# SUPPLEMENTAL MATERIAL

## Prediction Model Development

All adults who resided in Olmsted County in 2009 were identified by the Rochester Epidemiology Project (REP) Census[10] and used to develop a prediction model to select patients who had a high likelihood of initiating statin therapy in the next 3 years. Statin medication was chosen for the model because these are commonly prescribed and have an actionable PGx variant.[11] After excluding individuals who were deceased or had been prescribed statin therapy prior to January 1, 2009, the remaining subjects were followed up until December 31, 2011, for the initiation of statin therapy. As potential baseline risk factors, the model considered patient demographics including age on January 1, 2009, sex, race, and BMI. To identify a set of health conditions that predict future statin use, the Clinical Classifications Software (CCS) was used to cluster patient diagnoses and procedures into clinical categories (http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp). The Cox proportional hazards model was utilized to develop the statin prediction model using the variables selected through the Lasso shrinkage method. For continuous variables, the spline method was used to determine their functional form. Using the final model, the predicted risk of initiating statin therapy in the next 3 years was calculated for each individual. Discriminant ability of the model was assessed by the C-statistics. Using a predicted risk of 0.35 or higher and 0.15 or lower as predicted events and non-events, respectively, sensitivity and specificity were calculated.

*Results*: Among 92,712 Olmsted Country adult residents in 2009, 77899 (84%) individuals had BMI measurements within 3 years prior to the index date of January 1, 2009. A total of 65,665 individuals were included in the model after excluding subjects who were deceased or had been prescribed a statin prior to the index date. During the 3-year follow-up, 7533 (11.5%) subjects initiated statin therapy. The final multivariate predictive model included patient demographics (age in cubic spline, sex, and race) and 6 CCS codes (lipids, diabetes, peripheral atherosclerosis, disease of blood forming organs, coronary atherosclerosis and other heart diseases, and hypertension) as independent risk factors that are associated with statin use. The final model had a C-statistic of 0.74 as a measure of discrimination. Using 0.35 and 0.15 as a cutoff for determining predicted events and non-events, respectively, the sensitivity and specificity of the final model were 0.65 and 0.53, respectively. Among the 2009 Olmsted County adults studied, 28% had a predicted risk of at least 35% (Supplemental Figure 1).

## PGx Sequencing and Genotyping Methods

*CYP2D6* Test Methods*:* DNA was purified from EDTA blood by Qiagen EZ1®Advanced instruments in the PGL using manufacturer's guidelines. The Luminex

*CYP2D6* ASPE kit v2 was used per product manual to genotype all samples in this study. This kit detects the following *CYP2D6* recombinants and alleles: gene duplication, gene deletion (**5*), -1584C>G (**2A*), 100C>T (**4* and **10*), 124G>A (**12*), 138insT (**15*), 883G>C (**11*), 1023C>T(**17*), 1661G>C (**2*, **4*, **17*, and others), 1707delT (**6)*, 1758G>T/A (**8*, **14*), 1846G>A(**4*), 2549delA (**3*), 2613–2615 delAGA (**9*), 2850C>T (**2*, **17*, **41*, and others), 2935A>C (**7*), 2988G>A (**41*), and 4180G>C (**2*, **4*, **17*, and others). The *CYP2D6* nucleotide numbering used herein is based on a historic nomenclature model utilizing the first coding nucleotide as number "1" followed by consecutive numbering of each subsequent nucleotide including intronic nucleotides (http://www.cypalleles.ki.se/cyp2d6.htm). .

PGRN-Seq Methods: The PGRN-Seq capture reagent was designed to capture 84 PGx relevant genes using the NimbleGen in-solution custom capture method. This method allows for the selective enrichment of the desired genomic regions by hybridizing human genomic DNA to immobilized baits that have been designed to bind just the regions of interest. All other DNA is washed away, thus eliminating the 99% of the genome that is not of current interest. DNA sequencing is performed on an Illumina HiSeq using a protocol that sequences 101 bases on both ends of the captured material. Samples can be pooled into batches of 24 per lane of a HiSeq flow-cell. Following DNA sequencing, NGS data is passed through two custom bioinformatics pipelines, one for clinical variant calling of the four target genes and another for general research purposes.

**Bioinformatic Pipeline Methods**

The analysis pipeline for next-generation sequencing DNA data at Mayo Clinic entails three steps: alignment, single nucleotide and small insertion/deletion variant calling, and annotation. FASTQ files are aligned to the hg19 reference genome using Novoalign (VN:V2.07.13) with the following options: --hdrhd off -v 120 -c 4 -i PE 425,80 -x 5 -r Random. Realignment and recalibration was performed using GATK (VN:1.6-7-g2be5704) Best Practices version 3. Germline variations are called with GATK's UnifiedGenotyper. Variant quality score recalibration is also done with the following command line optimizations: for SNVs, -an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an DP -nt 2 --maxGaussians 4 --percentBadVariants 0.05 ; and for INDELS, -an QD -an FS -an HaplotypeScore -an ReadPosRankSum --maxGaussians 4 -nt 2 --percentBadVariants 0.12 -std 10.0. VCF files are subsequently annotated using Mayo Clinic's BioR annotation repository. This system includes structured data on variant population frequencies (HapMap, 1k Genomes, ESP6500), function predictions of variant effects, information on base conservation, occurrence of regulatory motifs, and summary data on the local occurrence of a mutation in Mayo Clinic processed samples. All information is consolidated and presented in a structured format for rapid mining.

### Clinical Sequence Database (Oracle TRC)

All variants that pass the bioinformatics pipeline QC thresholds are then loaded into the clinical sequence database. This database was implemented using Oracle's Translation Research platform. The system consists of the following components: 1) data extraction framework for extracting clinical data from multiple sources, 2) data normalization and quality processes, 3) standardized clinical data model to ensure data consistency, 4) genomic feature identification, 5) genomic normalization/standardization, 6) genomic data model for molecular features and annotation, and 7) user application to enable identification of cohorts based on both genomics features and clinical characteristics. This database provides Mayo the ability to normalize and integrate phenotypic data from various clinical source systems and genomics data. This provides investigators the ability to quickly identify and extract data on cohorts for prospective and retrospective studies.

SUPPLEMENTAL TABLE 1. 84 Genes Captured by the Next-Generation Sequencing Reagent Developed by the Pharmacogenomics Research Network

*ABCA1*
*ABCB1*
*ABCB11*
*ABCC2*
*ABCG1*
*ABCG2*
*ACE*
*ADRB1*
*ADRB2*
*AHR*
*ALOX5*
*APOA1*
*ARID5B*
*BDNF*
*CACNA1C*
*CACNA1S*
*CACNB2*
*CES1*
*CES2*
*COMT*
*CRHR1*
*CYP1A2*
*CYP2A6*
*CYP2B6*
*CYP2C19*
*CYP2C9*
*CYP2D6*
*CYP2R1*
*CYP3A4*
*CYP3A5*
*DBH*
*DPYD*
*DRD1*
*DRD2*
*EGFR*
*ESR1*
*FKBP5*
*G6PD*
*GLCCI1*
*GRK4*
*GRK5*
*HLAB*
*HLA-DQB3*
*HMGCR*
*HSD11B2*
*HTR1A*
*HTR2A*
*KCNH2*
*LDLR*
*MAOA*
*NAT2*
*NPPB*
*NPR1*
*NR3C1*
*NR3C2*
*NTRK2*
*PEAR1*
*POR*
*PTGIS*
*PTGS1*
*RYR1*
*RYR2*
*SCN5A*
*SLC15A2*
*SLC22A1*
*SLC22A2*
*SLC22A3*
*SLC22A6*
*SLC47A1*
*SLC47A2*
*SLC6A3*
*SLC6A4*
*SLCO1A2*
*SLCO1B1*
*SLCO1B3*
*SLCO2B1*
*TBXAS1*
*TCL1A*
*TPMT*
*UGT1A1*
*UGT1A4*
*VDR*
*VKORC1*
*ZNF423*

SUPPLEMENTAL TABLE 2: Phenotyping Algorithm for CYP2D6

| Predicted Phenotype[a] | Enzyme Activity | CYP2D6 alleles detected (No Duplication) | CYP2D6 alleles detected (With Duplication[b]) |
|---|---|---|---|
| Ultra-Rapid Metabolizer | Increased | 2 increased activity alleles | ≥ 3 normal alleles |
| Extensive Metabolizer | Normal | 2 normal alleles | 1 normal allele + 2 decreased activity alleles<br>2 normal alleles + 1 no activity allele |
| Intermediate Metabolizer | Decreased | 1 normal allele + 1 decreased/no activity allele<br>2 decreased activity alleles | ≥ 3 decreased activity alleles |
| Poor Metabolizer | No activity | Only no activity alleles detected | |

[a] There are instances where a phenotype prediction is not categorical and a range of possible phenotypes will be given (example: extensive to intermediate metabolizer). Other laboratories may use different phenotype prediction methods as there is no consensus on this in the literature.

[b] A duplication or multiplication of the *CYP2D6* gene is possible; examples of a multiplication for Extensive Metabolizer have not been included in this table.

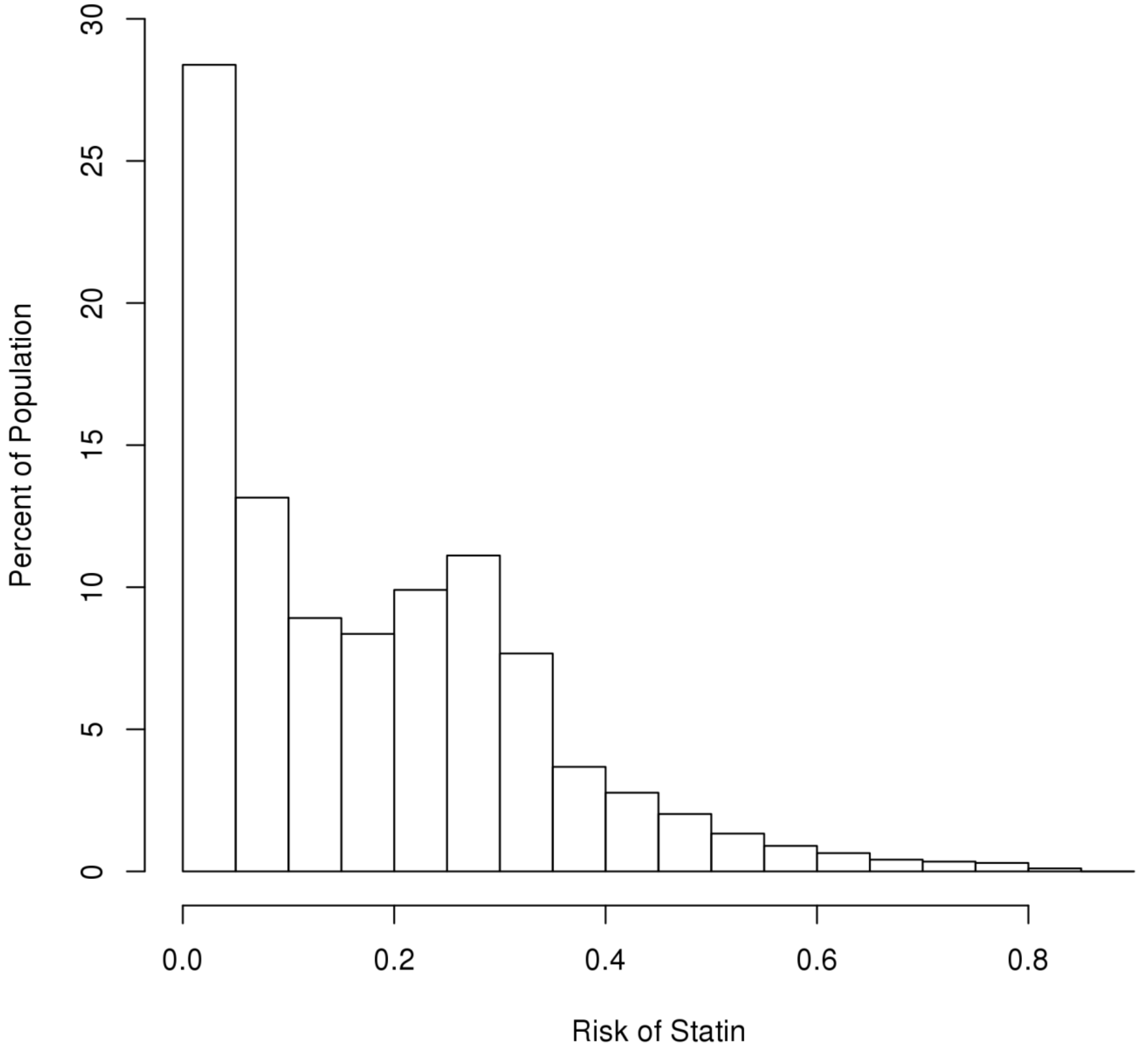
Normal allele = *1

Increased activity allele = *2A, gene duplication (depending on allele duplicated)

Decreased activity alleles = *2, *9, *10, *14B, *17, and *41

No activity alleles = *3, *4, *4N, *5, *6, *7, *8, *11, *12, *13, *14A, and *15

**SUPPLEMENTAL FIGURE 1.** Distribution of the Risk of initiating statin within 3 years within the Rochester Epidemiology Project Population.


Percent of Population
Risk of Statin