Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jeffrey Chou Date

Discovery of Airway Fluid Proteins Associated with Progressive Lung Disease and Damage in Early Childhood Cystic Fibrosis

By

Jeffrey Chou

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Limin Peng, Ph.D.

Committee Chair

_____

Rabindra Tirouvanziam, Ph.D.

Committee Member

_____

Joshua Chandler, Ph.D.

Committee Member

Discovery of Airway Fluid Proteins Associated with Progressive Lung Disease and Damage in Early Childhood Cystic Fibrosis

By

Jeffrey Chou

B.S.

University of Georgia

2013

Thesis Committee Chair: Limin Peng, Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2019

ABSTRACT

Discovery of Airway Fluid Proteins Associated with Progressive Lung Disease and Damage in Early Childhood Cystic Fibrosis

By Jeffrey Chou

Background: Lung disease is typically the landmark clinical presentation in cystic fibrosis (CF) patients. However, much is still unknown regarding which immunological signaling proteins truly contribute to airway pathology, especially in young children. To assess such relationships, bronchoalveolar lavage (BAL) samples and lung computed tomography (CT) scans were obtained from a longitudinal cohort of CF children at one, three, and five years of age. Concentrations of select BAL proteins were determined using the Olink® Immuno-Oncology assay and overall lung disease from CT scans was quantified by the PRAGMA-CF (Perth-Rotterdam Annotated Grid Morphometric Analysis) method. Components of the PRAGMA score include PRAGMA-%Dis for total airway disease and PRAGMA-%Bx scores for bronchiectasis.

Methods: Spearman correlation coefficients were used to evaluate cross-sectional relationships of individual protein correlations with PRAGMA scores. A random-intercept model with AR(1) covariance structure was then built, with guidance from contrast tests for evaluating the effects of each individual protein on PRAGMA score at given time points. Calculated p-values for Spearman correlations and contrast tests were adjusted with the Benjamini-Hochberg method to control for false discovery rate (FDR). Finally, a lasso penalized regression algorithm for mixed models was utilized to select proteins that lead to a parsimonious model for predicting PRAGMA scores.

Results: After FDR correction, three BAL proteins (ARG1, CCL4, and CSF1) exhibited significant positive correlations with PRAGMA-%Dis in the cross-sectional analysis of the five-year-old subset. These findings are corroborated by the contrast test results based on the linear mixed model with age set to five years. In addition, the lasso method suggests that higher HGF, lower LAMP3, and older age are predictive of increased PRAGMA-%Dis, whereas the selected predictors for increased PRAGMA-%Bx include higher ICOSLG, higher TNFRSF9, lower LAMP3, and older age. These proteins had significant effects in the linear mixed models' contrast tests except for the effect of LAMP3 on %Dis.

Conclusions: Our analyses demonstrate that select proteins have promising utility in predicting airway disease and damage in young CF children, both in a cross-sectional and longitudinal context. However, more research is needed to establish causal relationships for therapeutic drug development and improvement of precision medicine models.

Discovery of Airway Fluid Proteins Associated with Progressive Lung Disease and Damage in Early Childhood Cystic Fibrosis

By

Jeffrey Chou

B.S.

University of Georgia

2013

Thesis Committee Chair: Limin Peng, Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2019

ACKNOWLEDGMENTS

As an outdoor enthusiast, I liken the process of writing a thesis to hiking up a long mountain trail. It can be quite a daunting and arduous task, but nevertheless totally worth the panoramic views at the top. I could not have made it without all my thesis advisors, coworkers, friends, and family who helped me throughout the process. There were trying and difficult times, but at last I have made it to the finish line.

First and foremost, I would like to thank Dr. Limin Peng, Dr. Rabin Tirouvanziam, and Dr. Joshua Chandler for all their time in mentoring me throughout my time as a research assistant. I learned so much from all of them and definitely feel I have grown so much in terms of my statistical and collaboration skills. It was an absolute honor and privilege to have worked with them.

I would also like to thank Dr. Hamed Horati at Erasmus University in Rotterdam, The Netherlands for providing the dataset for this thesis. I much appreciate his time in answering my questions about the dataset and providing helpful discussion for our analysis.

I will fondly remember my days as a student at Emory's biostatistics department and will miss its tight-knit community. I believe everything I learn here will be instrumental in my future success and thank all my professors for being a part of it. There are so many exciting opportunities now in the "big data" era, and I look forward to my future adventures.

TABLE OF CONTENTS

1. INTRODUCTION

Cystic fibrosis (CF) is a lethal genetic disorder that severely hampers lung function as well as the gastrointestinal, endocrine, and reproductive systems.[1] CF is caused by a mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) protein that regulates secretion of vital bodily fluids such as mucus, sweat, saliva, and tears.[2] These secretions often become thickened and sticky in the loss of CFTR function, leading to chronic damage of associated internal organs.[3] In the respiratory system, the obstruction with mucopurulent sputum causes recurring bacterial infections, progressive inflammation, and physical exacerbations from coughing and wheezing.[3] The CFTR mutation also impairs digestive ability and pancreatic insulin production, resulting in symptoms such as malnutrition, diarrhea, weight loss, and CF-related diabetes (CFRD).[3] CF is inherited in an autosomal recessive manner and is primarily found in Caucasians.[1] Although incidence of cystic fibrosis in endemic populations is relatively rare (1 in 2,500 live births in the E.U. and 1 in 3,500 live births in the U.S.), it is still one of the deadliest genetically-inherited diseases.[1,4] While the average life expectancy has increased from less than five years in the 1940's to a median of 40.7 years in 2013, more research is needed to further improve quality of life and survival outcomes for the CF population.[5]

In particular, the inflammatory response in airways has long been the hallmark pathological feature of CF and is a key therapeutic target for limiting lung disease and damage.[6] Shortly after birth, individuals with CF experience chronic inflammation and bacterial infections due to poor sputum drainage in the lungs. The inflammation is not effective at clearing bacterial infections and may persist even in the absence of pathogens.[7,8] Several immunological cell subsets, including but not limited to neutrophils,

play a key role in CF inflammatory response. Neutrophils release powerful enzymes such as neutrophil elastase (NE) and myeloperoxidase that injure the respiratory lumen and have been shown to correlate with early structural damage of airways in young children (1-5 years old).[7,9-12] Recent findings indicate that neutrophils recruited to CF airways undergo pathological reprogramming caused by signaling proteins that allow them to survive longer and co-exist with pathogens, rather than directly attacking them.[10] However, much is still unknown about mechanistic pathways involving these proteins and which ones significantly correlate with CF airway disease and damage.[10,13]

To aid in the confirmation and discovery of such signaling mechanisms, protein biomarkers can be measured in clinical fluid samples such as bronchoalvelor lavage fluid (BALF). Children undergoing BAL are placed under sedation, after which a bronchoscope inserted into the lungs for a saline wash. The resulting BAL is then collected and sent for laboratory analysis, where it is centrifuged to produce cell-free fluid (BALF).[14] While the BAL is invasive, expensive, and difficult to perform, it is nonetheless regarded as the gold standard for airway biomarker measurement and provides excellent capture of inflammatory markers that correlate significantly with airway disease, such as NE and interleukin (IL)-8.[15] Sputum collection also offers good biomarker detection, but most young children have difficulty expectorating sputum even upon induction.[15] Blood plasma or serum is occasionally used, but this only captures biomarkers in the general circulation and may not reflect actual inflammatory mechanisms occurring in the airways.[15,16]

Airway disease and damage in young CF children can be assessed through a computed tomography (CT) scan and quantified by the Perth-Rotterdam Annotated Grid

Morphometric Analysis (PRAGMA).[17] In the PRAGMA procedure, physicians outline a grid over the CT scans and grade each cell for attributes such as percent bronchiectasis (structural damage of the airway) and percent airway disease (of any severity).[17] PRAGMA scores demonstrate excellent intra-class correlation coefficients (ICC, >90%) among physicians and perform well in quantifying the extent of lung disease and damage in children below six years of age.[17]

However, very few studies have reported on a broad analysis of signaling proteins in BAL samples that may associate with the inflammation response and airway damage in young CF children. One recent analysis by DeBoer et al.[18] attempted to measure 1,129 proteins from blood plasma using the SOMAScan® targeted proteomics platform, and then analyzed significant correlations with bronchiectasis. However the cohort included in that study consisted of older children (aged 7-15 years) and was only conducted cross-sectionally. As of today, there is a critical need to identify biomarkers that can aid in monitoring CF lung disease progression and spur development for novel inflammation therapies, especially in younger children (1-5 years).[16] Early detection and treatment is paramount to impeding the progression of inflammation-induced damage as much as possible. Furthermore, there is a clear need for a longitudinal cohort study in order to provide greater statistical power and the ability to monitor and predict disease progression for any particular child over time.

The objective of this thesis is to identify signaling proteins in BAL samples that correlate significantly with airway disease and bronchiectasis from a longitudinal cohort of young children with CF. To accomplish this, we obtained a dataset of protein concentrations and PRAGMA-CF Scores from the I-BALL early CF monitoring program

at Erasmus University in Rotterdam, Netherlands, a joint collaborative effort with Emory's own early CF monitoring program (IMPEDE-CF). The I-BALL cohort consists of young children with CF who had clinical measurements taken at bi-yearly intervals between one and five years of age. The Erasmus investigators utilized the Olink® high-throughput targeted proteomic assay, which employs protein extension assay (PEA) technology for detection of 92 biomarkers.[19] In PEA, patient samples are incubated with a pair of antibodies that target specific proteins.[20] If the DNA sequences of the antibodies matches with the protein, then amplification of the sequence occurs and final protein concentration is assessed by quantitative real-time polymerase chain reaction (qPCR).[19,20]

Our analyses aim at tackling the questions of whether and how airway disease and damage and their trajectory over time associate with BAL proteins and other potential demographic predictors, such as age. Narrowing down potential predictors is crucial due to the high cost of multiplex proteomic assays such as Olink® and SomaScan®. In this thesis project, the effects of each individual protein were first be evaluated in a both cross-sectional and longitudinal settings through the use of Spearman correlations and linear mixed models. This portion provided validation for known correlates of airway disease and damage, as well as guidance for building multivariate prediction models. Then, lasso penalized regression methods under linear mixed models were applied to select proteins that are predictive of airway disease and damage over time. This selection yielded a final multivariate longitudinal model that achieves easily interpretable prediction of PRAGMA scores from a parsimonious subset of proteins. Ultimately, it is expected that insights from this model will provide deeper understanding of the immunological signaling process at play in CF airways, help unlock key targets for

therapeutic drug development, and provide criteria for personalized diagnosis and treatment in a precision medicine context.

2. METHODS

2.1 Data Acquisition and Cleaning

Data for this study was integrated from three different sources: (1) I-BALL clinical demographics, (2) Olink® protein concentrations, and (3) PRAGMA-CF scores. All datasets shared a two-level hierarchical row structure such that measurements at each age (designated as 'Study Event' with level 'E2' for one year, 'E3' for two years, …, and 'E6' for five years) are nested within each patient (i.e., where *i* = 'Subject ID' and *j* = 'Study Event'). Each dataset was stored as a CSV file on the consortium's Isilon® server and then imported into RStudio GUI (Graphical User Interface) for data cleaning and analysis. The following subsections will detail specific steps for cleaning each of the three input datasets. After the cleaning process, all datasets were merged by a common identifier which includes the subject ID and the study event.

2.1.1 I-BALL Clinical Demographics

Clinical demographics from Erasmus' I-BALL cohort were originally recorded on OpenClinica® Data Management software. Clinical variables extracted from the file include the birth date, date of BAL, date of CT scan, and age in months. Patients who had more than a 30-day difference in date of BAL procedure vs. CT scan were excluded from analysis, since there can be dramatic discrepancies in airway pathology within that time frame. The mean difference between BAL and CT visits was 7.13 ± 7.58 days for the included data.

2.1.2 BALF & Olink Proteomics

BALF was collected for each patient at two different sites in the airways, designated as 'B2' and 'B4'. The 'B2' site corresponds to BALF collected from the right

middle lobe of the lungs, which is the default for all patients who underwent the BAL procedure. Some patients have a 'B4' measurement taken from various other parts of the lung such as the most affected lobe of the lung or the lingula (upper lobe of left lung), depending on clinician discretion. Protein concentrations of BALF from 'B2' and 'B4' (if available) were quantified using the Olink® assay. However, due to major lapses in data collection for 'B4' measurements, only the 'B2' protein samples will be included for this thesis.

Whereas the I-BALL and PRAGMA datasets have a two-level hierarchical row structure, the original Olink dataset has a third level to denote the site of BALF collection (i.e., where i = 'Subject ID', j = 'Study Event', and k = 'BAL Site'). Each column represents a different protein. In addition, each protein has a unique limit of detection, which is the minimum concentration that the assay can quantify reliably. The first data cleaning step involved imputation of protein measurements below the LOD as ½ of the LOD. Next, an indicator column for each protein was created such that '0' denotes a measurement at or above LOD and '1' denotes a measurement below LOD. These columns were used to calculate percent of measurements below LOD for each protein. Proteins that have no more than 20% measurements below LOD were included for analysis, and proteins that have between 20 and 50% measurements less than LOD were saved for a potential future analysis where proteins are analyzed as a categorical variable (above or at LOD vs. below LOD) instead of a continuous variable. All proteins are listed in Appendix B along with their calculated % below LOD values. Out of the 92 total proteins, 62 of them were selected for further analysis in this study.

2.1.3 PRAGMA Scores

Two types of PRAGMA scores, PRAGMA-%Dis (total airway disease) and PRAGMA-%Bx (airway structural damage), were recorded for each patient in three different batches by the same rater. Thus, some CT scans were scored more than once. If a patient has PRAGMA scores in batches one and two, the first batch was used for analysis since it had a larger sample size and will minimize variance contributed by batch effects. PRAGMA scores from the third batch was only used if no score was produced in the first or second batches. Thus, the PRAGMA scores primarily come from the first batch.

2.2 Summary Statistics

Summary statistics were calculated for PRAGMA scores (both %Dis and %Bx) and clinical variables of interest. The mean and standard deviations were reported for continuous variables such as age and PRAGMA scores. Frequencies were reported for categorical variables such as gender and infection status. All summary statistics were stratified by the study event to evaluate possible trends in variables across the different ages.

2.3 Cross-Sectional Spearman Correlations

At each age, the Spearman correlation coefficient (ρ) was used to determine which proteins have significant relationships with PRAGMA-%Dis or %Bx scores. This analysis will also reveal the directionality of the relationships at each time point and provide guidance for building longitudinal models. The Spearman Rho is a non-parametric rank-based algorithm that is robust against outliers. For this reason, the Spearman Rho was chosen over the parametric Pearson correlation coefficient (r) since

several proteins are nearly 20% below LOD and possibly left-skewed as a result. To address the multiplicity of comparisons across the 62 proteins, the associated p-values were adjusted by the Benjamini-Hochberg method to control for false discovery rate (FDR) when screening for significant correlates.[21]

2.4 Linear Mixed Models of PRAGMA Scores to Evaluate Individual Protein Main Effects

To evaluate the longitudinal relationship between individual protein concentrations and PRAGMA-%Dis or PRAGMA-%Bx scores while controlling for age, two different random-intercept linear mixed models were proposed with repeated measures of each subject as the random effect. The models were fit using the *lme()* function from the *nlme* package. Parameter beta coefficients and associated p-values were calculated using the residual maximum likelihood (REML) method. REML yields less biased estimates than normal maximum likelihood (ML), especially in smaller samples, since it accounts for the loss in degrees in freedom when estimating fixed effects.[22,23] In addition, the first-order autoregressive (AR(1)) structure was specified for the covariance matrix, which is suitable for longitudinal scenarios where observations closer in time are more strongly correlated than observations father apart in time.[22,23]

2.4.1 Model I: Protein concentration and age as a continuous covariate

Let i denote the Subject ID and j denote the study event. Model I includes the main effects of protein concentration (x_{ij}) and age (i.e. exact age in months at BAL procedure) on PRAGMA-%Dis or PRAGMA-%Bx (Y_{ij}) and is specified as follows:

$$Y_{ij} = \beta_0 + \theta_i + \beta_1 \cdot x_{ij} + \beta_2 \cdot age_{ij} + \epsilon_{ij}$$

The parameter θ_i is the random intercept for each i^{th} child that accounts for between-subject variance, and ϵ_{ij} is the error term that accounts for within-subject variance. For this analysis, the main effect of interest is that of protein concentration on PRAGMA score. In model I, the corresponding interpretation is that PRAGMA score increases by the constant β_1 for every one-unit increase in protein concentration, while keeping age fixed at a certain level. It is possible, however, that the relationship between protein concentration and PRAGMA score is not the same for all ages. Thus, a limitation of model I is that it assumes the same main effect of protein on PRAGMA score regardless of age. To remedy such an assumption, an interaction term between protein concentration can be added.

2.4.2 Model II: Protein concentration, age as a continuous covariate, and interaction of protein*age

Model II is similar to model I, but also includes an interaction term between protein concentration and age ($x_{ij} \cdot age_{ij}$):

$$Y_{ij} = \beta_0 + \theta_i + \beta_1 \cdot x_{ij} + \beta_2 \cdot age_{ij} + \beta_3 \cdot x_{ij} \cdot age_{ij} + \epsilon_{ij}$$

The interaction term allows the effect of protein on PRAGMA score to vary by different age levels. To illustrate, the model is rearranged in the following form where the protein terms are grouped together:

$$Y_{ij} = \beta_0 + \theta_i + (\beta_1 + \beta_3 \cdot age_{ij}) \cdot x_{ij} + \beta_2 \cdot age_{ij} + \epsilon_{ij}$$

Thus, the PRAGMA score now increases by $\beta_1 + \beta_3 \cdot age_{ij}$ for every one unit increase in protein concentration and can be different for each age category. Since nearly all patients had measurements at approximately 12 months (E2), 36 months (E4), or 60 months (E6), a few contrast tests were run to determine if the effects of the protein are

significantly different from zero at each age group. The contrast tests were conducted using the *glht()* function of the *multcomp* R package and were defined as follows:

$$H_{0a}: \beta_1 + 12 \cdot \beta_3 = 0$$

$$H_{0b}: \beta_1 + 36 \cdot \beta_3 = 0$$

$$H_{0c}: \beta_1 + 60 \cdot \beta_3 = 0$$

The p-values from the contrast tests were also FDR adjusted if the interest lies in inferring the protein has an effect on the PRAGMA score at one or more time points (or ages).

2.5 Lasso Penalized Regression for Building Multivariate Linear Mixed Models of PRAGMA Scores

The previous models can reveal whether a given protein has a significant longitudinal association with PRAGMA score. However, a limitation is that they only consider one protein at a time. Solely applying multiple comparison adjustment to such univariate analysis results often gives over-conservative results. In the context of the immune response in CF airways, it is well understood that multiple, likely correlated, signaling proteins play a role in airway disease and damage. Multivariate analysis can help reveal how they jointly impact the PRAGMA scores. However, the dimensionality issue need to be properly handled given the number of proteins is greater than the number of subjects.

Traditional model selection techniques such as forwards, stepwise, and backwards selection tend not work well in high-dimensional datasets (where $p > n$).[24] Instead, penalized regression methods such as the lasso (least absolute shrinkage and selection operator) was used to shrink parameter estimates of insignificant variables to zero.[24]

These methods can often gain estimation efficiency while achieving shrinkage estimation that leads to reasonable variable selection.[24]

In linear mixed models, the lasso method seeks to find the beta coefficients that minimize the quantity $-\log(L) + \lambda \cdot \sum_{j=1}^{p} |\beta_j|$, where L is the maximum likelihood estimate, λ is the shrinkage parameter, and $\sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 penalty. As λ increases, more beta coefficients are shrunken to zero. Model evaluation techniques such as the k-fold cross-validation (CV) and Bayesian Information Criteria (BIC) are used to choose the optimal λ parameter.

The R package *lmmlasso*, created by Schelldorfer et al[25], was utilized for protein selection in this study. The algorithm utilizes a maximum likelihood-based ℓ_1 penalization that is optimized for fixed effects in high-dimension datasets and predicts random effect coefficients via the maximum *a posteriori* (MAP) principle.[25] In addition, the package recommends choosing the optimal λ at the lowest BIC.[25]

2.6 Transparency Statement

To promote transparency and ensure reproducibility of results, we intend to make all datasets and R code publicly available online after peer-reviewed publication.

3. RESULTS

3.1 Summary Statistics

Table 1: Summary Statistics of Cohort

Variable	Study Event				
	E2 (n=9)	**E3 (n=1)**	**E4 (n=14)**	**E5 (n=1)**	**E6 (n=13)**
Age at BAL (months)					
Mean (SD)	13.2 (0.499)	25.0 (NA)	37.4 (0.767)	49.3 (NA)	61.4 (0.651)
PRAGMA-%Dis					
Mean (SD)	1.62 (0.768)	3.89 (NA)	2.57 (1.29)	2.33 (NA)	3.65 (2.47)
PRAGMA-%Bx					
Mean (SD)	0.226 (0.261)	1.23 (NA)	0.586 (0.495)	1.08 (NA)	1.30 (1.18)
Gender					
Female	5 (55.6%)	1 (100%)	9 (64.3%)	1 (100%)	5 (38.5%)
Male	4 (44.4%)	0 (0%)	5 (35.7%)	0 (0%)	8 (61.5%)
CFTR Mutation					
F508del Homozygous	4 (44.4%)	0 (0%)	7 (50.0%)	0 (0%)	6 (46.2%)
F508del Heterozygous	4 (44.4%)	1 (100%)	6 (42.9%)	1 (100%)	6 (46.2%)
Other	1 (11.1%)	0 (0%)	1 (7.1%)	0 (0%)	1 (7.7%)
Infection Present					
Yes	1 (11.1%)	0 (0%)	5 (35.7%)	0 (0%)	10 (76.9%)
No	8 (88.9%)	1 (100%)	9 (64.3%)	1 (100%)	3 (23.1%)

Of subjects that have complete Olink® and PRAGMA information, the Erasmus CF cohort has a total size of n=38 measurements, with n=9, n=14, and n=13 observations at the E2, E4, and E6 time points, respectively. There is one patient from whom clinical measurements were obtained at E3 and E5. Summary statistics for clinical variables of

interest are presented in table 1 and are stratified by study event. As expected, mean PRAGMA % Dis and % Bx scores increase as age increases. Overall frequencies of gender and CFTR mutation appear to be evenly distributed. However, the proportion of subjects with pathogenic bacterial infection increases with age (11.1% at E2, 35.7% at E4, and 76.9% at E6).

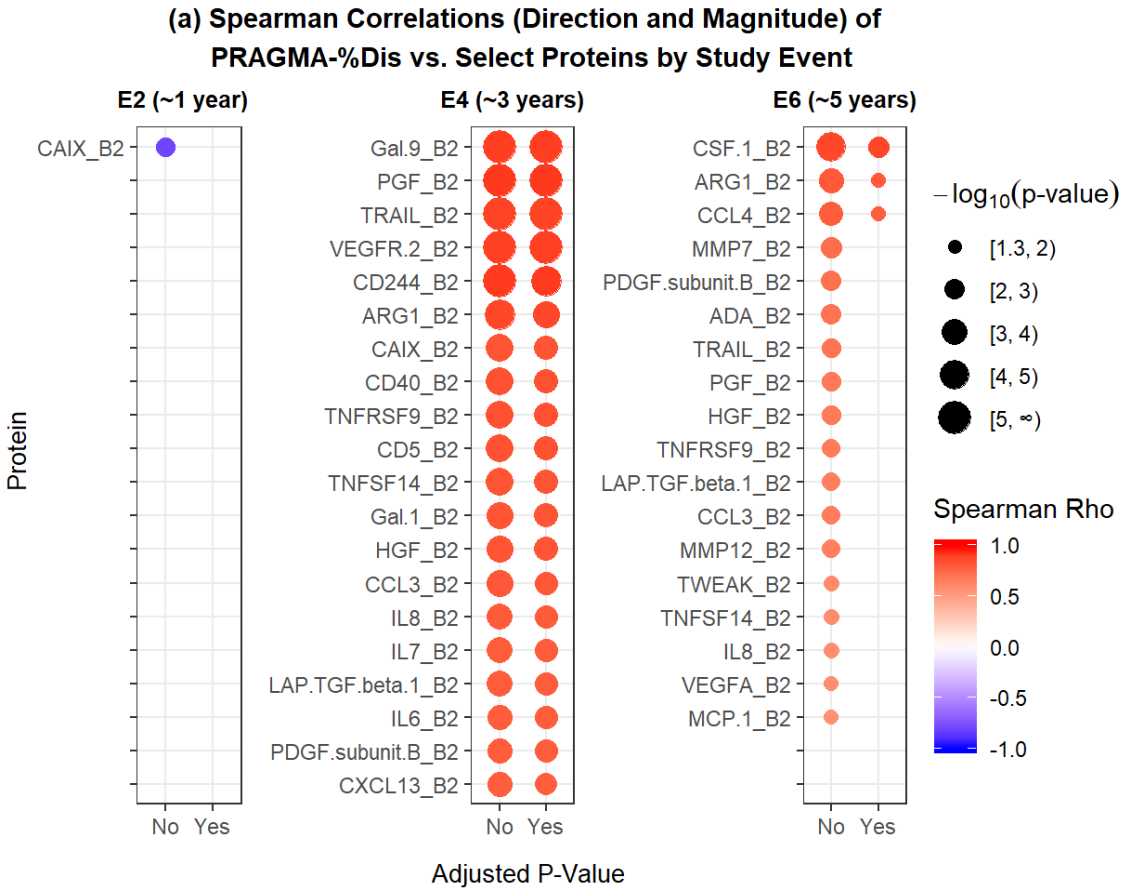
A potential limitation of this study is that only n=6 patients have two repeated measurements and most patients have only one measurement (n=26). No patients have complete records across E2, E4, and E6. However, this study is currently ongoing and more data will be collected in the future.

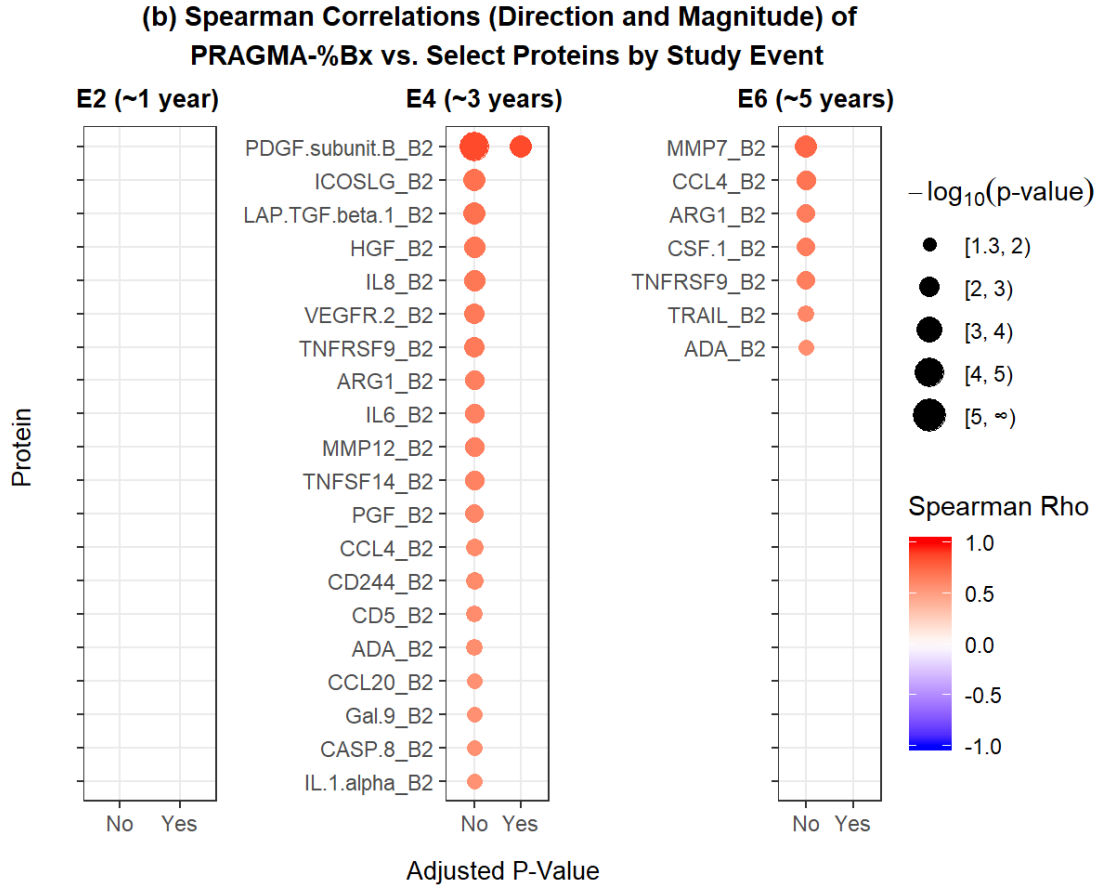
3.2 Cross-Sectional Spearman Correlations

Significant Spearman correlations between PRAGMA-%Dis or %Bx vs. select proteins across the time points are visualized below in figure 1a and 1b. The exact correlations of all proteins are printed in tables 4a and 4b of the appendix. At E2, nearly all proteins are negatively correlated but insignificant (at $\alpha = 0.05$) vs. PRAGMA scores, except for CAIX (insignificant after FDR adjustment). However at E4, a large majority of proteins are significantly (positively) correlated with PRAGMA-%Dis, even after FDR adjustment (46 out of 62 proteins). The correlations are also positive at E6, but after FDR adjustment, only CSF.1 (rho = 0.868, p = 0.008), ARG1 (rho = 0.808, p = 0.043), and CCL4 (rho = 0.791, p = 0.043) remain significant.

Similar trends are seen for the proteins vs. PRAGMA-%Bx. At E2, most of the proteins are insignificant, but instead have weak positive correlations. Stronger positive correlations are seen E4 and E6. However, there virtually all proteins are insignificant after FDR adjustment across each time point except for PDGF.subunit.B at E4.

Figure 1: Bubble plots to illustrate the direction and magnitude of cross-sectional spearman correlations of (a) PRAGMA-%Dis or (b) PRAGMA-%Bx versus protein concentrations, as stratified by study event. For each study event, the proteins were ranked by the strength of unadjusted p-value and the top 20 proteins were chosen for display. FDR adjusted p-values are also displayed for each study event. The magnitude of correlations is illustrated by the size of the bubbles, which is defined by the negative $\log_{10}$ of the p-value. Only significant correlations at $-log_{10}(0.05) \approx 1.3$ or greater are shown; otherwise the bubble is hidden. In addition, the directionality of the association of illustrated by the color of the bubble, where red indicates a positive correlation and blue indicates a negative correlation.





3.3 Linear Mixed Models of PRAGMA Scores to Evaluate Individual Protein Main Effects

3.3.1 Model I: Protein concentration and age as a continuous covariate

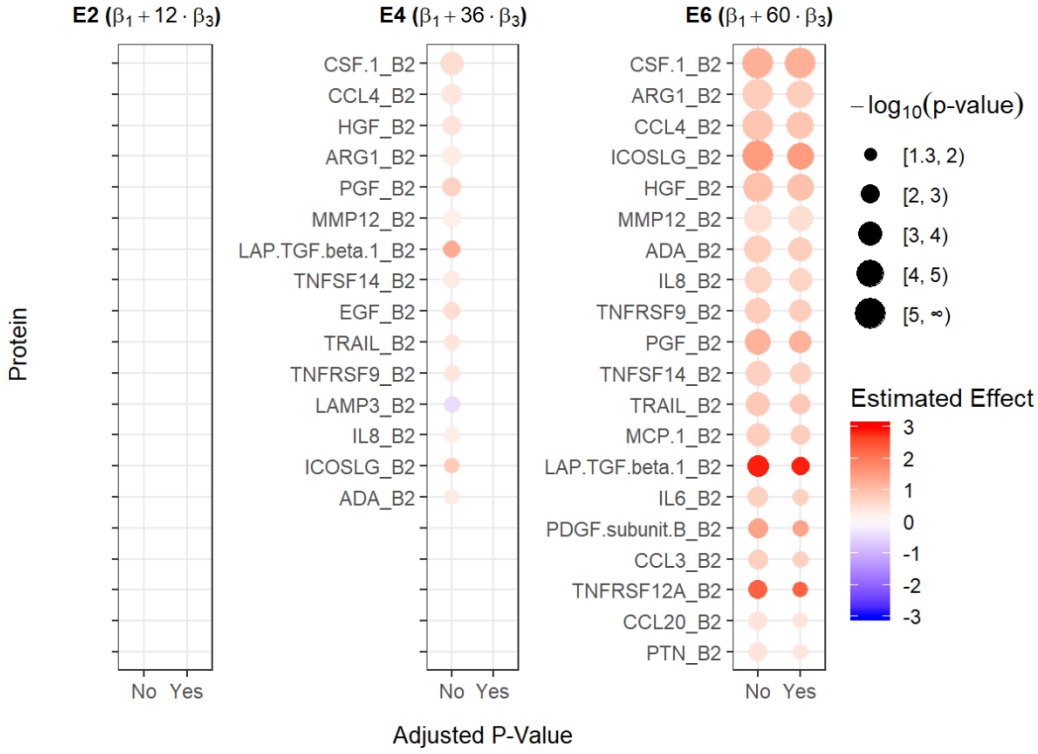
The coefficient estimates and p-values for β_1 (protein concentration) of model I are listed in table 5 for PRAGMA-%Dis and PRAGMA-%Bx. For both responses, there are several proteins that have significant β_1 coefficients (10 out of 62 for both % Dis and % Bx). However, model I may not adequately capture changes in strength of associations across the different time points. Model II will attempt to address this shortcoming.

3.3.2 Model II: Protein concentration, age as a continuous covariate, and interaction of protein*age

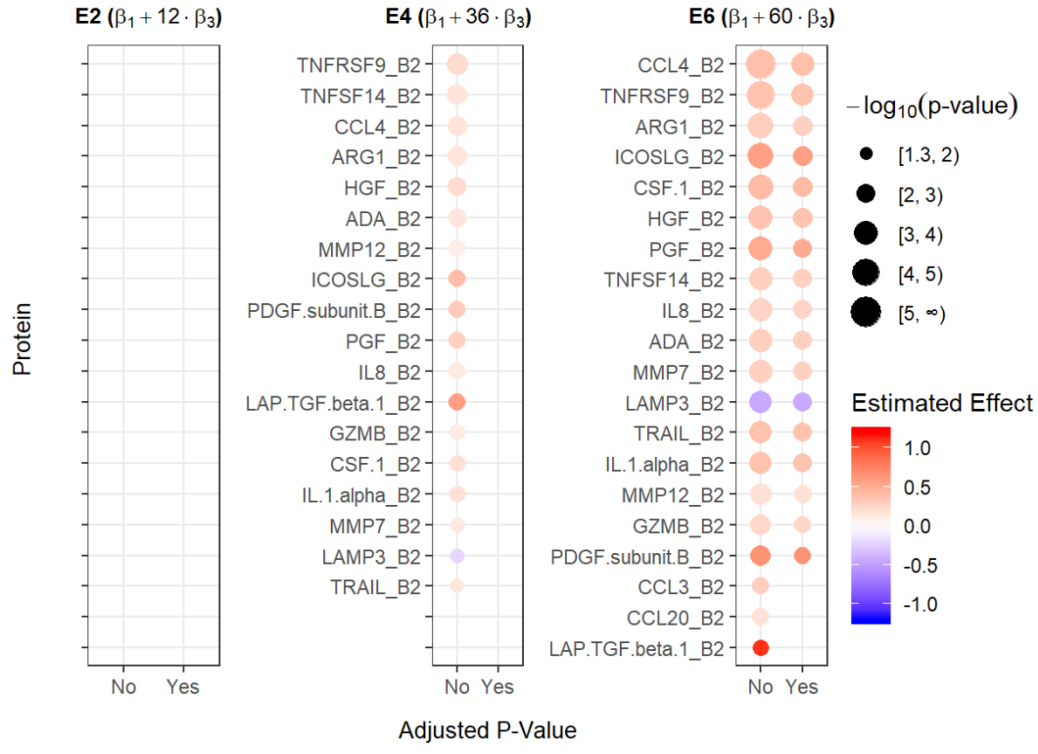
Table 6 lists the coefficient estimates and p-value for β_1 (protein concentration) and β_3 (interaction of protein*age) of PRAGMA-%Dis and PRAGMA-%Bx, respectively. Virtually none of the reported p-values are significant. However, in order to truly evaluate the significance of the main effect of each protein on PRAGMA score with an interaction term, the contrast tests were run with results reported in table 7a and 7b for PRAGMA-%Dis and %Bx, respectively. The results are also visualized in figure 2, below.

Figure 2: Bubble plots to illustrate the direction and magnitude of the linear mixed model contrast tests at each time point for (a) PRAGMA-%Dis and (b) PRAGMA-%Bx. Figure 2 is similar to figure 1 except that contrast estimates are displayed instead of Spearman rho. FDR adjusted p-values are also presented with the unadjusted p-values of contrast estimates. The color of the bubbles indicates the directionality of contrast estimates, where dark red bubbles are strong positive estimates and dark blue bubbles are strong negative estimates. In addition, the sizes of the bubbles illustrate the magnitude of contrast p-values, where larger bubbles indicate higher significance and smaller bubbles indicate lower significance. Insignificant contrasts with p-values (adjusted or unadjusted) below $-log_{10}(0.05) \approx 1.3$ are hidden from the display.

(a) Model II Contrasts (Direction and Magnitude) of PRAGMA-%Dis vs. Select Proteins by Study Event



(b) Model II Contrasts (Direction and Magnitude) of PRAGMA-%Bx vs. Select Proteins by Study Event



Overall, the results of the contrast tests, as illustrated in figure 2, somewhat reflect the overall trends seen in the cross-sectional analysis (figure 1). At E2, all of the effects of protein concentration on PRAGMA score are not significant. From E2 to E4, a change in magnitude of contrasts is also present, although not as pronounced as the cross-sectional correlations. A moderate number of proteins (18 out of 62) have significant effects at E4, but none of the p-values are significant after adjustment. Ending at E6, a larger number of proteins demonstrate significant effects (25 out of 62). After FDR adjustment a slightly smaller number of effects are significant (17 out of 62). Similar to the cross-sectional correlations, CSF.1 (estimated effect = 1.236, $p < 0.001$ for %Dis; estimated effect = 0.425, $p = 0.001$ for %Bx), CCL4 (estimated effect = 0.911, $p < 0.001$ for % -Dis; estimated effect = 0.391, $p < 0.001$ %Bx), and ARG1 (estimated effect = 0.787, $p < 0.001$ for %Dis; estimated effect = 0.300, $p < 0.001$ for %Bx) demonstrate significant effects at E6 in the longitudinal model. Other proteins of interest include LAMP3 (estimated effect = -0.442, $p = 0.009$ for %Bx, insignificant for %Dis), HGF (estimated effect = 0.981, $p < 0.001$ for %Dis; estimated effect = 0.374, $p = 0.008$ for %Bx), and ICOSLG (estimated effect = 1.536, $p < 0.001$ for %Dis; estimated effect = 0.593, $p = 0.006$ for %Bx), all of which are robust parameters in the upcoming lasso penalization regression. All aforementioned proteins have positive effects on PRAGMA score except for LAMP3, which has a negative effect.

3.4 Lasso Penalized Regression for Building Multivariate Protein Linear Mixed Models of PRAGMA Scores

Figures 3a-4b visualize the results of the lasso penalized regression of PRAGMA-% Dis and %Bx over proteins. For each outcome, the ‘a’ figures illustrate variation of

selection criteria AIC, BIC, and deviance across possible lambda values (i.e., values of the penalization parameter), and the 'b' figures illustrate penalization of beta estimates across lambda values. The selection of lambda does not start from zero (i.e., no penalization) because fitting the unpenalized model would fail given the sample size is less than the number of proteins. The selection of lambda starts at 4.2 for % Dis and 8.0 for %Bx.

Figure 3a: Selection criteria for lasso-shrinkage models of PRAGMA-%Dis. Estimated AIC, BIC, and deviance statistics are plotted possible values of the lambda shrinkage parameter. The *lmmlasso* package algorithm fails to produce models for lambda values less than 4.2, hence the non-zero starting value of lambda. It appears that the lowest BIC occurs at 5.8, which is denoted by the dashed gray line.

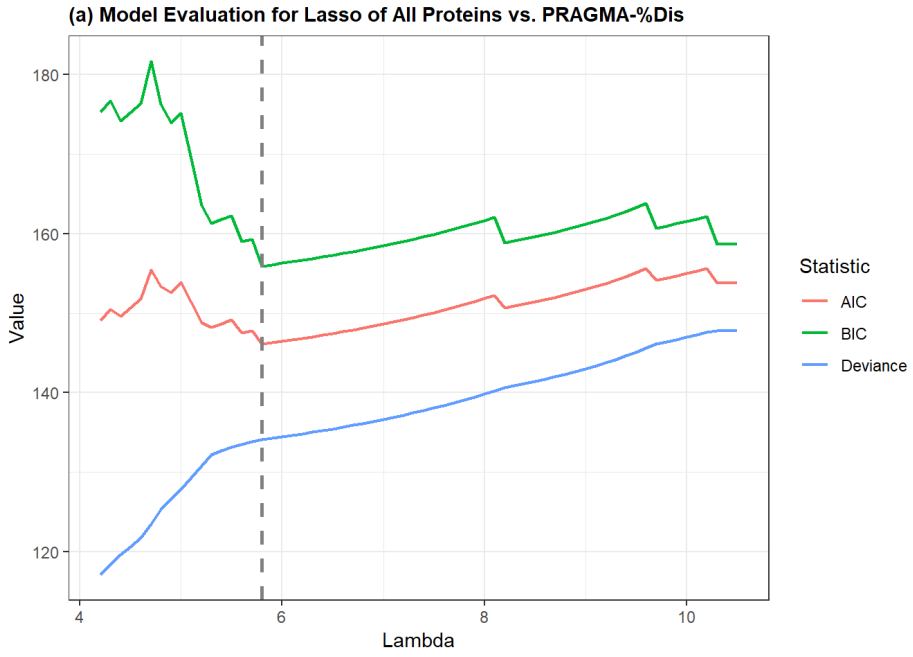


Figure 3b: Parameter estimates plotted against possible lambda values. As lambda increases, the parameter estimates shrink closer to zero and the number of parameters in the overall model decreases. The dashed line indicates the model with the lowest BIC, which includes the parameters HGF, age, and LAMP3.

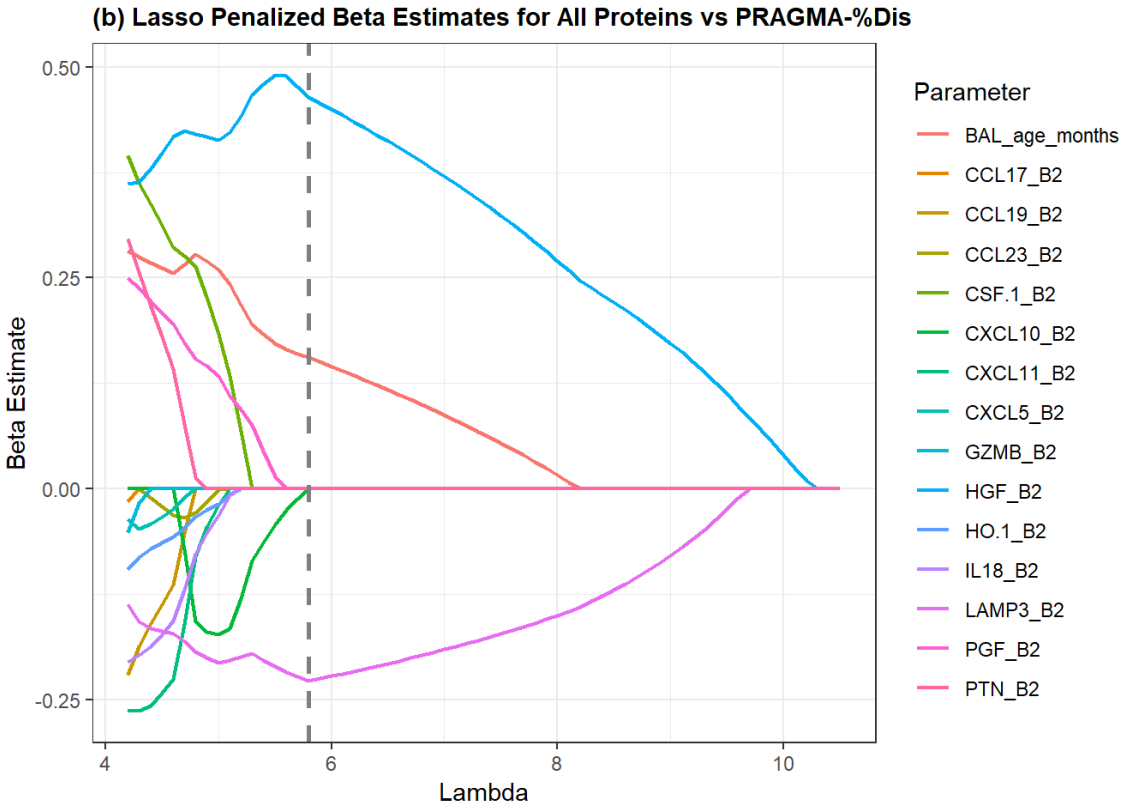


Figure 4a: Lasso selection criteria for PRAGMA-%Bx. The model fails to converge for lambda values less than 18.0, but the *lmmlasso* algorithm still prints results. The lowest BIC occurs at 14.9. While the model does not converge at 14.9 (4 parameters), the lowest acceptable lambda of 18.0 still has 4 parameters in the model (+ HGF, + ICOSLG, + BAL_age_months, -LAMP3).

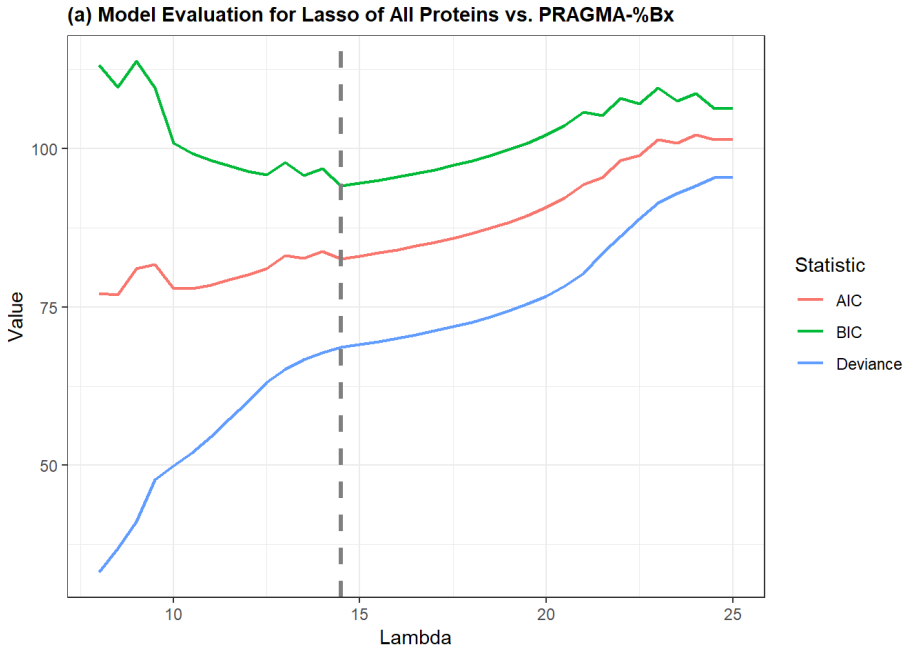
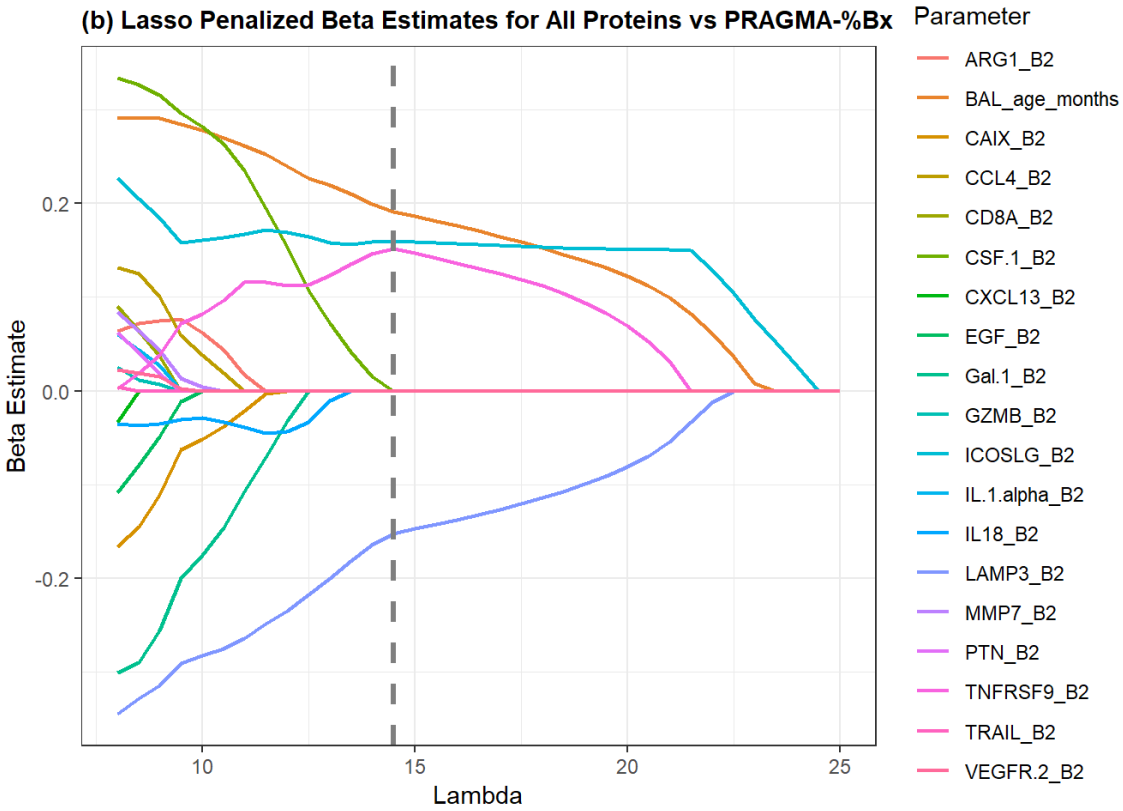


Figure 4b: Same as figure 3b, but for response variable of PRAGMA-%Bx.



Judging from figures 3a and 3b, the lowest BIC occurs at 5.8 and 14.9 for % Dis and % Bx, respectively. For % Dis, the lowest BIC corresponds to a final mixed model comprised of independent variables age, HGF, and LAMP3 (figure 4a). A lack of convergence warning occurred at lambda = 14.9 for %Bx (corresponding to a 4 parameter-model), but the issue is resolved by setting lambda to at least 18.0, which still retains 4-parameter model of age, ICOSLG, LAMP3, and HGF (figure 4b). All selected parameters were refitted into a linear mixed model with AR(1) covariance structure. Parameter estimates, 95% confidence intervals, associated p-values are reported in table 2 below. Note that LAMP3 has a negative parameter estimate for both % Dis and % Bx, while all other estimates are positive.

Table 2: Fitted multivariate linear mixed models for PRAGMA-%Dis and PRAGMA-%Bx as chosen by the lasso. For each independent variable, the parameter estimate is displayed first, followed by the 95% confidence interval and the p-value.

	PRAGMA-%Dis		
Parameter	**Estimate**	**95% CI**	**p-value**
(Intercept)	0.567	(-2.846, 3.980)	0.737
BAL_age_months	0.022	(-0.014, 0.058)	0.143
HGF_B2	0.546	(-0.015, 1.106)	0.053
LAMP3_B2	-0.302	(-0.899, 0.294)	0.205
Observations	38		
Unique Subjects	32		
BIC	165.874		

	PRAGMA-%Bx		
Parameter	**Estimate**	**95% CI**	**p-value**
(Intercept)	0.584	(-0.630, 1.797)	0.334
BAL_age_months	0.016	(-0.009, 0.040)	0.109
ICOSLG_B2	0.210	(-0.594, 1.014)	0.378
LAMP_3B2	-0.208	(-0.570, 0.153)	0.131
TNFRSF9_B2	0.180	(-0.253, 0.612)	0.216
Observations	38		
Unique Subjects	32		
BIC	111.581		

4. DISCUSSION

4.1 Interpretation of Results

The results of the cross-sectional analyses demonstrate significant differences in the association between protein concentrations and PRAGMA scores across the three time points of the study. At E2, nearly all of the correlations are insignificant. While PRAGMA-% Dis was negatively correlated and %Bx was positively correlated with the proteins, there is not enough evidence to ascertain the true directionality of the associations. However, it is reasonable to infer that the associations are weak due to the lack of observable inflammatory pathology at such a young age. As the children progress into the E4 time point (approximately 3 years old), most proteins become significantly positively correlated with %Dis. The change in magnitude of correlation coincides with the establishment of CF pathology as the children develop. At the final E6 time point (approx. 5 years old), only a handful of proteins are significant, indicating a possible stabilization in inflammatory response from three to five years of age. However, further research is needed to infer the causality of such a change in relations.

With regards to screening possible biomarkers, the E6 proteins seem most helpful as a starting point. False discovery rate p-value adjustment can further help in determining reliable predictors of PRAGMA scores while reducing Type I error. At E4, a relatively large number of significant proteins remain significant after FDR adjustment, while the FDR adjustment at E6 yields only CSF1, ARG1, and CCL4 as proteins that are significantly associated with PRAGMA-%Dis. The proteins have been of high interest in recent CF literature and could aid in corroborating current evidence that they associate with airway damage in young children. However, the proteins do not exhibit significant

correlation with PRAGMA-%Bx after FDR adjustment, so there is not enough evidence to determine if they also correlate with airway structural damage. The lack of significance in the correlations may also be caused by the low range of values PRAGMA-% Bx scores (min = 0.000, max = 3.76, median = 0.788; table 1) as opposed to the PRAGMA-%Dis (min = 0.871, max = 9.86, median = 3.08) across all ages, which is expected due to lack of severe airway damage in young children as opposed to teenage or adult CF populations. However, the results are valuable in determining earliest causes of inflammation and therapeutic targets to impede progression before onset of severe pathology.

The trends seen in the cross-sectional analysis also provide insight into building appropriate longitudinal models to assess marginal effects of each protein on PRAGMA scores while accounting for age and potential interaction between age and protein concentration. Due to changes in magnitudes of the protein vs. PRAGMA associations at E2 (all insignificant and weakly positive or negative), E4 (mostly significant and strongly positive), and E6 (select few significant and moderately positive), there is *a priori* reasoning to infer the need for an interaction term of age and protein concentration. The significant results of most of the protein contrast with model II especially at E6, which confirm such a hypothesis, indicating that the main effects of protein on PRAGMA score is significantly different at each age. Thus, the interaction term successfully encapsulates such differences in main effects. In addition, the longitudinal models have an advantage over cross-sectional correlations in that it has more power in differentiating proteins that truly have significant long-term airway effects in young CF children versus just a significant effect solely at E6. Significant proteins of interest included ARG1, CSF1, and

CCL4 for both PRAGMA-%Dis and %Bx, which corroborates the cross-sectional findings and lends further evidence that they may be associated with longitudinal long-term airways disease and damage in young children, not just a cross-sectional benchmark of five-year-olds.

A limitation of the cross-sectional association and the previous linear mixed models is that they only consider the marginal effects of each individual protein, rather than all of them simultaneously. Immunological responses are regulated by intricate molecular pathways that comprise of complex interactions within and among signaling proteins, and cells such as neutrophils, macrophages, epithelial cells and other key players. While it is not possible to capture all such relationships in one statistical model, the building of a multivariate protein model can help to determine which proteins demonstrate significant impact on PRAGMA scores. This can help eliminate proteins that appear to be significant in marginal analyses, but whose effects on PRAGMA scores may be mainly mediated through other proteins. In the lasso for both PRAGMA-%Dis and %Bx, age and LAMP3 were most robust against penalized regression. For PRAGMA-%Dis only, HGF was also a robust parameter, and for % Bx only, ICOSLG and TNFRSF9 were the additional robust parameters. Since the selected proteins differ for % Dis or %Bx, it may suggest possible heterogeneity in protein subsets that associate with total airway disease versus airway structural damage. However, the %Bx results also must be interpreted with caution since the range of values is low.

The p-values for the lasso selected parameters in the final linear mixed model are all insignificant. This may be due to the small sample sizes of the cohort. However, the selected proteins can still be potential topics of future research based on their robustness

against lasso penalization versus many other proteins. In addition, while they are not significant in the cross-sectional analysis (after FDR adjustment), they all are significant in the contrast tests under linear mixed models except for the effect of LAMP3 on PRAGMA-%Dis. LAMP3 is also unique in that it is the only protein with a significant negative effect on PRAGMA-%Bx in longitudinal model II, whereas a majority of significant proteins demonstrate positive directionality in many of the other figures and tables. Further research is needed to ascertain the negative contribution of LAMP3 to PRAGMA scores, as it was recently discovered that LAMP3 (also referred to as CD63) on neutrophil surfaces contribute to destruction of extracellular matrix in the lung.[26] There also may differences in clinical outcomes of cell-bound CD63 versus free CD63 that warrant future investigation.

It may be peculiar that CSF.1, CCL4, and ARG1, proteins that were notably significant in the both cross-sectional correlations and mixed model contrast tests, were not selected by the lasso. While they demonstrably contribute positively to airway disease when considered individually, this suggests that they may not be reliable predictors of PRAGMA scores when controlling for the marginal effects of all other proteins in a multivariate setting. As a disclaimer, none of the findings are indicative of a true causal relationship between protein and PRAGMA scores considered in either an individual or multivariate context. However, significant results can yield insights as to what strongly correlates with PRAGMA scores and to generate hypothesis as to potential targets for drug development or endpoints for clinicians to optimize CF pulmonology treatment for different children.

4.2 Limitations

As previously mentioned, the small number of repeated measurements can limit the power and stability of longitudinal mixed models and penalized regression. With more repeated measurements collected in the future, we expect to have better capacity to validate the current findings and reveal a clearer and more robust picture regarding the roles of proteins on airway disease progression. A larger sample size would also aid in predicting subject-level PRAGMA score trajectories as well as discovering possible heterogeneities in PRAGMA outcomes among the CF population.

Another limitation of the study is the possibility of batch effects among the three sets of PRAGMA scores generated for the subjects. No batch effect correction was employed in this thesis, but in a planned future analysis, all of subjects' CT scans will be rescored in one batch by one physician. This can help eliminate unwanted batch-wise variances.

4.3 Suggestions for Further Research

Due to the possibility of highly correlated proteins in the study, a future analysis can consider using penalized regression with elastic net penalty, which assigns an evenly-distributed shrinkage to highly correlated predictors rather than the lasso's behavior of arbitrarily favoring one over another.[27] The elastic net was not attempted here due to the unavailability of appropriate statistical software. Furthermore, it may be of interest to evaluate groups of proteins that work or interact together in known pathways.[27] This can be achieved through the grouped lasso algorithm, which allows one to penalize subsets of variables in groups rather than individually.[27] The results can be especially useful in

analysis and discovery of protein clusters that work in similar molecular pathways to affect PRAGMA scores.

5. REFERENCES

1. O'Sullivan, B. P., & Freedman, S. D. (2009). Cystic fibrosis. *The Lancet, 373*(9678), 1891-1904. doi:10.1016/s0140-6736(09)60327-5
2. CFTR gene - Genetics Home Reference - NIH. (n.d.). Retrieved January 21, 2019, from https://ghr.nlm.nih.gov/gene/CFTR#conditions
3. Cystic fibrosis - Genetics Home Reference - NIH. (n.d.). Retrieved January 21, 2019, from https://ghr.nlm.nih.gov/condition/cystic-fibrosis
4. Genes and human disease. (n.d.). Retrieved January 21, 2019, from https://www.who.int/genomics/public/geneticdiseases/en/index2.html#CF
5. Spoonhower, K. A., & Davis, P. B. (2016). Epidemiology of Cystic Fibrosis. *Clinics in Chest Medicine, 37*(1), 1-8. doi:10.1016/j.ccm.2015.10.002
6. Nichols, D. P., & Chmiel, J. F. (2015). Inflammation and its genesis in cystic fibrosis. *Pediatric Pulmonology, 50*(S40). doi:10.1002/ppul.23242
7. Cantin, A. M., Hartl, D., Konstan, M. W., & Chmiel, J. F. (2015). Inflammation in cystic fibrosis lung disease: Pathogenesis and therapy. *Journal of Cystic Fibrosis, 14*(4), 419-430. doi:10.1016/j.jcf.2015.03.003
8. Rosen, B. H., Evans, T. I., Moll, S. R., Gray, J. S., Liang, B., Sun, X., . . . Engelhardt, J. F. (2018). Infection Is Not Required for Mucoinflammatory Lung Disease in CFTR-Knockout Ferrets. *American Journal of Respiratory and Critical Care Medicine, 197*(10), 1308-1318. doi:10.1164/rccm.201708-1616oc
9. Sly, P. D., Gangell, C. L., Chen, L., Ware, R. S., Ranganathan, S., Mott, L. S., . . . Stick, S. M. (2013). Risk Factors for Bronchiectasis in Children with Cystic

Fibrosis. *New England Journal of Medicine, 368*(21), 1963-1970. doi:10.1056/nejmoa1301725

10. Margaroli, C., & Tirouvanziam, R. (2016). Neutrophil plasticity enables the development of pathological microenvironments: Implications for cystic fibrosis airway disease. *Molecular and Cellular Pediatrics, 3*(1). doi:10.1186/s40348-016-0066-2
11. Chandler, J. D., Margaroli, C., Horati, H., Kilgore, M., Veltman, M., Liu, H. K., . . . Janssens, H. M. (2018). Myeloperoxidase oxidation of methionine associates with early cystic fibrosis lung disease. *European Respiratory Journal*, *52*(4), 1801118. https://doi.org/10.1183/13993003.01118-2018
12. Margaroli, C., Garratt, L. W., Horati, H., Dittrich, A. S., Rosenow, T., Montgomery, S. T., . . . Tirouvanziam, R. (2019). Elastase Exocytosis by Airway Neutrophils Is Associated with Early Lung Damage in Children with Cystic Fibrosis. *Am J Respir Crit Care Med, 199*(7), 873-881. doi:10.1164/rccm.201803-0442OC
13. Gharib, S. A., Vaisar, T., Aitken, M. L., Park, D. R., Heinecke, J. W., & Fu, X. (2009). Mapping the Lung Proteome in Cystic Fibrosis. *Journal of Proteome Research, 8*(6), 3020-3028. doi: 10.1021/pr900093j
14. ERS Task Force On Bronchoalveolar Lavage In Children, & Members Of The Task Force: J. De Blic And F. Midulla (Co-Chairmen), A. Barbato, A. Clement, I. Dab, E. Eber, C. Green, J. Grigg, S. Kotecha, G. Kurland, P. Pohunek, F. Ratjen And G. Rossi. (2000). Bronchoalveolar lavage in children. *European Respiratory Journal, 15*(1), 217. doi:10.1034/j.1399-3003.2000.15a40.x

15. Giddings, O., & Esther, C. R. (2017). Mapping targetable inflammation and outcomes with cystic fibrosis biomarkers. *Pediatric Pulmonology, 52*(S48). doi:10.1002/ppul.23768

16. Ramsey, K. A., Schultz, A., & Stick, S. M. (2015). Biomarkers in Paediatric Cystic Fibrosis Lung Disease. *Paediatric Respiratory Reviews, 16*(4), 213-218. doi:10.1016/j.prrv.2015.05.004

17. Rosenow, T., Oudraad, M. C., Murray, C. P., Turkovic, L., Kuo, W., Bruijne, M. D., . . . Stick, S. M. (2015). PRAGMA-CF. A Quantitative Structural Lung Disease Computed Tomography Outcome in Young Children with Cystic Fibrosis. *American Journal of Respiratory and Critical Care Medicine, 191*(10), 1158-1165. doi:10.1164/rccm.201501-0061oc

18. DeBoer, E. M., Kroehl, M. E., Wagner, B. D., Accurso, F. J., Harris, J. K., Lynch, D. A., . . . Deterding, R. R. (2017). Proteomic profiling identifies novel circulating markers associated with bronchiectasis in cystic fibrosis. *PROTEOMICS - Clinical Applications, 11*(9-10), 1600147. doi:10.1002/prca.201600147

19. Olink Immuno-Oncology panel. (n.d.). Retrieved January 28, 2019, from https://www.olink.com/products/immuno-oncology/#

20. Assarsson, E., Lundberg, M., Holmquist, G., Björkesten, J., Thorsen, S. B., Ekman, D., . . . Fredriksson, S. (2014). Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLoS ONE, 9*(4). doi:10.1371/journal.pone.0095192

21. Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

22. West, B. T., Welch, K. B., & Gałecki, A. T. (2015). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: Chapman & Hall/CRC.

23. Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis*. Hoboken: Wiley.

24. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.

25. Schelldorfer, J., Bühlmann, P., & Geer, S. V. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ1-Penalization. *Scandinavian Journal of Statistics, 38*(2), 197-214. doi:10.1111/j.1467-9469.2011.00740.x

26. Genschmer, K. R., Russell, D. W., Lal, C., Szul, T., Bratcher, P. E., Noerager, B. D., . . . Blalock, J. E. (2019). Activated PMN Exosomes: Pathogenic Entities Causing Matrix Destruction and Disease in the Lung. *Cell, 176*(1-2). doi:10.1016/j.cell.2018.12.002

27. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY, USA: Springer.

6. APPENDIX A: Olink® Immuno-Oncology Panel Protein Names and Abbreviations

Abbreviation	Full Name	LOD value	% below LOD	< 20 % LOD
ADA	Adenosine Deaminase	-1.980	0.014	Yes
ADGRG1	Adhesion G-protein coupled receptor G1	-0.045	0.919	No
ANG.1	Angiopoietin-1	0.198	0.203	No
ANGPT2	Angiopoietin-2	-1.159	0.500	No
ARG1	Arginase-1	0.715	0.095	Yes
CAIX	Carbonic anhydrase IX	-1.128	0.108	Yes
CASP.8	Caspase-8	-0.246	0.014	Yes
CCL17	C-C motif chemokine 17	-1.040	0.027	Yes
CCL19	C-C motif chemokine 19	-0.589	0.041	Yes
CCL20	C-C motif chemokine 20	-0.123	0.014	Yes
CCL23	C-C motif chemokine 23	-0.004	0.014	Yes
CCL3	C-C motif chemokine 17	-0.785	0.014	Yes
CCL4	C-C motif chemokine 3	0.224	0.014	Yes
CD244	Natural killer cell receptor 2B4	-0.289	0.068	Yes
CD27	CD27 antigen	0.065	0.014	Yes
CD28	T-cell-specific surface glycoprotein CD28	0.174	0.932	No
CD4	T-cell surface glycoprotein CD4	-2.362	0.027	Yes
CD40	CD40 ligand	-0.185	0.014	Yes
CD40.L	CD40L receptor	-0.097	0.662	No
CD5	T-cell surface glycoprotein CD5	-1.099	0.014	Yes
CD70	CD70 antigen	-0.256	0.216	No
CD83	CD83 antigen	-0.654	0.014	Yes
CD8A	T-cell surface glycoprotein CD8 alpha chain	-0.058	0.122	Yes
CRTAM	Cytotoxic and regulatory T-cell molecule	-0.614	0.703	No
CSF.1	Macrophage colony-stimulating factor 1	-1.462	0.014	Yes
CX3CL1	Fractalkine	-0.484	0.027	Yes
CXCL1	C-X-C motif chemokine 1	0.423	0.014	Yes
CXCL10	C-X-C motif chemokine 10	-1.472	0.027	Yes
CXCL11	C-X-C motif chemokine 11	-0.267	0.081	Yes
CXCL12	Stromal cell-derived factor 1	0.073	0.986	No
CXCL13	C-X-C motif chemokine 13	-0.160	0.014	Yes

CXCL5	C-X-C motif chemokine 5	0.070	0.027	Yes
CXCL9	C-X-C motif chemokine 9	-0.339	0.014	Yes
DCN	Decorin	-0.392	0.378	No
EGF	Pro-epidermal growth factor	0.008	0.027	Yes
FASLG	Fas antigen ligand	-0.415	0.014	Yes
FGF2	Fibroblast growth factor 2	-0.953	0.257	No
Gal.1	Galectin-1	-0.225	0.014	Yes
Gal.9	Galectin-9	-0.206	0.014	Yes
GZMA	Granzyme A	-0.988	0.014	Yes
GZMB	Granzyme B	-0.256	0.014	Yes
GZMH	Granzyme H	0.156	0.027	Yes
HGF	Hepatocyte growth factor	-0.698	0.014	Yes
HO.1	Heme oxygenase 1	0.164	0.014	Yes
ICOSLG	ICOS ligand	-0.336	0.014	Yes
IFN.beta	Interferon beta	-0.076	0.932	No
IFN.gamma	Interferon gamma	-0.101	0.932	No
IL.1.alpha	Interleukin-1 alpha	-0.110	0.054	Yes
IL.21	Interleukin-21	0.057	0.946	No
IL.35	Interleukin-35	-0.739	0.905	No
IL10	Interleukin-10	-0.158	0.757	No
IL12	Interleukin-12	-0.939	0.014	Yes
IL12RB1	Interleukin-12 receptor subunit beta-1	-0.196	0.716	No
IL13	Interleukin-13	-0.846	0.784	No
IL18	Interleukin-18	-0.005	0.014	Yes
IL2	Interleukin-2	-0.002	0.932	No
IL33	Interleukin-33	0.051	0.568	No
IL4	Interleukin-4	-1.018	0.595	No
IL5	Interleukin-5	0.872	0.986	No
IL6	Interleukin-6	0.057	0.014	Yes
IL7	Interleukin-7	-0.301	0.027	Yes
IL8	Interleukin-8	0.469	0.014	Yes
KLRD1	Natural killer cells antigen CD94	-0.400	0.432	No
LAMP3	Lysosome-associated membrane glycoprotein 3	-0.148	0.014	Yes
LAP.TGF.beta.1	Latency-associated peptide transforming growth factor beta-1	-1.623	0.122	Yes
MCP.1	Monocyte chemotactic protein 1	-0.607	0.014	Yes
MCP.2	Monocyte chemotactic protein 2	-0.169	0.041	Yes
MCP.3	Monocyte chemotactic protein 3	-0.423	0.068	Yes
MCP.4	Monocyte chemotactic protein 4	-0.735	0.027	Yes

MIC.A.B	MHC class I polypeptide-related sequence A/B	-0.828	0.068	Yes
MMP12	Matrix metalloproteinase-12	0.050	0.014	Yes
MMP7	Matrix metalloproteinase-7	-0.041	0.014	Yes
NCR1	Natural cytotoxicity triggering receptor 1	0.003	0.892	No
NOS3	Nitric oxide synthase, endothelial	-0.570	0.743	No
PD.L1	Programmed cell death 1 ligand 1	-0.003	0.041	Yes
PD.L2	Programmed cell death 1 ligand 2	-0.594	0.230	No
PDCD1	Programmed cell death protein 1	0.124	0.716	No
PDGF.subunit.B	Platelet-derived growth factor subunit B	-1.008	0.068	Yes
PGF	Placenta growth factor	0.224	0.014	Yes
PTN	Pleiotrophin	-0.789	0.027	Yes
TIE2	Angiopoietin-1 receptor	-0.241	0.284	No
TNF	Tumor necrosis factor	-0.226	0.500	No
TNFRSF12A	Tumor necrosis factor receptor superfamily member 12A	-0.594	0.027	Yes
TNFRSF21	Tumor necrosis factor receptor superfamily member 21	-0.161	0.014	Yes
TNFRSF4	Tumor necrosis factor receptor superfamily member 4	-0.165	0.338	No
TNFRSF9	Tumor necrosis factor receptor superfamily member 9	-1.299	0.014	Yes
TNFSF14	Tumor necrosis factor ligand superfamily member 14	-0.434	0.027	Yes
TRAIL	TNF-related apoptosis-inducing ligand	-0.793	0.014	Yes
TWEAK	Tumor necrosis factor (Ligand) superfamily, member 12	-0.750	0.014	Yes
VEGFA	Vascular endothelial growth factor A	-0.082	0.014	Yes
VEGFC	Vascular endothelial growth factor C	-1.898	0.216	No
VEGFR.2	Vascular endothelial growth factor receptor 2	-1.566	0.068	Yes

7. APPENDIX B: Additional Tables

Table 4a: Cross-Sectional Correlations of PRAGMA-%Dis vs. All Proteins

	Study Event								
	E2 (n=9)			E4 (n=14)			E6 (n=13)		
Protein	Spearman Rho	P-Value	FDR P-Value	Spearman Rho	P-Value	FDR P-Value	Spearman Rho	P-Value	FDR P-Value
ADA_B2	-0.567	0.121	0.587	0.776	0.002	0.005	0.703	0.010	0.089
ARG1_B2	-0.683	0.050	0.587	0.868	0.000	0.001	0.808	0.001	0.043
CAIX_B2	-0.800	0.014	0.587	0.827	0.000	0.002	0.380	0.201	0.369
CASP.8_B2	-0.367	0.336	0.596	0.767	0.002	0.005	0.379	0.202	0.369
CCL17_B2	-0.183	0.644	0.753	0.068	0.820	0.820	0.176	0.566	0.650
CCL19_B2	-0.333	0.385	0.646	0.539	0.050	0.062	0.264	0.384	0.495
CCL20_B2	-0.233	0.552	0.728	0.758	0.003	0.006	0.528	0.067	0.182
CCL23_B2	-0.400	0.291	0.587	0.196	0.502	0.537	0.082	0.793	0.806
CCL3_B2	-0.500	0.178	0.587	0.815	0.001	0.003	0.659	0.017	0.089
CCL4_B2	-0.200	0.613	0.746	0.754	0.003	0.006	0.791	0.002	0.043
CD244_B2	-0.483	0.194	0.587	0.911	0.000	0.000	0.319	0.289	0.426
CD27_B2	-0.417	0.270	0.587	0.710	0.006	0.011	0.319	0.289	0.426
CD4_B2	-0.483	0.194	0.587	0.675	0.010	0.017	0.148	0.630	0.697
CD40_B2	-0.250	0.521	0.702	0.837	0.000	0.002	0.550	0.055	0.180
CD5_B2	-0.467	0.213	0.587	0.833	0.000	0.002	0.236	0.437	0.542
CD83_B2	-0.383	0.313	0.587	0.556	0.042	0.054	0.506	0.081	0.194
CD8A_B2	-0.417	0.270	0.587	0.139	0.638	0.648	-0.096	0.754	0.780
CSF.1_B2	-0.033	0.948	0.980	0.714	0.006	0.010	0.868	0.000	0.008
CX3CL1_B2	-0.467	0.213	0.587	0.741	0.004	0.007	0.132	0.669	0.703
CXCL1_B2	-0.217	0.581	0.746	0.604	0.025	0.035	0.528	0.067	0.182
CXCL10_B2	-0.317	0.410	0.669	0.481	0.084	0.100	0.209	0.494	0.588
CXCL11_B2	-0.417	0.270	0.587	0.393	0.165	0.183	0.050	0.878	0.878

CXCL13_B2	-0.517	0.162	0.587	0.789	0.001	0.004	0.143	0.643	0.699
CXCL5_B2	-0.450	0.230	0.587	0.626	0.019	0.031	0.352	0.239	0.401
CXCL9_B2	-0.467	0.213	0.587	0.736	0.004	0.008	0.396	0.182	0.369
EGF_B2	-0.250	0.521	0.702	0.763	0.002	0.006	0.319	0.289	0.426
FASLG_B2	-0.433	0.250	0.587	0.574	0.035	0.047	0.308	0.306	0.442
Gal.1_B2	-0.100	0.810	0.881	0.824	0.001	0.002	0.423	0.152	0.336
Gal.9_B2	0.017	0.982	0.982	0.899	0.000	0.000	0.390	0.189	0.369
GZMA_B2	-0.617	0.086	0.587	0.393	0.165	0.183	0.302	0.315	0.444
GZMB_B2	-0.533	0.148	0.587	0.604	0.025	0.035	0.544	0.058	0.180
GZMH_B2	-0.400	0.291	0.587	0.407	0.151	0.173	0.214	0.482	0.586
HGF_B2	-0.667	0.059	0.587	0.824	0.001	0.002	0.670	0.015	0.089
HO.1_B2	-0.133	0.744	0.823	0.754	0.003	0.006	-0.385	0.196	0.369
ICOSLG_B2	-0.483	0.194	0.587	0.640	0.016	0.027	0.539	0.061	0.180
IL.1.alpha_B2	-0.383	0.313	0.587	0.744	0.002	0.006	0.511	0.078	0.192
IL12_B2	-0.067	0.880	0.925	0.389	0.170	0.185	0.280	0.353	0.476
IL18_B2	-0.267	0.493	0.695	0.596	0.028	0.038	0.352	0.239	0.401
IL6_B2	-0.417	0.270	0.587	0.793	0.001	0.004	0.423	0.152	0.336
IL7_B2	-0.200	0.613	0.746	0.798	0.001	0.004	0.170	0.579	0.652
IL8_B2	-0.683	0.050	0.587	0.802	0.001	0.004	0.582	0.040	0.155
LAMP3_B2	-0.367	0.336	0.596	0.182	0.532	0.559	-0.264	0.384	0.495
LAP.TGF.beta.1_B2	-0.133	0.744	0.823	0.798	0.001	0.004	0.649	0.016	0.089
MCP.1_B2	-0.283	0.463	0.695	0.609	0.024	0.035	0.566	0.047	0.163
MCP.2_B2	-0.467	0.213	0.587	0.556	0.042	0.054	0.132	0.669	0.703
MCP.3_B2	-0.200	0.613	0.746	0.541	0.046	0.058	0.352	0.239	0.401
MCP.4_B2	-0.017	0.982	0.982	0.415	0.141	0.165	0.187	0.541	0.633
MIC.A.B_B2	0.067	0.880	0.925	0.169	0.563	0.582	-0.324	0.280	0.426
MMP12_B2	-0.667	0.059	0.587	0.728	0.005	0.009	0.648	0.020	0.093
MMP7_B2	-0.267	0.493	0.695	0.767	0.002	0.005	0.725	0.007	0.089

PD.L1_B2	-0.383	0.313	0.587	0.609	0.024	0.035	0.324	0.280	0.426
PDGF.subunit.B_B2	-0.533	0.148	0.587	0.793	0.001	0.004	0.714	0.008	0.089
PGF_B2	-0.483	0.194	0.587	0.912	0.000	0.000	0.676	0.014	0.089
PTN_B2	-0.150	0.708	0.813	0.723	0.005	0.009	0.412	0.163	0.349
TNFRSF12A_B2	-0.267	0.493	0.695	0.604	0.025	0.035	0.511	0.078	0.192
TNFRSF21_B2	-0.333	0.385	0.646	0.675	0.010	0.017	0.280	0.353	0.476
TNFRSF9_B2	-0.617	0.086	0.587	0.837	0.000	0.002	0.665	0.016	0.089
TNFSF14_B2	-0.600	0.097	0.587	0.829	0.000	0.002	0.588	0.038	0.155
TRAIL_B2	-0.300	0.437	0.677	0.886	0.000	0.000	0.692	0.011	0.089
TWEAK_B2	-0.417	0.270	0.587	0.771	0.002	0.005	0.593	0.036	0.155
VEGFA_B2	-0.300	0.437	0.677	0.521	0.059	0.072	0.577	0.043	0.155
VEGFR.2_B2	-0.183	0.644	0.753	0.895	0.000	0.000	0.242	0.426	0.538

Table 4b: Cross-Sectional Correlations of PRAGMA-%Bx vs. All Proteins

	Study Event								
	E2 (n=9)			**E4 (n=14)**			**E6 (n=13)**		
Protein	**Spearman Rho**	**P-Value**	**FDR P-Value**	**Spearman Rho**	**P-Value**	**FDR P-Value**	**Spearman Rho**	**P-Value**	**FDR P-Value**
ADA_B2	0.270	0.483	0.863	0.577	0.031	0.117	0.582	0.040	0.324
ARG1_B2	0.174	0.654	0.863	0.648	0.012	0.084	0.659	0.017	0.227
CAIX_B2	-0.017	0.965	0.980	0.435	0.120	0.213	0.179	0.559	0.737
CASP.8_B2	-0.035	0.929	0.976	0.563	0.036	0.117	0.242	0.426	0.729
CCL17_B2	0.348	0.359	0.863	-0.165	0.573	0.629	0.000	1.000	1.000
CCL19_B2	0.531	0.141	0.863	0.130	0.658	0.692	0.154	0.617	0.797
CCL20_B2	0.192	0.622	0.863	0.568	0.034	0.117	0.407	0.170	0.454
CCL23_B2	0.174	0.654	0.863	-0.015	0.958	0.964	-0.126	0.683	0.804
CCL3_B2	-0.148	0.704	0.873	0.508	0.064	0.158	0.522	0.071	0.324
CCL4_B2	-0.122	0.755	0.906	0.594	0.025	0.109	0.692	0.011	0.227
CD244_B2	0.235	0.543	0.863	0.594	0.025	0.109	0.181	0.554	0.737

CD27_B2	0.400	0.286	0.863	0.530	0.051	0.144	0.247	0.415	0.729
CD4_B2	0.017	0.965	0.980	0.372	0.191	0.285	0.022	0.949	0.998
CD40_B2	0.348	0.359	0.863	0.449	0.107	0.202	0.456	0.120	0.354
CD5_B2	0.296	0.439	0.863	0.590	0.027	0.109	0.099	0.751	0.846
CD83_B2	0.279	0.468	0.863	0.354	0.214	0.309	0.286	0.344	0.710
CD8A_B2	-0.183	0.638	0.863	-0.224	0.441	0.536	-0.124	0.687	0.804
CSF.1_B2	0.540	0.134	0.863	0.469	0.091	0.182	0.659	0.017	0.227
CX3CL1_B2	0.340	0.372	0.863	0.308	0.284	0.367	0.071	0.821	0.909
CXCL1_B2	0.000	1.000	1.000	0.370	0.193	0.285	0.401	0.176	0.454
CXCL10_B2	0.279	0.468	0.863	0.167	0.568	0.629	0.143	0.643	0.797
CXCL11_B2	0.148	0.704	0.873	0.189	0.517	0.617	-0.011	0.978	1.000
CXCL13_B2	0.322	0.398	0.863	0.524	0.055	0.147	-0.033	0.921	0.984
CXCL5_B2	0.174	0.654	0.863	0.350	0.220	0.310	0.099	0.751	0.846
CXCL9_B2	0.340	0.372	0.863	0.493	0.073	0.175	0.341	0.255	0.586
EGF_B2	0.174	0.654	0.863	0.332	0.246	0.331	0.231	0.448	0.729
FASLG_B2	0.383	0.309	0.863	0.414	0.142	0.244	0.247	0.415	0.729
Gal.1_B2	-0.104	0.789	0.906	0.541	0.046	0.135	0.126	0.683	0.804
Gal.9_B2	0.061	0.876	0.937	0.566	0.035	0.117	0.264	0.384	0.729
GZMA_B2	0.061	0.876	0.937	0.165	0.573	0.629	0.236	0.437	0.729
GZMB_B2	0.261	0.497	0.863	0.464	0.095	0.183	0.495	0.089	0.324
GZMH_B2	0.113	0.772	0.906	0.136	0.642	0.686	0.231	0.448	0.729
HGF_B2	0.218	0.574	0.863	0.687	0.007	0.076	0.533	0.064	0.324
HO.1_B2	-0.087	0.824	0.912	0.409	0.146	0.245	-0.456	0.120	0.354
ICOSLG_B2	0.374	0.321	0.863	0.713	0.004	0.076	0.517	0.074	0.324
IL.1.alpha_B2	0.279	0.468	0.863	0.558	0.038	0.118	0.495	0.089	0.324
IL12_B2	0.644	0.061	0.863	0.163	0.578	0.629	0.187	0.541	0.737
IL18_B2	0.148	0.704	0.873	0.471	0.089	0.182	0.220	0.470	0.729
IL6_B2	0.218	0.574	0.863	0.638	0.014	0.084	0.319	0.289	0.639

IL7_B2	0.322	0.398	0.863	0.308	0.284	0.367	0.187	0.541	0.737
IL8_B2	0.270	0.483	0.863	0.682	0.007	0.076	0.467	0.110	0.354
LAMP3_B2	0.644	0.061	0.863	-0.013	0.964	0.964	-0.203	0.505	0.737
LAP.TGF.beta.1_B2	0.261	0.497	0.863	0.709	0.005	0.076	0.520	0.069	0.324
MCP.1_B2	0.218	0.574	0.863	0.471	0.089	0.182	0.385	0.196	0.485
MCP.2_B2	0.340	0.372	0.863	0.381	0.179	0.278	0.044	0.892	0.970
MCP.3_B2	0.313	0.412	0.863	0.398	0.159	0.260	0.192	0.529	0.737
MCP.4_B2	0.566	0.112	0.863	0.169	0.563	0.629	0.006	0.993	1.000
MIC.A.B_B2	0.618	0.076	0.863	-0.031	0.917	0.947	-0.231	0.448	0.729
MMP12_B2	0.104	0.789	0.906	0.638	0.014	0.084	0.500	0.085	0.324
MMP7_B2	0.461	0.211	0.863	0.475	0.086	0.182	0.753	0.004	0.227
PD.L1_B2	0.313	0.412	0.863	0.277	0.337	0.418	0.198	0.517	0.737
PDGF.subunit.B_B2	0.200	0.606	0.863	0.858	0.000	0.005	0.544	0.058	0.324
PGF_B2	0.296	0.439	0.863	0.616	0.019	0.098	0.511	0.078	0.324
PTN_B2	0.096	0.806	0.909	0.343	0.230	0.316	0.374	0.209	0.499
TNFRSF12A_B2	0.331	0.385	0.863	0.389	0.169	0.268	0.291	0.334	0.710
TNFRSF21_B2	0.279	0.468	0.863	0.482	0.081	0.182	0.148	0.630	0.797
TNFRSF9_B2	0.244	0.527	0.863	0.671	0.009	0.076	0.654	0.018	0.227
TNFSF14_B2	0.261	0.497	0.863	0.634	0.015	0.084	0.495	0.089	0.324
TRAIL_B2	0.348	0.359	0.863	0.519	0.057	0.147	0.610	0.030	0.313
TWEAK_B2	0.522	0.149	0.863	0.438	0.117	0.213	0.456	0.120	0.354
VEGFA_B2	0.575	0.106	0.863	0.277	0.337	0.418	0.445	0.130	0.366
VEGFR.2_B2	0.392	0.297	0.863	0.673	0.008	0.076	0.223	0.464	0.729

Table 5: Model I Results for All Proteins vs. PRAGMA-%Dis and %Bx

	%Dis (n=38, repeats=6)		%Bx (n=38, repeats=6)	
Protein	**Beta-1**	**P-Value**	**Beta-1**	**P-Value**
ADA_B2	0.417	0.043	0.205	0.031
ARG1_B2	0.333	0.067	0.178	0.035
CAIX_B2	0.318	0.339	0.018	0.903
CASP.8_B2	0.152	0.468	0.09	0.354
CCL17_B2	-0.198	0.437	-0.052	0.647
CCL19_B2	-0.005	0.98	0.017	0.841
CCL20_B2	0.173	0.208	0.082	0.179
CCL23_B2	-0.317	0.223	-0.077	0.517
CCL3_B2	0.247	0.2	0.107	0.216
CCL4_B2	0.462	0.029	0.217	0.025
CD244_B2	0.24	0.417	0.068	0.611
CD27_B2	0.462	0.162	0.207	0.172
CD4_B2	0.012	0.969	-0.028	0.835
CD40_B2	0.368	0.171	0.163	0.174
CD5_B2	0.2	0.288	0.068	0.423
CD83_B2	0.271	0.472	0.089	0.596
CD8A_B2	-0.132	0.367	-0.075	0.278
CSF.1_B2	0.614	0.029	0.228	0.062
CX3CL1_B2	0.164	0.519	0.058	0.615
CXCL1_B2	0.049	0.841	0.012	0.911
CXCL10_B2	-0.095	0.443	0.009	0.865
CXCL11_B2	-0.051	0.744	0.023	0.74
CXCL13_B2	0.094	0.56	0.05	0.495
CXCL5_B2	-0.026	0.86	0.027	0.692
CXCL9_B2	0.239	0.192	0.13	0.124
EGF_B2	0.569	0.047	0.14	0.224

FASLG_B2	0.114	0.794	0.136	0.506
Gal.1_B2	0.315	0.273	0.048	0.701
Gal.9_B2	0.699	0.112	0.188	0.312
GZMA_B2	0.162	0.373	0.066	0.431
GZMB_B2	0.145	0.338	0.128	0.09
GZMH_B2	0.101	0.554	0.05	0.522
HGF_B2	0.603	0.026	0.263	0.028
HO.1_B2	-0.07	0.582	-0.036	0.545
ICOSLG_B2	1.096	0.022	0.503	0.023
IL.1.alpha_B2	0.291	0.214	0.223	0.055
IL12_B2	0.218	0.559	0.097	0.569
IL18_B2	0.059	0.827	0.031	0.802
IL6_B2	0.282	0.161	0.118	0.183
IL7_B2	0.343	0.339	0.076	0.629
IL8_B2	0.348	0.042	0.167	0.033
LAMP3_B2	-0.455	0.082	-0.251	0.061
LAP.TGF.beta.1_B2	0.716	0.245	0.366	0.197
MCP.1_B2	0.391	0.097	0.148	0.147
MCP.2_B2	-0.017	0.925	0.058	0.477
MCP.3_B2	0.205	0.378	0.091	0.387
MCP.4_B2	0.012	0.968	0.025	0.86
MIC.A.B_B2	-0.201	0.471	-0.062	0.62
MMP12_B2	0.311	0.032	0.129	0.041
MMP7_B2	0.304	0.12	0.163	0.077
PD.L1_B2	0.165	0.665	0.081	0.637
PDGF.subunit.B_B2	0.533	0.177	0.34	0.076
PGF_B2	0.873	0.027	0.372	0.029
PTN_B2	0.32	0.06	0.115	0.122
TNFRSF12A_B2	0.754	0.249	0.174	0.537
TNFRSF21_B2	0.047	0.915	0.055	0.788

TNFRSF9_B2	0.537	0.036	0.28	0.018
TNFSF14_B2	0.366	0.057	0.192	0.031
TRAIL_B2	0.501	0.052	0.202	0.068
TWEAK_B2	0.42	0.221	0.128	0.389
VEGFA_B2	0.434	0.35	0.031	0.877
VEGFR.2_B2	0.307	0.527	0.275	0.229

Table 6: Model II Parameter Estimates for All Proteins vs. PRAGMA-%Dis and %Bx

	%Dis (n= 38, repeats=6)				%Bx (n=38, repeats=6)			
	Beta 1		Beta 3		Beta 1		Beta 3	
Protein	**Estimate**	**P-Value**	**Estimate**	**P-Value**	**Estimate**	**P-Value**	**Estimate**	**P-Value**
ADA_B2	-0.383	0.396	0.019	0.109	-0.03	0.883	0.005	0.275
ARG1_B2	-0.428	0.24	0.02	0.055	-0.068	0.658	0.006	0.147
CAIX_B2	-0.801	0.307	0.029	0.156	-0.141	0.699	0.005	0.588
CASP.8_B2	-0.421	0.405	0.017	0.238	-0.154	0.508	0.007	0.286
CCL17_B2	-0.694	0.155	0.017	0.215	-0.048	0.819	0	0.983
CCL19_B2	-0.229	0.537	0.006	0.483	-0.036	0.838	0.001	0.728
CCL20_B2	-0.277	0.335	0.012	0.129	-0.104	0.44	0.005	0.181
CCL23_B2	-0.577	0.26	0.01	0.412	0.006	0.981	-0.002	0.69
CCL3_B2	-0.456	0.292	0.02	0.117	-0.179	0.375	0.008	0.164
CCL4_B2	-0.428	0.324	0.022	0.075	-0.166	0.394	0.009	0.093
CD244_B2	-0.342	0.59	0.016	0.332	-0.14	0.643	0.006	0.442
CD27_B2	-0.408	0.53	0.023	0.184	-0.051	0.862	0.007	0.362
CD4_B2	-0.283	0.666	0.008	0.618	-0.083	0.787	0.002	0.844
CD40_B2	-0.316	0.562	0.018	0.213	-0.088	0.734	0.007	0.322
CD5_B2	-0.108	0.804	0.008	0.457	-0.056	0.788	0.003	0.529
CD83_B2	-1.04	0.193	0.044	0.091	-0.166	0.629	0.008	0.419
CD8A_B2	-0.052	0.871	-0.002	0.779	-0.123	0.425	0.001	0.717
CSF.1_B2	-0.549	0.249	0.03	0.043	-0.148	0.498	0.01	0.118
CX3CL1_B2	-0.278	0.629	0.011	0.423	-0.012	0.963	0.002	0.77
CXCL1_B2	-0.887	0.149	0.03	0.102	-0.351	0.239	0.011	0.188
CXCL10_B2	-0.224	0.353	0.004	0.488	-0.066	0.568	0.002	0.463
CXCL11_B2	-0.109	0.722	0.002	0.826	0.006	0.968	0.001	0.886
CXCL13_B2	-0.195	0.568	0.007	0.36	-0.031	0.843	0.002	0.573
CXCL5_B2	-0.351	0.264	0.011	0.228	-0.091	0.533	0.004	0.375

CXCL9_B2	-0.081	0.837	0.007	0.399	-0.006	0.975	0.003	0.442
EGF_B2	0.474	0.407	0.002	0.874	0.125	0.687	0	0.952
FASLG_B2	-0.381	0.692	0.014	0.566	-0.17	0.698	0.009	0.453
Gal.1_B2	-0.507	0.483	0.019	0.25	-0.106	0.749	0.004	0.62
Gal.9_B2	-0.41	0.71	0.026	0.315	-0.128	0.806	0.007	0.529
GZMA_B2	-0.128	0.807	0.007	0.563	-0.1	0.682	0.004	0.478
GZMB_B2	-0.217	0.511	0.01	0.243	-0.06	0.667	0.005	0.185
GZMH_B2	-0.129	0.745	0.006	0.533	-0.073	0.691	0.003	0.473
HGF_B2	-0.345	0.466	0.022	0.09	-0.017	0.94	0.007	0.248
HO.1_B2	-0.143	0.634	0.002	0.776	0.022	0.881	-0.001	0.688
ICOSLG_B2	-0.297	0.772	0.031	0.196	0.15	0.742	0.007	0.425
IL.1.alpha_B2	-0.276	0.597	0.014	0.272	-0.083	0.713	0.008	0.202
IL12_B2	-0.063	0.935	0.007	0.683	-0.081	0.822	0.005	0.588
IL18_B2	-0.557	0.292	0.017	0.201	-0.06	0.804	0.003	0.656
IL6_B2	-0.34	0.383	0.018	0.124	-0.078	0.661	0.005	0.268
IL7_B2	-0.044	0.96	0.009	0.639	-0.137	0.752	0.005	0.607
IL8_B2	-0.334	0.384	0.017	0.112	-0.073	0.676	0.006	0.212
LAMP3_B2	-0.426	0.449	-0.001	0.96	0.147	0.585	-0.01	0.169
LAP.TGF.beta.1_B2	-1.119	0.284	0.067	0.082	-0.171	0.712	0.021	0.197
MCP.1_B2	-0.355	0.425	0.019	0.119	-0.061	0.767	0.005	0.307
MCP.2_B2	-0.174	0.612	0.005	0.586	-0.002	0.991	0.002	0.657
MCP.3_B2	-0.215	0.671	0.011	0.372	-0.009	0.969	0.002	0.635
MCP.4_B2	-0.209	0.771	0.006	0.722	-0.091	0.787	0.003	0.702
MIC.A.B_B2	0.048	0.932	-0.007	0.629	0.121	0.655	-0.005	0.468
MMP12_B2	-0.183	0.481	0.012	0.102	-0.001	0.991	0.003	0.318
MMP7_B2	-0.082	0.825	0.009	0.289	-0.083	0.62	0.006	0.157
PD.L1_B2	-0.53	0.521	0.019	0.348	-0.075	0.841	0.004	0.65
PDGF.subunit.B_B2	-0.917	0.299	0.039	0.118	-0.167	0.66	0.014	0.2
PGF_B2	-0.052	0.942	0.021	0.229	-0.044	0.903	0.009	0.278
PTN_B2	0.033	0.926	0.006	0.418	-0.06	0.733	0.004	0.325

TNFRSF12A_B2	-1.277	0.296	0.06	0.106	-0.025	0.965	0.006	0.683
TNFRSF21_B2	-0.817	0.403	0.026	0.336	-0.245	0.585	0.008	0.483
TNFRSF9_B2	-0.166	0.714	0.016	0.162	0.003	0.988	0.006	0.212
TNFSF14_B2	-0.281	0.449	0.017	0.117	-0.027	0.872	0.005	0.224
TRAIL_B2	-0.229	0.628	0.018	0.156	-0.162	0.499	0.009	0.159
TWEAK_B2	-1.129	0.277	0.043	0.148	-0.563	0.249	0.018	0.172
VEGFA_B2	-1.131	0.278	0.044	0.129	-0.474	0.322	0.013	0.274
VEGFR.2_B2	-0.223	0.808	0.015	0.521	-0.052	0.902	0.009	0.404

Table 7a: Model II Contrasts for Main Effects of Proteins vs PRAGMA-%Dis

	Contrast								
	E2 ($\beta_1 + 12 \cdot \beta_3$)			**E4 ($\beta_1 + 36 \cdot \beta_3$)**			**E6 ($\beta_1 + 60 \cdot \beta_3$)**		
Protein	**Estimate**	**P-Value**	**FDR P-Value**	**Estimate**	**P-Value**	**FDR P-Value**	**Estimate**	**P-Value**	**FDR P-Value**
ADA_B2	-0.151	0.606	0.976	0.313	0.03	0.124	0.777	0	0.001
ARG1_B2	-0.185	0.403	0.976	0.301	0.01	0.117	0.787	0	0.000
CAIX_B2	-0.457	0.353	0.976	0.232	0.399	0.626	0.92	0.028	0.059
CASP.8_B2	-0.216	0.493	0.976	0.192	0.309	0.565	0.6	0.091	0.161
CCL17_B2	-0.49	0.075	0.976	-0.083	0.735	0.859	0.325	0.443	0.520
CCL19_B2	-0.154	0.547	0.976	-0.003	0.985	0.989	0.147	0.578	0.652
CCL20_B2	-0.136	0.46	0.976	0.145	0.192	0.424	0.427	0.01	0.032
CCL23_B2	-0.453	0.152	0.976	-0.206	0.378	0.617	0.042	0.912	0.943
CCL3_B2	-0.218	0.409	0.976	0.259	0.091	0.291	0.735	0.006	0.023
CCL4_B2	-0.16	0.556	0.976	0.375	0.005	0.117	0.911	0	0.000
CD244_B2	-0.147	0.731	0.976	0.244	0.352	0.589	0.634	0.139	0.233
CD27_B2	-0.127	0.771	0.976	0.434	0.102	0.291	0.995	0.014	0.038
CD40_B2	-0.101	0.785	0.976	0.33	0.131	0.324	0.76	0.021	0.046
CD4_B2	-0.182	0.681	0.976	0.021	0.941	0.989	0.224	0.646	0.715
CD5_B2	-0.015	0.96	0.976	0.17	0.31	0.565	0.356	0.146	0.238
CD83_B2	-0.509	0.258	0.976	0.553	0.103	0.291	1.615	0.011	0.032
CD8A_B2	-0.077	0.729	0.976	-0.127	0.338	0.589	-0.177	0.372	0.452
CSF.1_B2	-0.192	0.51	0.976	0.522	0.001	0.069	1.236	0	0.000
CX3CL1_B2	-0.141	0.719	0.976	0.133	0.571	0.748	0.406	0.262	0.353
CXCL10_B2	-0.17	0.264	0.976	-0.062	0.6	0.755	0.046	0.823	0.865
CXCL11_B2	-0.088	0.675	0.976	-0.046	0.756	0.859	-0.004	0.986	0.986
CXCL13_B2	-0.106	0.654	0.976	0.073	0.621	0.755	0.252	0.226	0.342
CXCL1_B2	-0.522	0.113	0.976	0.209	0.351	0.589	0.939	0.03	0.062
CXCL5_B2	-0.223	0.24	0.976	0.032	0.825	0.882	0.287	0.248	0.348
CXCL9_B2	0.008	0.976	0.976	0.186	0.249	0.510	0.364	0.067	0.122

EGF_B2	0.494	0.203	0.976	0.535	0.017	0.117	0.575	0.021	0.046
FASLG_B2	-0.212	0.745	0.976	0.125	0.762	0.859	0.462	0.497	0.570
Gal.1_B2	-0.275	0.574	0.976	0.188	0.476	0.682	0.65	0.05	0.098
Gal.9_B2	-0.103	0.893	0.976	0.509	0.185	0.424	1.122	0.019	0.044
GZMA_B2	-0.049	0.893	0.976	0.108	0.56	0.748	0.265	0.242	0.348
GZMB_B2	-0.096	0.661	0.976	0.144	0.272	0.528	0.385	0.058	0.109
GZMH_B2	-0.059	0.83	0.976	0.08	0.618	0.755	0.219	0.344	0.435
HGF_B2	-0.079	0.803	0.976	0.451	0.01	0.117	0.981	0	0.000
HO.1_B2	-0.118	0.558	0.976	-0.066	0.576	0.748	-0.015	0.94	0.956
ICOSLG_B2	0.07	0.923	0.976	0.803	0.026	0.117	1.536	0	0.000
IL.1.alpha_B2	-0.108	0.764	0.976	0.23	0.253	0.510	0.567	0.043	0.086
IL12_B2	0.023	0.968	0.976	0.194	0.579	0.748	0.366	0.445	0.520
IL18_B2	-0.349	0.304	0.976	0.066	0.789	0.859	0.482	0.194	0.301
IL6_B2	-0.126	0.611	0.976	0.301	0.052	0.203	0.729	0.005	0.021
IL7_B2	0.064	0.919	0.976	0.279	0.414	0.626	0.495	0.253	0.348
IL8_B2	-0.134	0.584	0.976	0.264	0.026	0.117	0.662	0	0.001
LAMP3_B2	-0.434	0.23	0.976	-0.45	0.025	0.117	-0.466	0.183	0.290
LAP.TGF.beta.1_B2	-0.313	0.618	0.976	1.298	0.016	0.117	2.909	0.003	0.013
MCP.1_B2	-0.128	0.664	0.976	0.324	0.062	0.215	0.776	0.002	0.008
MCP.2_B2	-0.117	0.62	0.976	-0.003	0.986	0.989	0.111	0.682	0.742
MCP.3_B2	-0.086	0.808	0.976	0.172	0.411	0.626	0.43	0.139	0.233
MCP.4_B2	-0.141	0.781	0.976	-0.004	0.989	0.989	0.133	0.747	0.799
MIC.A.B_B2	-0.035	0.929	0.976	-0.203	0.427	0.630	-0.37	0.367	0.452
MMP12_B2	-0.044	0.8	0.976	0.233	0.014	0.117	0.511	0	0.001
MMP7_B2	0.03	0.909	0.976	0.256	0.108	0.291	0.481	0.018	0.044
PD.L1_B2	-0.298	0.593	0.976	0.167	0.636	0.758	0.631	0.24	0.348
PDGF.subunit.B_B2	-0.451	0.409	0.976	0.481	0.122	0.314	1.413	0.006	0.023
PGF_B2	0.198	0.696	0.976	0.697	0.011	0.117	1.196	0	0.002
PTN_B2	0.109	0.667	0.976	0.261	0.057	0.206	0.413	0.011	0.032
TNFRSF12A_B2	-0.558	0.465	0.976	0.881	0.102	0.291	2.32	0.009	0.031

TNFRSF21_B2	-0.507	0.412	0.976	0.112	0.79	0.859	0.731	0.326	0.425
TNFRSF9_B2	0.025	0.937	0.976	0.409	0.023	0.117	0.792	0	0.002
TNFSF14_B2	-0.082	0.737	0.976	0.318	0.017	0.117	0.717	0.001	0.003
TRAIL_B2	-0.014	0.964	0.976	0.415	0.022	0.117	0.844	0.001	0.005
TWEAK_B2	-0.618	0.31	0.976	0.404	0.145	0.346	1.426	0.015	0.039
VEGFA_B2	-0.605	0.345	0.976	0.445	0.255	0.510	1.496	0.019	0.044
VEGFR.2_B2	-0.044	0.946	0.976	0.314	0.484	0.682	0.672	0.329	0.425

Table 7b: Model II Contrasts for Main Effects of Proteins vs PRAGMA-%Bx

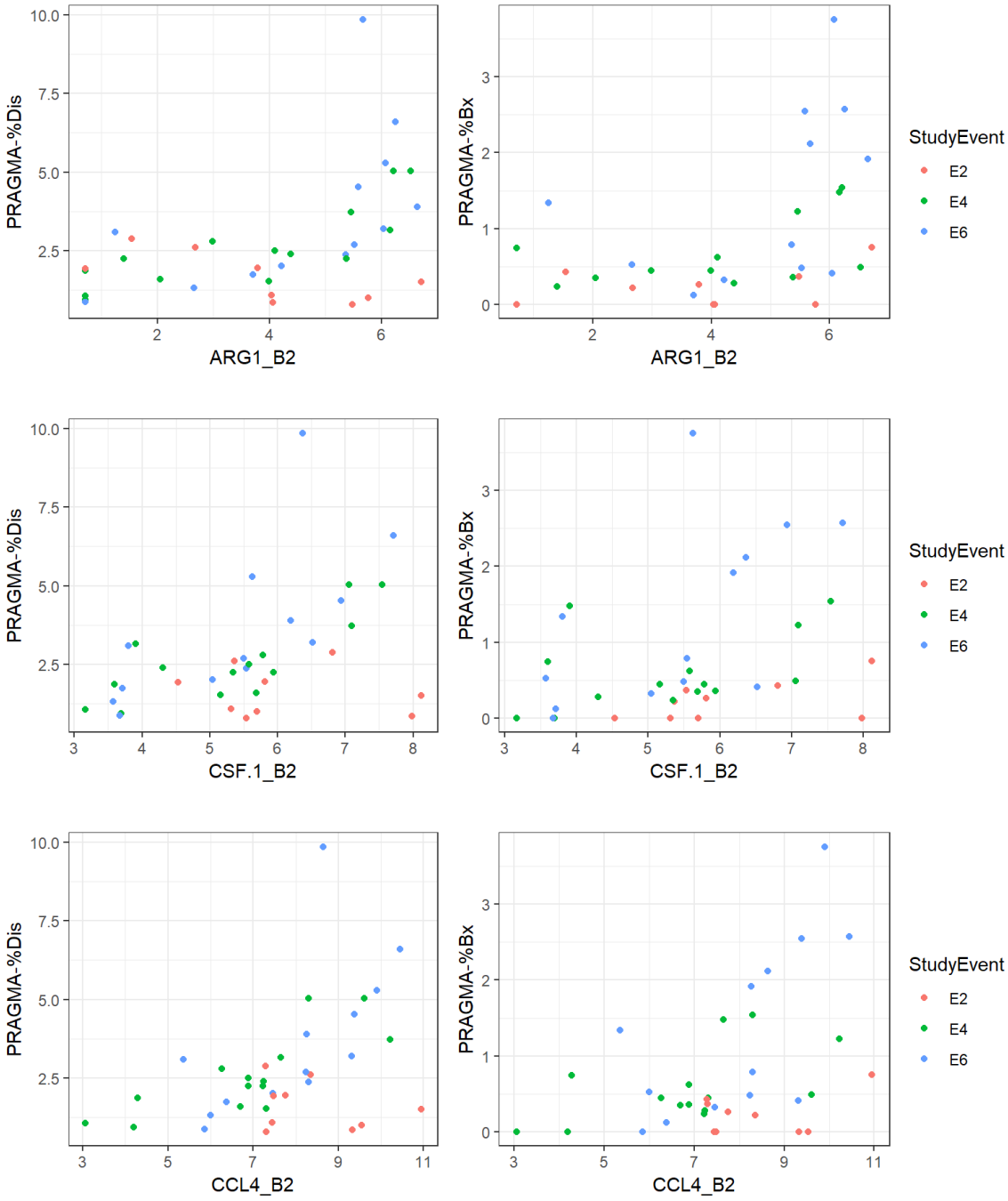
	Contrast								
	E2 ($\beta_1 + 12 \cdot \beta_3$)			**E4 ($\beta_1 + 36 \cdot \beta_3$)**			**E6 ($\beta_1 + 60 \cdot \beta_3$)**		
Protein	**Estimate**	**P-Value**	**FDR P-Value**	**Estimate**	**P-Value**	**FDR P-Value**	**Estimate**	**P-Value**	**FDR P-Value**
ADA_B2	0.036	0.800	0.997	0.166	0.015	0.095	0.297	0.002	0.009
ARG1_B2	0.006	0.957	0.997	0.153	0.007	0.095	0.300	0.000	0.006
CAIX_B2	-0.083	0.738	0.997	0.033	0.816	0.878	0.149	0.518	0.594
CASP.8_B2	-0.069	0.642	0.997	0.101	0.236	0.488	0.271	0.100	0.218
CCL17_B2	-0.050	0.730	0.997	-0.053	0.638	0.796	-0.056	0.776	0.789
CCL19_B2	-0.018	0.884	0.997	0.018	0.821	0.878	0.054	0.666	0.700
CCL20_B2	-0.047	0.598	0.997	0.069	0.170	0.407	0.184	0.018	0.058
CCL23_B2	-0.024	0.880	0.997	-0.084	0.452	0.774	-0.143	0.448	0.558
CCL3_B2	-0.082	0.514	0.997	0.111	0.113	0.326	0.305	0.017	0.058
CCL4_B2	-0.054	0.664	0.997	0.168	0.006	0.095	0.391	0.000	0.002
CD244_B2	-0.067	0.740	0.997	0.078	0.529	0.781	0.223	0.286	0.455
CD27_B2	0.031	0.879	0.997	0.196	0.115	0.326	0.361	0.058	0.133
CD40_B2	-0.008	0.962	0.997	0.150	0.129	0.333	0.309	0.048	0.118
CD4_B2	-0.064	0.757	0.997	-0.027	0.835	0.878	0.010	0.965	0.965
CD5_B2	-0.018	0.898	0.997	0.057	0.470	0.774	0.132	0.267	0.455
CD83_B2	-0.066	0.769	0.997	0.135	0.415	0.774	0.336	0.282	0.455

CD8A_B2	-0.108	0.288	0.997	-0.078	0.197	0.436	-0.048	0.593	0.645
CSF.1_B2	-0.033	0.822	0.997	0.196	0.023	0.102	0.425	0.001	0.008
CX3CL1_B2	0.011	0.955	0.997	0.057	0.600	0.795	0.103	0.551	0.621
CXCL10_B2	-0.037	0.625	0.997	0.021	0.707	0.829	0.078	0.428	0.558
CXCL11_B2	0.012	0.900	0.997	0.025	0.709	0.829	0.038	0.740	0.765
CXCL13_B2	-0.006	0.958	0.997	0.045	0.502	0.777	0.096	0.335	0.490
CXCL1_B2	-0.214	0.206	0.997	0.059	0.568	0.795	0.333	0.116	0.240
CXCL5_B2	-0.047	0.622	0.997	0.042	0.525	0.781	0.130	0.273	0.455
CXCL9_B2	0.033	0.797	0.997	0.112	0.116	0.326	0.190	0.043	0.112
EGF_B2	0.130	0.550	0.997	0.139	0.214	0.457	0.148	0.263	0.455
FASLG_B2	-0.066	0.824	0.997	0.142	0.449	0.774	0.351	0.266	0.455
Gal.1_B2	-0.063	0.786	0.997	0.022	0.862	0.891	0.107	0.499	0.584
Gal.9_B2	-0.042	0.909	0.997	0.130	0.480	0.774	0.302	0.184	0.368
GZMA_B2	-0.054	0.750	0.997	0.038	0.655	0.796	0.130	0.230	0.445
GZMB_B2	0.003	0.978	0.997	0.128	0.022	0.102	0.254	0.004	0.017
GZMH_B2	-0.035	0.780	0.997	0.040	0.580	0.795	0.116	0.282	0.455
HGF_B2	0.061	0.702	0.997	0.218	0.009	0.095	0.374	0.001	0.008
HO.1_B2	0.004	0.969	0.997	-0.032	0.560	0.795	-0.068	0.491	0.584
ICOSLG_B2	0.238	0.465	0.997	0.416	0.017	0.095	0.593	0.000	0.006
IL.1.alpha_B2	0.008	0.960	0.997	0.188	0.025	0.104	0.369	0.002	0.010
IL12_B2	-0.027	0.917	0.997	0.083	0.603	0.795	0.192	0.395	0.544
IL18_B2	-0.027	0.872	0.997	0.039	0.738	0.847	0.105	0.572	0.633
IL6_B2	-0.012	0.920	0.997	0.120	0.104	0.326	0.251	0.043	0.112
IL7_B2	-0.078	0.796	0.997	0.038	0.810	0.878	0.155	0.450	0.558
IL8_B2	-0.005	0.965	0.997	0.131	0.018	0.095	0.268	0.001	0.009
LAMP3_B2	0.029	0.873	0.997	-0.206	0.035	0.129	-0.442	0.002	0.009
LAP.TGF.beta.1_B2	0.087	0.774	0.997	0.602	0.018	0.095	1.118	0.021	0.064
MCP.1_B2	0.002	0.990	0.997	0.128	0.125	0.333	0.255	0.033	0.094
MCP.2_B2	0.020	0.852	0.997	0.064	0.402	0.774	0.108	0.392	0.544
MCP.3_B2	0.021	0.898	0.997	0.081	0.406	0.774	0.140	0.295	0.457

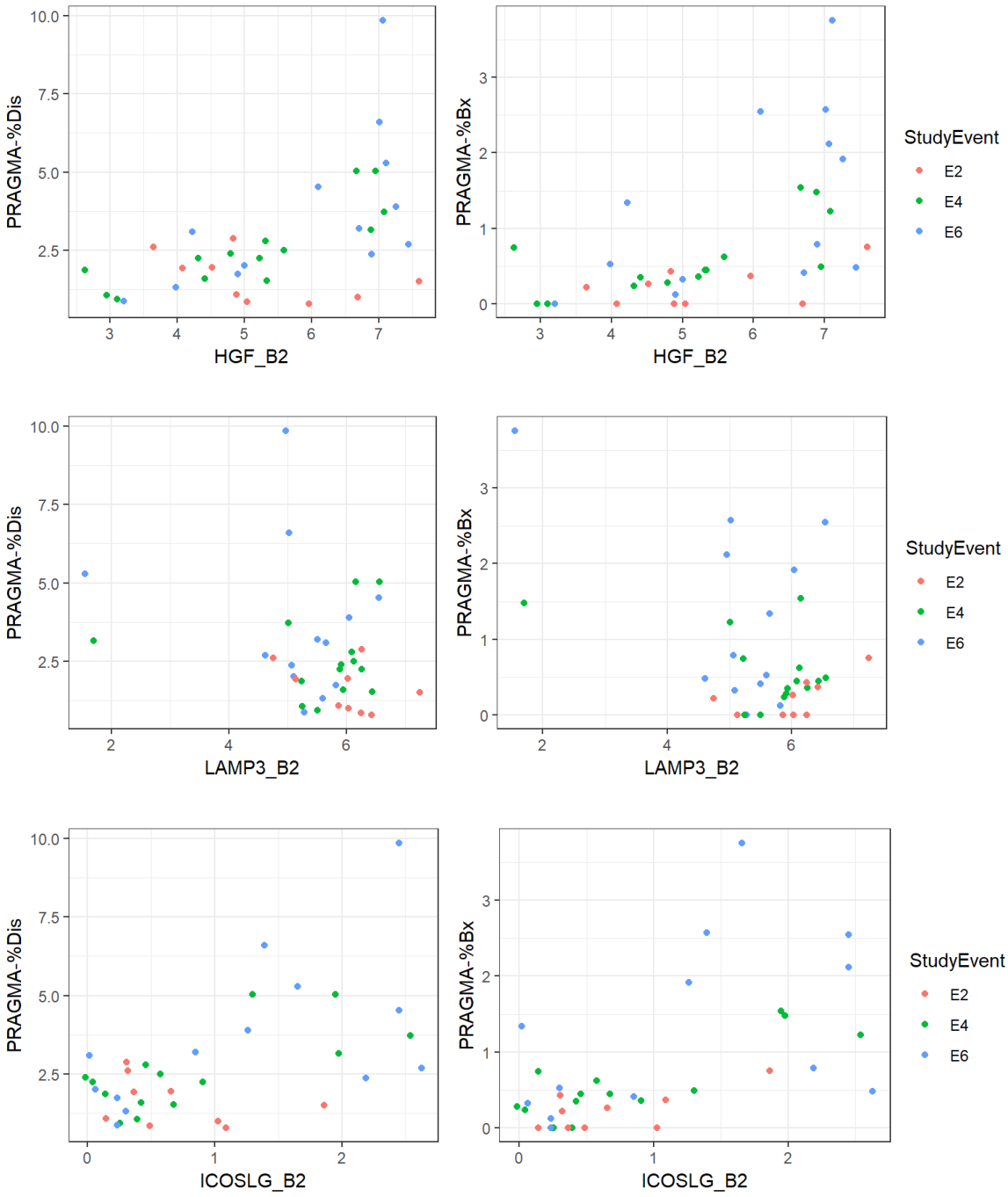
MCP.4_B2	-0.056	0.813	0.997	0.014	0.921	0.936	0.083	0.664	0.700
MIC.A.B_B2	0.061	0.739	0.997	-0.057	0.622	0.796	-0.176	0.340	0.490
MMP12_B2	0.036	0.689	0.997	0.110	0.016	0.095	0.185	0.003	0.014
MMP7_B2	-0.008	0.947	0.997	0.143	0.035	0.129	0.293	0.002	0.009
PD.L1_B2	-0.025	0.923	0.997	0.074	0.648	0.796	0.173	0.490	0.584
PDGF.subunit.B_B2	-0.001	0.997	0.997	0.330	0.018	0.095	0.661	0.005	0.019
PGF_B2	0.069	0.785	0.997	0.296	0.018	0.095	0.522	0.001	0.009
PTN_B2	-0.012	0.923	0.997	0.085	0.185	0.425	0.182	0.024	0.071
TNFRSF12A_B2	0.046	0.904	0.997	0.189	0.474	0.774	0.331	0.449	0.558
TNFRSF21_B2	-0.143	0.630	0.997	0.061	0.755	0.851	0.265	0.443	0.558
TNFRSF9_B2	0.078	0.589	0.997	0.229	0.003	0.095	0.379	0.000	0.002
TNFSF14_B2	0.038	0.739	0.997	0.168	0.005	0.095	0.299	0.001	0.009
TRAIL_B2	-0.054	0.735	0.997	0.163	0.046	0.160	0.379	0.002	0.010
TWEAK_B2	-0.344	0.223	0.997	0.094	0.487	0.774	0.531	0.055	0.132
VEGFA_B2	-0.316	0.294	0.997	-0.002	0.992	0.992	0.312	0.302	0.457
VEGFR.2_B2	0.059	0.842	0.997	0.279	0.151	0.375	0.499	0.102	0.218

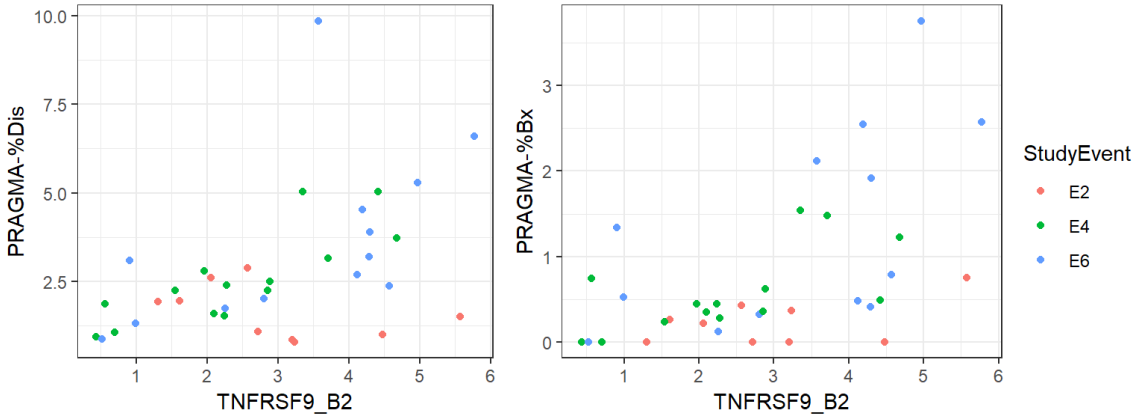
8. APPENDIX C: Scatterplots of Significant Proteins

C.1 Proteins Significant in Cross-Sectional Analysis



C.2 Proteins Significant in Linear Mixed Models and Lasso Penalized Regression




PRAGMA-%Dis
10.0
7.5
5.0
2.5
1
2
3
4
5
6
TNFRSF9_B2
PRAGMA-%Bx
3
2
1
0
1
2
3
4
5
6
TNFRSF9_B2
StudyEvent
E2
E4
E6