**RESEARCH STRATEGY**

1. **SIGNIFICANCE**

Molecular-based early prediction of complex diseases is a major goal of medical research as it both provides an avenue to curtail pathologies through timely interventions and may elucidate underlying disease mechanisms. Numerous studies have demonstrated beneficial alterations of disease trajectories with early medical interventions in diseases such as systemic inflammatory conditions1-5, diabetes6-7, cancers8-10, cardiovascular disease11-13, and obesity.14-16 The use of biomarkers, such as genetic variants, circulating cytokines, metabolic markers, and antibodies to pathogens, has shown increasing utility for disease prediction and prognosis.17-20 Furthermore, careful examination of predictive sets of molecular biomarkers can highlight underlying perturbations in critical pathways involved in pathogenesis.

The overarching goal of this study is to use powerful machine learning techniques to create predictive models for eight common diseases using a combination of exome data, circulating cytokines, metabolic markers, and endogenous antibodies to selected epitopes (for two diseases). Our long-term aim is to produce clinically useful, molecular-based predictive models of diseases and to show the utility of using large biobanks linked to electronic medical record information. By screening multiple diseases in our population-based biobank, we increase the likelihood of identifying one or more diseases that exhibit excellent predictive accuracy that will serve as the basis for subsequent replication and translational investigations. The exome data will capture signals from coding variants across the genome that confer increased inherited susceptibility to the diseases studied. Pivotal signaling in the immuno-metabolic axis will be measured in the circulating proteins selected. Lastly, autoantibodies representing disruption in immune tolerance and antibodies to infectious agents that may trigger pathogenesis will be interrogated in the multiplexed antigen assay. The diseases interrogated in this study are: **rheumatoid arthritis**, **systemic lupus erythematosus**, **relapsing-remitting multiple sclerosis**, **premature myocardial infarction**, **chronic lymphocytic leukemia**, **obsessive compulsive disorder**, **autoimmune thyroid disease**, and **axial spondyloarthritis**. Importantly, we are proposing to interrogate biological samples collected prior to clinical diagnosis. Hence, this study has the strong potential for aiding wide array of individuals through providing predictive information early in disease processes where medical interventions are the most efficacious.

**A1. Importance of results from this study:** The development of molecular-based predictive models for common diseases would represent a dramatic advancement in medical diagnostics, enabling prompt therapeutic intervention and potentially curtailing morbidity. High throughput molecular screens coupled with advanced statistical and computational approaches have set the stage for assessment of the viability of such predictive models. Several valuable results will be generated from this study: (i) we will determine the diagnostic efficacy of classifiers for eight common diseases using the combined signals from genetic data, circulating cytokines, metabolic markers, and antibodies to autoantigens and pathogens, (ii) we will gain understanding of the role of these various molecular markers in each disease studied, (iii) we will elucidate the proportion of disease-affected individuals who harbor predictive molecular signals, (iv) by assaying molecular panels prior to clinical diagnoses using biobanked samples, we will help delineate the molecular changes that precede clinical diagnosis, and (v) the study results may suggest specific biological pathways that play a role in the etiology of these diseases.

**A2. Critical barriers:** The large majority of common diseases have been relatively resistant to prognostic efforts. Four important impediments have conspired to hinder advances in the development of viable predictive models for risk of the diseases proposed to be studied: 1) The use of dynamic biomarkers assayed in samples collected after clinical diagnosis has produced signals that can be driven by advanced disease processes and therapeutic interventions; 2) The use of a single type of biomarker (e.g., solely genetic markers) has generally produced classifiers with insufficient diagnostic utility; 3) The heterogeneity in clinically-defined phenotypes reduces the efficacy and applicability of predictive models and 4) The use of genetic-based predictive models that are restricted to GWAS-significant markers undermines the performance of such models for common diseases.21

Prediction of disease onset using banked samples and interrogating a large combined set of inflammatory cytokines, metabolic markers, exome variants, and circulating antibodies has not been performed due to the lack of availability of such data, previous technological impediments and complexity of analyses. The difficult access to large biorepositories with linked, longitudinal electronic medical records, high-throughput protein-based assays, and exome genotyping data has also substantially hindered advancement of disease prediction studies. In addition, the analysis of data generated by a large-scale study using samples with complicated medical record information and high density genomic/protein/antibody measurements requires expertise in complex statistical and machine learning methods as well as highly developed informatics skills. Assembling a seasoned team with expertise in these areas is difficult.

**A3. Overcoming the critical barriers:** (1) To better determine biomarkers that are antecedent to disease onset, we will use biobanked DNA and plasma collected prior to clinical diagnosis. The use of the Personalized Medicine Research Project (PMRP)—a 14 year-old biobank of DNA, plasma and serum from 20,000 adults with linked electronic medical records—at the Marshfield Clinic enables us to determine sets of individuals with biological samples collected before a physician diagnoses an individual with disease. (2) To harness orthogonal signals across molecular data types, we will interrogate the combined signals from (i) exome-wide genotype data, (ii) circulating inflammatory cytokines and metabolic markers, and (iii) the presence of antibodies to a variety of autoantigens and pathogen epitopes; (3) Using electronic medical record information consisting of laboratory test results, diagnostic codes and medications along with the genetic, protein and antibody data, we will perform dimensional reduction techniques such as PCA and binary PCA designed to identify significant subgroups within a heterogeneous disease state. (4) To capture both GWAS-significant and sub-GWAS-significant predictive signals from the genome, we will use the entire exome genotype array data measured on individuals in the study.

The Marshfield Clinic Research Institute (MCRI) has collected, maintained, and enhanced the PMRP, which has played an integral role in numerous disease susceptibility studies.17,22-27 The Integrated Research & Development Laboratory, a service center at MCRI under the direction of Dr. Meece, has implemented a high-throughput pipeline for utilization of plasma samples from the PMRP. Dr. Schrodi has exploited this resource in quantifying inflammatory cytokines on >2,000 samples and immune-metabolic markers on >600 samples, resulting in a high utility Bayesian Network classifier for type 2 diabetes in samples obtained prior to clinical diagnosis (AUC=0.91). The MCRI houses the Biomedical Informatics Research Center, a service center with considerable computational infrastructure consisting of a research data warehouse built to store health and biological data and the High Performance Computing Cluster designed to perform complex analyses of high dimensional data. Preliminary queries have been performed by this Center under the guidance of Dr. Schrodi to identify individuals satisfying a rigorous set of criteria involving ICD codes and clinical laboratory measurements for each disease studied. Dr. Schrodi has 17 years of post-PhD experience performing human genomic experiments and statistical analyses on genetic and biological data using Bayesian methods, machine learning algorithms, permutation routines, Monte Carlo simulations, and various frequentist tests. He has designed and managed several large-scale genetic mapping studies of inflammatory diseases (~$1.6M/study) in the biotechnology industry for over a decade. Recently, using the PMRP samples, he has conducted studies of genetic variants underlying cytokine expression and recessive diplotypes across the diseases proposed in this study.

Dr. Judy Smith, a collaborator with Dr. Schrodi on several studies including a Rheumatology Research Foundation-funded study on ankylosing spondylitis, has considerable clinical expertise and expert knowledge of immunology. Dr. Smith is a tenured associate professor in the University of Wisconsin-Madison School of Medicine & Public Health. Drs. Smith and Schrodi recently published the results of a study showing how genetic variants and expression of Tnfaip3 contribute to differential macrophage responses in individuals with ankylosing spondylitis compared to controls.28

Dr. Shicheng Guo has also collaborated with Dr. Schrodi. Dr. Guo has training and expertise in human genetics and computational genetics. Dr. Guo has conducted research in genetic analyses, epigenetics, computational biology, and bioinformatics. Dr. Guo is currently a postdoctoral fellow in Dr. Schrodi’s laboratory.

Drs. Trappl-Kimmons and Liang have considerable expertise in developing, assaying and analyzing high-throughput antigen/antibody assays.

Dr. Maadooliat, a tenure-track assistant professor in Mathematics, Statistics and Computer Science at Marquette University, is also a close collaborator with Dr. Schrodi. Dr. Maadooliat has spent the last three years working from the Center for Human Genetics at the Marshfield Clinic Research Institute where he is a faculty member, sponsored by Dr. Schrodi. Dr. Maadooliat has considerable expertise in statistical/computational methods applied to protein structure and functional data analysis. Drs. Maadooliat and Schrodi have jointly conducted research in the areas of linkage disequilibrium patterns from disease loci29, Bayesian multiple testing30, and identifying disease susceptibility genes through shared chromosomal regions. Two published manuscripts have resulted from the collaboration.

**A4. Scientific knowledge gained:** The development of predictive models for one or more common diseases would provide substantial insight into how molecular intermediate phenotypes can be used to predict disease states and could significantly facilitate early treatment of those conditions. Importantly, (1) This study will evaluate each disease against matched controls and test for significant signals from genetic markers, circulating proteins and antigen/antibody reactivity profiles. Doing so will provide evidence for or against known and novel molecular markers in correlation with each disease entity and identify specific pathways that contribute to disease etiologies. (2) This study will characterize the level of classification performance for molecular-based predictive models. (3) This study will show the relative contributions of genetics, key cytokines and metabolic proteins, and antibodies to the prediction of diseases; and (4) We will assess whether the panel of molecular markers partitions patients into significant subgroups within any of the diseases investigated. Importantly, we will validate an initial cross-validated predictive model for rheumatoid arthritis in an independent set of rheumatoid arthritis and matched controls. In potential situations where we cannot develop a viable, cross-validated predictive model, then we will have generated evidence of either the lack of classifier signal from the combined action of exome genetics, dynamic biomarkers and antibodies for our sample sizes and disease definitions and/or the machine learning algorithms employed. Such negative results would suggest that new approaches to disease prediction are needed.

**A5. Potential implications:** There has been a rapid adoption of the use of large patient populations with linked electronic medical records for the purpose of understanding the molecular pathogenesis and pathophysiology across a spectrum of disease phenotypes. This work will demonstrate how studies utilizing these resources can serve to better understand how to create predictive models for diseases. Once predictive models for early disease states are validated, implementation in clinical settings will enable physicians to identify individuals early in their disease course and provide timely therapeutic interventions.

1. **INNOVATION**

The proposed study has several important aspects that, in combination, result in a highly innovative study. Our group is well-positioned and experienced to harness of three key advances to create molecular-based predictive models for the diseases studied here. (i) We have access to 20,000 biobanked samples with linked longitudinal electronic medical records (PMRP). As the patients for the biobank were recruited over 14 years ago and the average individual has in excess of 34 years of electronic medical record data, these samples enable us to utilize molecular profiles from samples obtained prior to disease diagnosis. (ii) The recent and explosive growth of high-throughput genomic and proteomic technologies has enabled cost-effective large-scale screens of important types of molecular markers. The exome genotyping array, applied to the PMRP samples, interrogates the large majority of genome-wide significant SNPs, identify-by-descent polymorphisms, ancestry informative SNPs, and ~250k nonsynonymous coding variants. The Luminex multiplex immunoassay technology enables the concurrent measurement of dozens of key inflammatory cytokines, chemokines and metabolic proteins with high sensitivity and wide dynamic range. The Antigen Discovery platform provides a method to interrogate circulating antibodies to hundreds of different autoantigens and pathogen epitopes. Lastly, (iii) the recent advances in computational statistics and machine learning algorithms allows for the rigorous development of robust and accurate predictive models from high dimensional data sets. The team assembled will consolidate all three aspects into a study that dives deeper into extensive molecular profile data to produce predictive models of early disease using data obtained prior to diagnosis.

**B1. Shift in current research paradigm through novel approaches**

Often, with dynamic biomarkers such as the inflammatory cytokine and metabolic protein panels proposed here, retrospective studies can face difficulties in delineating the effects of molecular pathogenesis from effects generated by the disease process and/or medical treatment. This issue hinders the mechanistic interpretation of biomarkers and may reduce the robustness of predictive models using these markers. The experimental design proposed here is quasi-prospective in that plasma samples from the biobank were obtained ~14 years ago and subsequent disease diagnosis occurred in a substantial fraction of those with the diseases studied. Hence, we are restricting our analyses to those cases that had samples collected prior to clinical diagnosis. While not truly a prospective design, this design reduces identifying case/control molecular differences that result from disease treatment or the molecular perturbations that occur with advanced disease states. By adopting such an experimental design, we are better positioned to interrogate molecular effects in dynamic biomarkers more likely to underlie pathogenesis. While still correlative, we argue that results from this quasi-prospective approach yield more reliable evidence for causal relationships than purely retrospective designs. We have generated data for metabolic dysfunction using this design, interrogating immuno-metabolic markers. Our results show that, for example, C-peptide is dramatically increased (54%, on average) in individuals who were undiagnosed with type 2 diabetes at the time of measurement, but have subsequently transitioned to type 2 diabetes following enrollment in PMRP (P=1.18E-13).

Another advance that this study employs is the concurrent measurement of ‘multi-omics’ data on individuals studied. For complex diseases, sole reliance on single types of molecular data can produce limited results for predictive models. For example, reliance on GWAS-significant SNPs for disease prediction has repeatedly shown limited diagnostic utility.21 By expanding the investigation to include key inflammatory cytokines, critically important metabolic markers, antibody binding to hundreds of antigens, and exome data, we harness predictive signal from disparate biochemical processes and susceptibility factors. This technique has recently produced several insights into various clinical traits, including hepatocellular carcinoma32, Alzheimer’s disease33, and optimal nutrition.34 As preliminary data for our ability to assay protein levels and measure genome-wide genetics, we measured 9 inflammatory cytokines in stored plasma samples of over 2,000 PMRP individuals. Our results show, for example, that (i) a polymorphism in the *LAIR1* gene, an inhibitory receptor expressed on PBMCs and reportedly involved in systemic lupus erythematosus35, is strongly associated with IL-17A levels (P=1.33E-08). Further, we have shown interesting co-regulatory cytokine patterns, such as IL-10 levels are highly correlated with IL-12p70 levels (P-value Spearman’s rho = 1.91E-67).

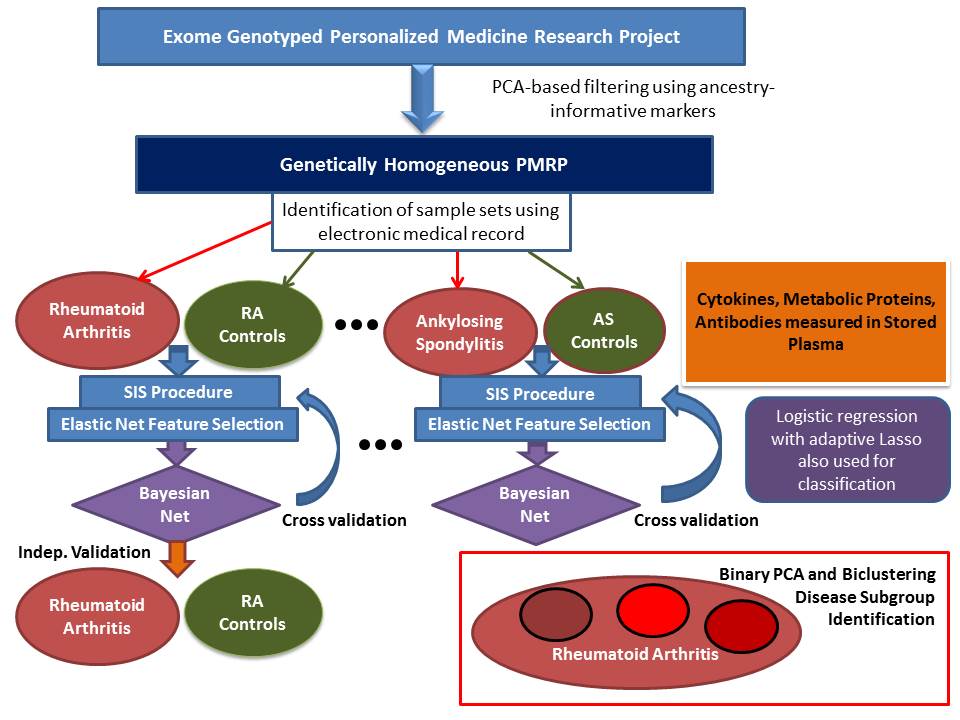
Moreover, our predictive models will use all exome array data instead of focusing on simply the GWAS-significant markers. Theoretical and empirical research in livestock and crop phenotypic traits and breeding values had suggested that using genome-wide SNP data with generalized linear mixed models (e.g., best linear unbiased prediction) approach has considerable merit.36-38 Building on this work, researchers in human genetics have provided evidence that the prediction of traits in humans is well-informed by using all SNPs across the genome.39-41 Logistic regression with adaptive Lasso (one of the two classification methods proposed) is an approach that can utilize all of the data on the exome array in combination with the circulating protein and antibody data.

**B2. Shift in current research paradigm through methodology**

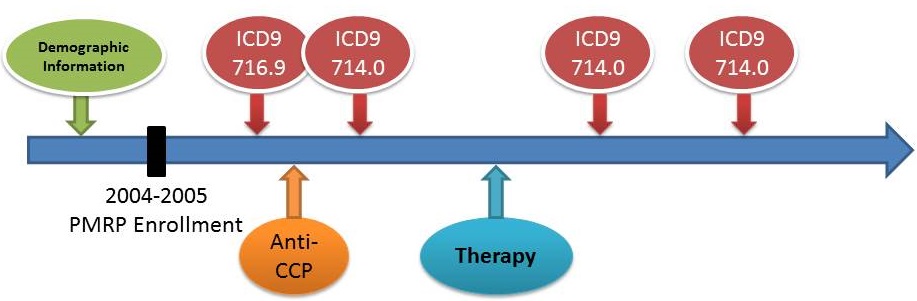
Several types of statistical methods and machine learning algorithms will be used in this study. Three researchers involved in the study, Drs. Maadooliat, Schrodi and Guo, have expertise in the development of a variety of quantitative methods including Bayesian networks and regression methods, multiplicity approaches, transforms, permutation methods, functional data analysis, Monte Carlo simulations, clustering methods, resampling methods, and general linear mixed models.29,30,42-52 Aside from the use of standard methods across the study, three areas of methodology research will be investigated: 1) Due to heavy reliance on observations of gross physiology in a clinical setting, the correspondence between molecular pathogenic effects and clinical disease states is not a one-to-one mapping for common, complex diseases. Indeed, there are multiple examples of molecular etiologies that were previously classified as single disease entities (e.g., breast cancer can meaningfully subdivided into estrogen receptor positive/negative disease, with profound implications for prognosis, trajectory, and treatment decisions).53-55 To better identify molecular-based disease subgroups within each disease examined, we will use and extend binary PCA and biclustering methods.56,57 2) The use of multiple ‘omics’ data types in predictive models necessitates sophisticated machine learning approaches. We will refine and tailor a Sure Independence Screening (SIS) procedure for priming data for feature selection from these molecular data. SIS is a simple and elegant tool that is widely used to reduce the dimensionality of data to a manageable scale.58 It is well known that the screening and commonly used penalties (e.g., adaptive Lasso, Elastic net, and SCAD) can produce valuable results for variable selection in high dimensional frameworks. SIS and its extensions are useful techniques to improve the computational efficiency of these feature selection algorithms, as well as the accuracy in extracting informative signals from very large biological datasets. Using the SIS/Elastic Net-determined features, we will construct Bayesian Network-based classifiers59,60 for use on the generated exome genotype data, protein biomarkers, and antibody data. Bayesian Networks are widely considered to produce highly competitive predictive performance when compared to other approaches. In addition, Bayesian Networks can (i) be easily motivated by first principles of probabilistic laws applied to graphs, and (ii) produce highly transparent results showing the level at which each feature drives the classification and the correlative structure between features. In addition to Bayesian Networks, we will also evaluate high dimensional logistic regression with adaptive Lasso as an alternative classifier. 3) Characterizing results from machine learning classifiers in a way that both captures the performance of the prediction as well as the clinical utility is challenging. While we will use the standard area under the Receiver Operating Characteristic curve (AUC)61 (by averaging AUC values over 10-fold cross validations) to evaluate classifier performance, we will also explore the use of the posterior probability of disease (PPD) (probability of disease conditioned on the molecular features used in the predictive model), where we investigate metrics that compare the probability mass of the PPD that exceeds a high threshold and the probability mass of PPD below a low threshold.21 We believe this type of analysis method both measures the performance of a classifier and has direct applicability in a clinical setting as physicians often tailor clinical decisions based on whether they believe an individual’s probability of disease exceeds a particular value.

1. **APPROACH**

**Experimental Design Overview**

****

The first step in the experimental design is to utilize the panel of ancestry informative markers on the exome beadchip to conduct a PCA analysis removing genetic-background outliers within the set of 20,000 PMRP individuals. From the resulting set of individuals, we will identify individuals from each of the diseases studied using the electronic medical record information. We have found that for many diseases using solely ICD9 codes can often generate inaccurate identification (i.e., low concordance rates with manual adjudication from clinical professionals). Therefore, multiple incidences of ICD9 codes (not simply one instance of a code) coupled with laboratory test, prescription data, and/or imaging data is used to identify cases. The impact of the number of ICD9 instances on the power of genetic association studies has been previously investigated and has informed this work.62 Much of the identification of cases from the eight diseases investigated has been performed by as part of a different study from Schrodi, Maadooliat and Guo. Manual adjudication of cases defined in this manner has seen very high validation rates (100% for multiple sclerosis and >96% in other phenotypes). Further, these have been used in genetic studies where known positive controls (e.g., HLA-DRB1 in RA) are significantly associated. Controls will be defined as those individuals without any evidence of the disease studied or similar diseases and matched based on age and sex. Dr. Smith will provide clinical expertise aiding in determining case/control definitions. Below is an example timeline of the medical trajectory of an individual diagnosed with RA:

****

Importantly, we will restrict the case definitions to those with clinical diagnoses occurring after sample collection/enrollment in PMRP. Once case/control sets are identified, stored plasma will be pulled and plated for interrogation for cytokines/chemokines and metabolic proteins on two Luminex multiplex assays by Eve Technologies (the 42-plex Human Cytokine Array/Chemokine Array and the Human Metabolic Hormone array 9-plex). To correct for total protein level variation across samples, we will run an inexpensive BSA-based total protein assay. This has already been accomplished for 1,000 PMRP samples where moderate differences were observed for results obtained under total protein-adjusted cytokine concentration values compared to unadjusted cytokine levels. Genes identified using the BSA-adjusted data were more likely to fall in pathways that corresponded to what is known about the biology of these cytokines. Preliminary data on IL-6 and IL-17A cytokine concentrations for 2,015 PMRP samples are shown below.

In addition, stored plasma for rheumatoid arthritis, systemic lupus erythematosus and controls will be sent to Antigen Discovery Inc. (ADI) to interrogate antibody reactivity to 800 autoantigens and pathogen antigens. ADI has produced a high throughput protein microarray for the identification of immune signatures of humoral response to hundreds of proteins on a single assay. The microarray is produced by coupled *in vitro* transcription/translation of pathogen and human genetics, followed by printing onto nitrocellulose bonded to glass slides. ADI has demonstrated the utility of this platform. Only a few microliters of plasma are necessary for screening. Proteins include established autoantigens & pathogen epitopes.

**IL-6 distribution**

**IL-17A distribution**





Proteins with no known immune reactivity serve as technical controls. Using a fluorescence scanner, data from the microarrays is acquired and checked for quality, followed by statistical analysis. Drs. Liang and Trappl-Kimmons from ADI will be Co-Investigators on this study.

**Central Wisconsin Population and the PMRP**

The genetically homogeneous population in rural Central Wisconsin is the source population for the PMRP.63 All PMRP participants are adults. This highly stationary population is largely derived from Bavarian migrants from the late 1800s and carried high utility for genetic and biomarker studies through avoidance of confounding by population stratification and reduction in allelic heterogeneity. Environmental exposures are through to be relatively uniform across this population and migration rates are low. The PMRP has been effectively used in numerous genetic and clinical research studies including the NHGRI-funded eMERGE network.64,65

**Plasma Collection**

30mL (3x10mL K3EDTA purple-top tubes) of blood was collected for plasma from each PMRP participant. Tubes were immediately placed on ice or refrigerated before processing. Within 24 hours, plasma cryovials were stored at -80oC.

**Selection of Study Individuals**

PMRP samples were collected ~14 years ago. ~19,000 individuals were recently genotyped on the Illumina Human Exome + Human Core Exome Beadchips. This genotyping platform has excellent coverage of coding variants across the exome and also has GWAS-significant markers. Sample randomization was used for plating samples for genotyping. The exome genotyping beadchips used have ancestry informative marker (AIM) content. The Schrodi Lab used this set of AIMs to run a standard PCA on the full set of 19,000 individuals and genetic background outliers (~180 individuals) based on the first 3 PCs were removed from subsequent analyses. In an attempt to diminish effects from transient inflammation stimulated from acute infection, only individuals without evidence of acute infection +/- 2 weeks of PMRP enrollment will be included. Evidence of acute infection consisted of either a temperature reading of >37.7 oC, an hs-CRP lab test >3.0, or an abnormal white blood cell count of <4.3 or >10.8 within this time duration. 249 individuals satisfied one or more of these acute infection conditions and will be excluded from the study. Similarly, a record of immunization within 2 weeks of PMRP enrollment will be an exclusion criterion. Record of immuosuppressive drug prescription within 3 months prior to enrollment will also be an exclusion criterion.

**Disease Phenotypes**

Predictive models each of eight diseases will be developed for this study. Cases and Controls will be selected from the PMRP biobank using the electronic medical record, inclusive of ICD9/ICD10 codes, laboratory tests and medications. Over the past 4 years, Dr. Schrodi has consulted with several physicians at the Marshfield Clinic and UW-Madison to refine definitions of these 8 diseases. Using the CDC statistics, we estimate that over 4% of the United States population suffers from one or more of the diseases examined. The table below gives the results of informatics queries on the estimated number of individuals satisfying the disease criteria:

|  |  |
| --- | --- |
| Disease | **Estimated Number with Samples Prior to Dx** |
| Rheumatoid Arthritis | **194 total (97 for Classifier Development / 97 for Validation)** |
| Systemic Lupus Erythematosus | **86** |
| Multiple Sclerosis | **31** |
| Premature Myocardial Infarction | **88** |
| Chronic Lymphocytic Leukemia | **84** |
| Obsessive Compulsive Disorder | **112** |
| Autoimmune Thyroid Disease  Axial Spondyloarthritis | **95**  **30** |

An *a priori* guess at the effect size from the combined effects of the multi-omics data investigated here is extremely difficult. Hence, power calculations may not be informative. We note that for our pilot study of type 2 diabetes, 56 pre-T2D cases and 445 matched controls yielded an average cross-validated AUC of 0.912 using our elastic net logistic regression feature selection coupled with a Bayesian Network classifier (feature selection was included in the cross validation procedure). This impressive result suggests that the sample sizes for this study should be sufficient to produce accurate predictive models in at least 6 of the 8 diseases.

**Rheumatoid arthritis**. Rheumatoid arthritis is a debilitating systemic autoimmune disease targeting the synovium, but affecting numerous organ systems. There is evidence that rheumatoid arthritis may be triggered in susceptibility individuals by exposure to one or more pathogens. To identify individuals with a diagnosis of rheumatoid arthritis within the PMRP, we selected patients satisfying all of the following criteria: (i) three or more instances of ICD9 714.0 for rheumatoid arthritis, (ii) a positive rheumatoid factor or anti-CCP antibody test, and (iii) having received at least one prescription of methotrexate and/or anti-TNF medication. Manual spot-checking of the medical records of the selected individuals validated the approach. In addition, we extracted exome genotype data from these individuals and compared against controls. A gene-based test was applied to these genetic data to test for compound heterozygosity effects as part of a related pilot study. The positive control *HLA-DRB1* was clearly significant, following correction for multiple testing. Controls for rheumatoid arthritis will be composed of PMRP individuals without any instances of ICD9 codes for systemic inflammatory diseases, arthropathies, and similar conditions. We expect to use 194 RA cases and 400 controls, to be split evenly into a 97/200 training set and 97/200 test set (to be used as verification of the top performing classifier, SA3).

**Systemic lupus erythematosus.** Multiple instances of the 710.0 code will be used in conjunction with a positive lupus anticoagulant, ACA, anti-ENA, or ANA test results will define cases. We will use controls matched on age and sex and without any autoimmune disease diagnoses or chronic inflammatory diseases. We will select twice the number of cases as controls.

**Multiple sclerosis**. Multiple sclerosis is also a systemic autoimmune disease characterized by demyelination of the central nervous system. Motility impairment, vision loss, pain, muscle weakness and fatigue are all frequent symptoms of the disease. Within the Marshfield Clinic system, we have found excellent accuracy of the ICD9 codes and imaging data for multiple sclerosis. Indeed, of a sample of 20 PMRP individuals determined to have multiple sclerosis from their ICD9 code information, all 20 were verified by a Marshfield Clinic neurologist to have relapsing-remitting multiple sclerosis from chart review. We will use 100 controls matched on age and sex for multiple sclerosis.

**Myocardial infarction.** Any 410 ICD9 code (410, 410.x, 410.xx) coupled with a high-sensitivity cardiac troponin T (hs-cTnT) test greater than 51 ng/L within 1 week of the ICD9 code. Earlier age of onset will be prioritized for inclusion in the case group. The top priority group for inclusion will be males ages <50 and females <55. Between 400 and 500 controls, matched on age and sex, will be used for myocardial infarction.

**Chronic Lymphocytic Leukemia.** CLL cases are defined as PMRP individuals with two or more instances of the 204.1 ICD9 code and prescriptions for relevant chemotherapies/immunotherapies (e.g., rituximab). Controls will be matched on age and sex and not have any neoplasm diagnoses or systemic inflammatory diseases. We will select twice the number of cases as controls.

**Obsessive compulsive disorder.** Two or more instances of the 300.3 ICD9 code will serve at the criterion for case inclusion. We will use 200 age and sex-matched controls without any psychiatric disease diagnosis.

**Autoimmune thyroid disease.** Multiple instances of the 245.2 or 242.0 ICD9 code coupled with abnormal TSH, anti-TPO, anti-Tg, FT4 or TSI laboratory tests will serve as the criterion for cases. We will use 200 age and sex-matched controls without any autoimmune diagnosis or systemic inflammatory disease diagnosis.

**Axial spondyloarthritis.** Cases will be defined as having the 720.2 ICD9 code or multiple instances of the 720.0 ICD9 code coupled with a positive HLA-B27 test and having received a prescription for anti-TNF medication or methotrexate. We will use 100 age and sex-matched controls, without any systemic inflammatory disease diagnosis.

**Specific Aim 1**

**To systematically characterize the levels of 48 key circulating inflammatory and metabolic proteins and exome variants in eight common diseases and matched control groups. Additionally screen antibodies to 800 autoantigens and pathogen epitopes in rheumatoid arthritis and systemic lupus erythematosus.**

Our approach to SA1 will be to measure (1) 42 cytokines and chemokines on an Eve Technologies (Alberta, Canada) Luminex multiplex assay in 50uL of plasma from all cases and controls selec0ted from the PMRP biobank; (2) 6 metabolic proteins on an Eve Technologies Luminex multiplex assay on 50uL of plasma from all cases and controls selected from the PMRP biobank; Eve technologies Luminex multiplex arrays have been successfully used on human samples to investigate numerous different inflammatory and infectious diseases66-68; and (3) All individuals studied will have the Illumina exome beadchip data previous collected and stored within the PMRP database at MCRI. Extraction of exome variants involves identifying case and control individuals, then merging the phenotype files with the exome variant file by the BIRC center at MCRI.

Eve Technology Human Cytokine array/Chemokine Array 42-plex: EGF, Eotaxin-1, FGF-2, Flt-3L, Fractalkine, G-CSF, GM-CSF, GRO(alpha), IFNalpha2, IFNgamma, IL-1alpha, IL-1beta, IL-1ra, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-12 (p40), IL-12 (p70), IL-13, IL-15, IL-17A, IL-18, IP-10, MCP-1, MCP-3, MDC, MIP-1alpha, MIP-1beta, PDGF-AA, PDGF-AB/BB, RANTES, sCD40L, TGFalpha, TNFalpha, TNFbeta, VEGF-A

Eve Technology Human Metabolic Hormone Array: C-Peptide, GIP, Insulin, Leptin, MCP-1, PYY (total)

Lastly, (4) a custom 800 antigen high-density protein microarray panel from Antigen Discovery using plasma from 120 rheumatoid arthritis cases, 80 systemic lupus erythematosus cases and 100 shared controls from the PMRP biobank. This antigen/antibody protein microarray has been successfully used to interrogate human humoral immune response to infectious agents and autoantigen activity.69-72

**Peptide Quantification Preliminary Data**

We measured C-peptide and adiponectin in a cohort of individuals without type 2 diabetes at the time of blood draw, then partitioned individuals who had met the formal criteria for diabetes in the years following the blood draw. These immune-metabolic markers were strongly correlated with prospective disease as determined by a Kolmogorov-Smirnov test.

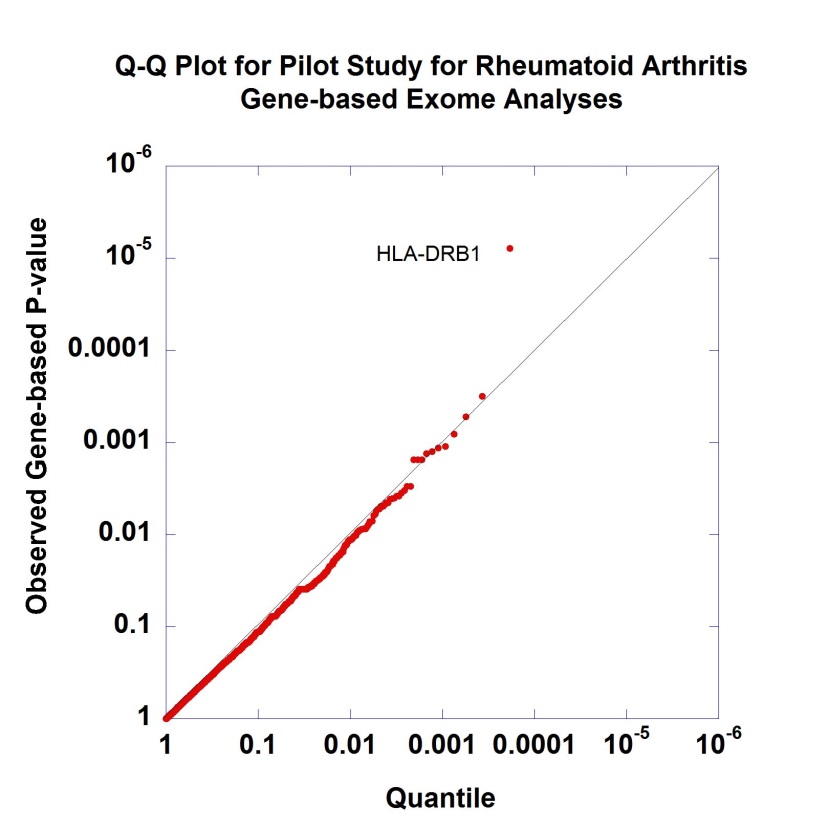
PKS <0.0001

PKS <0.0001





**Exome Genotyping Array Preliminary Data**

****

We have analyzed the exome beadchip data as part of several PMRP-based studies. ~250k nonsynonymous coding variants, >5k GWAS-significant SNPs, >5k grid of common SNPs, AIM SNPs, HLA tagging SNPs, and IBD SNPs are included on the beadchip. Various QC procedures have been applied to these data including PCA on AIM SNPs to identify genetic background outliers, determination of variants that pass quality control criteria (PASS), calculation of Hardy-Weinberg exact p-values, and kinship coefficient calculations to identify cryptic relatedness. In addition, we have phased all of these data using Beagle73 and used the resulting haplotypes to conduct scans of disease association. This Quantile-Quantile plot shows one such scan investigating rheumatoid arthritis under a recessive mode of inheritance using gene-based data. The positive control, HLA-DRB1 showed significant association.

**Specific Aim 2**

**To optimally classify cases and controls using genetics and circulating proteins in DNA and plasma obtained prior to clinical diagnosis.**

**Predictive Modeling:** Machine learning approaches have revolutionized the analysis of high dimensional data sets and this is particularly true with regard to developing robust classifiers. Bayesian Networks are an important, well-studied class of predictive modeling algorithms which, in general, produce accurate and transparent results. As other types of predictive models may perform well on these data, we will be flexible in our classifier approaches and also perform logistic regression with adaptive Lasso. Dr. Maadooliat has considerable expertise in employing high-dimensional logistic regression with L1 penalty procedures.

**Feature Selection:** To prime the molecular data for feature selection and greatly improve computational efficiency on these large data sets, we will apply Sure Independence Screening to the full set of molecular variables. Following SIS, we will use an Elastic Net logistic regression for feature selection across exome variants, GWAS-significant SNPs, cytokines, chemokines, metabolic proteins and antibodies.

**Bayesian Network Classifier:** Following feature selection, we will primarily use Bayesian Network classifiers to generate predictive models. Bayesian Networks are directed acyclic graphs, often used to model and analyze biological data. In essence, within the context of disease classification, Bayesian Networks calculate the PPD through Bayes theorem:

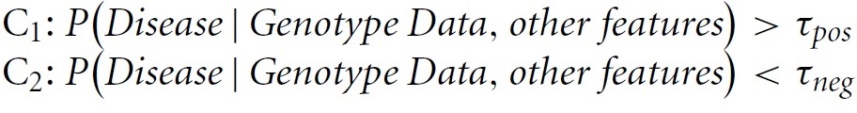
; where is the ith feature in the classifier and is an indicator variable for disease status. The degree to which the joint densities are factorized depends on the covariance structure of the set of features. Applying a given threshold for the PPD enables the classification decision. Drs. Maadooliat, Guo and Schrodi have experience using Bayesian Networks through the machine learning software Weka, Mathematica, as well as the R programming language.

**Penalized Logistic Regression with Adaptive Lasso:** As an alternative classification approach, we will use penalized logistic regression with adaptive lasso. High dimensional logistic regression with adaptive lasso has several compelling statistical features for analysis of these data. Logistic regression is a likelihood-base predictive model that is commonly used across numerous fields including genetics and bioinformatics for analysis of binary outcomes. A penalized logistic regression is incorporated for instances with a large number of features and small number of subjects.74 A popular penalty for model selection and estimation is lasso.75 In some scenarios with high-dimensional input, such as is the case with this study, the lasso variable selection can be inconsistent.76 To avoid this potential inconsistency, we will adopt an adaptive lasso, which is consistent under very general conditions.74,76

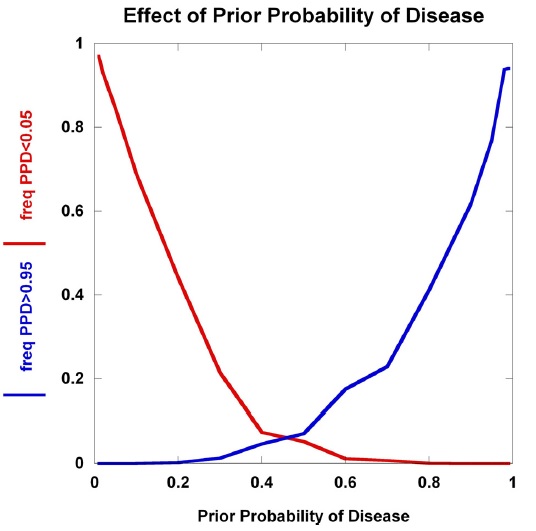
**Control for Overfitting:** A common pitfall encountered in predictive modeling is overfitting the data. 10-fold cross validation will be employed to reduce the effect of overfitting. The feature selection procedure will be embedded in the cross validation routine.

**Incorporation of Antigen/Antibody Profiles:** For rheumatoid arthritis and systemic lupus erythematosus sample sets, we will incorporate the results derived from the ADI antigen *in vitro* transcription/translation system as features for the Bayesian network and penalized logistic regression classifiers.

**Evaluating Diagnostic Utility of Classifiers:** Proper and informative evaluation of the performance of classifiers is critically important to understanding the potential utility of disease predictive models. The primary method that will be employed to do so will be the area under the ROC curve, or AUC. That said, we will explore methods to evaluate the diagnostic utility of the Bayesian Network classifier results that both capture the predictive capacity of the classification and have interpretability in a clinical setting. One such approach is to calculate a posterior probability of disease (PPD) using a prior density that reflects the risk in a population from which potential patients may be derived for initial evaluation by physicians (e.g., those with symmetrical joint pain being evaluated for possible rheumatoid arthritis). Setting thresholds for the density mass that exceeds the threshold (pos) and does not exceed a lower threshold (neg) is one such approach:

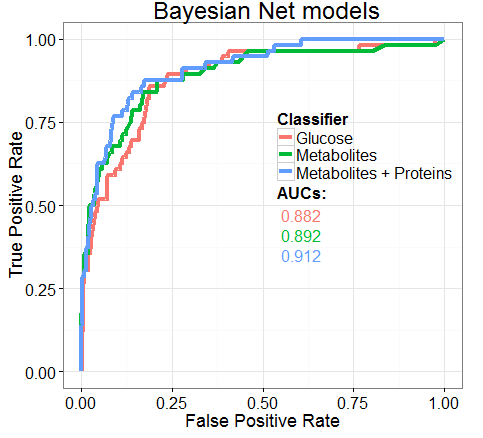
****

**Example:** Using the molecular data, we can calculate the posterior probability of RA for each individual. Thus, we would generate a distribution of posterior probabilities for RA. With perfect information, the PPD mass is concentrated at 1 and 0 (everyone either has or does not have RA). Suppose that clinicians would likely intervene therapeutically if their assessment of the probability of RA exceeds 0.95. That is, suppose a common clinical threshold for employing a particular therapeutic (e.g., methotrexate) is 0.95 probability of RA. Then, the proportion of individuals with PPD>0.95 should be close to the frequency of RA in the population evaluated for a classifier with high predictive capacity. Similarly, one can argue for the lower threshold. The plot to the left shows the impact of the expectation of the prior probability of disease on the PPD mass.

****

**Testing for Significant Clustering Within Each Disease:** To test for significant molecular subgroups within each disease, we will use a generalization of logistics PCA77 and biclustering approaches.78 The probabilistic nature of the logistic PCA model will allow us to extend the methodology to obtain a dimension reduction tool that handles both binary and continuous variables in a unified framework.77 Dr. Maadooliat is an expert in PCA methods and dimensional reduction.

**Preliminary Data on Type 2 Diabetes:** The Schrodi lab has recently conducted a predictive study of type 2 diabetes (T2D) using a panel of metabolites, immunological and metabolic markers, and T2D-associated genetic markers. The samples were obtained from PMRP and all case individuals (n=56) were selected to have a T2D diagnosis within 18 months following enrollment in the biobank. T2D diagnosis was determined by the Marshfield Clinic electronic medical record information and was defined as fasting glucose > 125 mg/dL, random plasma glucose > 200 mg/dL, and/or HbA1c exceeding or equal to 6.5%. Controls totaled 445 individuals who did not have any ICD9 codes for diabetes and all measurements of fasting glucose, random glucose and HbA1c were within the normal range. Using the stored plasma samples on all T2D cases and controls, a metabolite panel was measured by Metanomics Health. We measured concentrations of 6 inflammatory and metabolic proteins: C-peptide, adiponectin, IL-6, insulin, glucagon, and leptin. Using a combination of (i) forward selection that maximizes orthogonality between features, and (ii) an elastic net logistic regression model selecting features with the highest correlation with residuals, we select features to be used in the classifier. A Bayesian Network was generated for the selected features within each cross-validation run and an AUC calculated (avg AUC 0.912).17

****

The results of the Bayesian Network classifier for early T2D prediction shows high sensitivity and specificity using an elastic net for feature selection, followed by calculation of PPD incorporating correlative structure across the features. The addition of other data types, in this example proteins and additional metabolites, improves the averaged cross validated AUC values.

**Specific Aim 3**

**To validate the predictive model for rheumatoid arthritis in an independent sample set.**

The PMRP houses sufficient RA patients with diagnoses following sample collection and controls to split the sample into training and test sample sets. The same predictive modeling procedure will be used for the initial set of RA and controls as will be used for the remaining 7 diseases (i.e., SIS, elastic net feature selection, classifier development, and 10-fold cross validation incorporating the SIS/feature selection steps). The predictive model with the highest average cross validation AUC values will be tested in the independent hold-out sample set of 97 cases and 200 controls. A portion of these RA samples (n=60) and controls (n=100) will be classified by the predictive model utilizing the exome variants, circulating cytokines/metabolic proteins as well as the ADI protein microarray features. This will further test if the molecular panels measured and machine learning/statistical procedures have the capacity to produce a robust, accurate classifier for RA.

**Timeline**

**Year 1:** Obtain Marshfield IRB approval for the study. Partition exome data for the PMRP into data types (exome variants, GWAS-significant SNPs, AIMs). Define the PMRP cohort to be studied by applying exclusion criteria and performing the PCA analysis on the exome beadchip data to remove outliers. Identify cases and controls for all diseases with aid from Dr. Smith. Send plasma from 50% of the samples to Eve Technologies to run the Human Cytokine Array/Chemokine Array 42-Plex (HD42) and the Human Metabolic Hormone Array 9-Plex. Send plasma from 50% of the samples to Antigen Discovery for interrogation on the 800 antigen protein microarray platform.

**Year 2:** Send plasma from 50% of the samples to Eve Technologies to run the Human Cytokine Array/ Chemokine Array 42-Plex with IL-18 (HD42) and the Human Metabolic Hormone Array 9-Plex. Send plasma from 50% of the samples to Antigen Discovery for interrogation on the 800 antigen microarray platform. Retrieve and merge data from Eve Technologies and Antigen Discovery with exome variant data and phenotype data. With aid from Dr. Maadooliat and Dr. Guo, set up computational software and algorithms for running the association analyses, SIS, feature selection, and classifier development on the high performance computing cluster.

**Year 3:** Perform statistical and machine learning analyses on the merged data files. Set up binary PCA and biclustering algorithms for analysis of disease subgroups through clustering. Dr. Guo will also receive training from Drs. Schrodi and Maadooliat in analysis methods.

**Year 4:** Complete disease prediction for eight diseases and evaluate diagnostic utility measures. Select predictive model for validation sample set of rheumatoid arthritis. Perform analysis to identify molecular-based disease subgroups. Mentor and train Dr. Guo (postdoctoral fellow on the study). Development of novel methods for prediction, clustering through dimensional reduction, and investigate new diagnostic utility metrics. Work with Dr. Smith to interpret biological implications of results.

**Year 5:** Complete analysis of disease subgroups. Write manuscripts and present at national conferences. Coordinate with Drs. Maadooliat, Guo, Trappl-Kimmons and Liang on analysis results, the performance of novel methods. Identify new hypotheses that are generated from these results and apply for funding to continue investigation in new direction. Work with Dr. Smith to interpret clinical implications of results.