Title: *FMR1* Premutation Phenotypes in a Population-Based Sample

Authors: Arezoo Movaghar1,2, David Page3, Murray Brilliant4, Mei Wang Baker5, Jan Greenberg1, Jinkuk Hong1, Leann Smith Dawalt1, Krishanu Saha1,2, Marsha Mallick1

1Waisman Center, University of Wisconsin-Madison, Madison, WI, United States
2Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI, United States
3Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States
4Marshfield Clinic Research Institute, Marshfield, WI, United States
5Wisconsin State Laboratory of Hygiene, Madison, WI, United States

Introduction: Despite the high prevalence of *FMR1* premutation, the potential impact of this genetic variant on human health has not been fully explored. We have used a novel, unbiased and robust approach to discover the pattern of health problems in female premutation carriers by employing computational phenotyping methods on Marshfield Clinic electronic health records (EHR). We examined the EHR data from 72 female permutation carriers and 507 controls, selected from general population and who are unaware of their *FMR1* genotype. Our extensive phenotyping provides strong evidence that female *FMR1* premutation carriers experience higher burden of disease from younger age and through the lifespan comparing to the normal population.

Method: Using a machine learning method, the participants’ EHRs were used to predict their *FMR1* premutation status. For the present study, we used a supervised machine learning method called random forest. Diagnostic codes from Marshfield Clinic’s EHR system were utilized as the input variables. The participants were matched on year of birth and duration of receiving care from the clinic. We examined the codes with various age thresholds including all of the codes that were received at least twice (rule of 2) before age of 40, 60, 80 or lifetime diagnoses. We pruned the diagnostic codes to the ones that were observed in at least 5 participants. To measure the success of classification, the area under receiver operating characteristic curve (AUROC) was reported. To select the candidate phenotypes, we have used a measure called mean decrease in impurity based on Gini score (MDG). Variables with higher MDGs had higher contribution in creating the prediction model. After identifying these variables, we examined the EHRs of participants regarding to 1) the percentage of cases and controls who received these diagnoses, 2) number of visits for those conditions and 3) the age of participants when they received the diagnosis for the first time.

Results: We were able to successfully differentiate premutation carriers from controls using the diagnostic codes in their EHRs. Our random forest classifiers could predict the premutation status using diagnostic codes received before age of 40, 60, 80 or lifetime diagnoses with AUROC of 0.63, 0.65, 0.65 and 0.6, respectively. These results provide evidence that the health profiles of female premutation carriers are significantly different from normal population.

Examination of the first 100 variables with highest MDG showed that premutation carriers have higher frequency of visits for those diagnoses and also were diagnosed at a younger age compared to the normal population. Categorizing the diagnostic codes revealed that the most frequent groups of diagnostic codes that differentiate the two groups are related to diseases of the musculoskeletal system and connective tissue (i.e., fibromyalgia and limb pain), diseases of the genitourinary systems (i.e., hypertrophy of uterus and female infertility) and mental, behavioral and neurodevelopmental disorders (i.e., anxiety and depression).

Discussion: Although there have been multiple studies investigating health and well-being of *FMR1* premutation carriers, almost all of them focused on clinically-derived samples, which are inherently biased towards individuals who experience higher level of stress. In addition these studies are usually limited to a small range of targeted phenotypes. Our discovery-oriented approach is the first study to investigate the health characteristics of female *FMR1* premutation carriers in a population setting incorporating the entire spectrum of available health diagnoses. While our computational phenotyping results showed that premutation carriers suffer from more health difficulties over their life span, our sample size limits our ability to examine individual diagnostic codes. Additional studies on larger populations are required to replicate the results.

Acknowledgement: This research was supported by grant from NICHD (R01 HD082110) to M. Mallick and by the core grant of the Waisman Center IDCRC (U54 HD090256) to Q. Chang. We also acknowledge the Morse Fellows program (for support of Arezoo Movaghar). We would like to than Emily Allen, Don Bailey, Deborah Hall, Jessica Hunter, Melissa Raspa, Stephanie Sherman, Peter Todd, and Anne Wheeler. We are very grateful to the Marshfield Clinic Research Institute and all individuals who have participated in our research study.