly promising since many genetic risk loci have been identified for these diseases8, making synovitis the ideal symptom to test our proof of principle.  We developed a genetic probability score where each patient gets a probability for each of the included disease. When a probability is sufficiently high or low a clinician may prioritize diagnostic testing if it is consistent with clinical signs and symptoms.  **This is a proof of principle study testing the performance of genetics to discriminate patients with synovitis.** |
| **Outline of Project** | The genetic probability score combines population-level genetic risk and a sex adjusted symptom-level disease risk. We first calculate the population-based probabilities from a weighted-GRS for each disease with a sex-adjusted disease risk. These GRSs are a summation of the number of risk alleles (both single nucleotide polymorphisms (SNPs) and HLA amino-acids or haplotypes) weighted by the disease susceptibility effect size. Finally, we calculate the within-case probability, by normalizing the population-based probabilities such that the maximum reaches 1 in each patient.  The performance of the genetic probability score will be tested in multiple phases:   1. Simulation of 1 million people with genetic-based case selection (n= 49,151) to test theoretical framework 2. Partners Biobank (n=15,043) with 245 clinical disease cases identified through chart review. 3. eMERGE patient data (n=83,717) with ICD based case identification.   We have finalized phase I and II have, which both showed a good discriminatory ability of G-Prob AUC=0.84 in the simulated data and 0.81 on the clinical data.  We intend to use the eMERGE as final replication step. |
| **Desired Variables**  *(essential for analysis*  *indicated by* ***\*****)* | ICD9 or 10 codes\*  Demographics: age and gender\*  Common Variable set: autoantibody status for anti-CCP, ANA, rheumatoid factor |
| **Desired Data** | Imputed genetic data |
| **Planned Statistical Analyses** | Building the genetic probability score consist of a combination of   * Weighted genetic risk score (GRS) * Logistic regression * Newton-Raphson method to estimate the ideal intercept for the   Testing the models performance:   * Repeated measure correlation (R package rmcorr)8 * Area Under the receiver operating Curve (R package pROC)9,10 |
| **Ethical Considerations** | There are no additional risks involved. The analyses will be performed on de-identified data. The data will be stored at Dr. Raychaudhuri’s servers at Partners’ HealthCare.  This study makes use of data that is already obtained from electronic health records |
| **Target Journal** | TBD |
| **Milestones\*\*** | June 2018: eMERGE case-selection + genetic probability score calculation  July 2018: writing the manuscript |

***\*\**** *This section should include the timeline for completion of project, including: approval, project duration, first and second draft of the paper and submission.*

References:

1. Evans, J. P., et al. (2017). "Introducing “Genomics and Precision Health”." JAMA 317(18): 1904-1905.
2. <https://ignite-genomics.org/about-ignite/>
3. JAMA Insights: Genomics and Precision Health series + B.M., , Kuehn. (2017). "Pilot Programs Seek to Integrate Genomic Data Into Practice." JAMA 318(5): 410-412.
4. Combe, B., et al. (2016). "2016 update of the EULAR recommendations for the management of early arthritis." Ann Rheum Dis 0: 1–12.
5. EULAR early arthritis, vdHelm and Karim Raza
6. Zink A, Listing J, Klindworth C, Zeidler H. The national database of the German Collaborative Arthritis Centres: I. Structure, aims and patients. Ann Rheum Dis 2001; 60(3): 199-206. doi: 10.1136/ard.60.3.199
7. Miedema HS, van der Linden SM, Rasker JJ, Valkenburg HA. National database of patients visiting rheumatologists in the Netherlands: the standard diagnosis register of rheumatic diseases. A report and preliminary analysis. Br J Rheumatol 1998; 37(5): 555-61. doi: 10.1093/rheumatology/37.5.555.
8. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 77.
9. Metz CE (1978) Basic principles of ROC analysis. Seminars in Nuclear Medicine 8: 283–298.*],*
10. Lu Q, Elston RC (2008) Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet 82: 641–651