Biomarkers Consortium Project
Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in Alzheimer's Disease

Statistical Analysis Plan

1 Introduction

The Analysis Plan described within this document represents the work of the Biomarkers Consortium Project "**Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in Alzheimer's Disease"**. This project was submitted to the Biomarkers Consortium Neuroscience Steering Committee by a subgroup of the Alzheimer's Disease Neuroimaging Initiative (ADNI) Industry Scientific Advisory Board (ISAB) for execution and was managed by a Biomarkers Consortium Project Team that includes members from academia, government and the pharmaceutical industry; funding for this project was provided through an overage of funds raised by the Foundation for NIH for the ADNI partnership, as well as Pfizer Inc. This project is intended to be the first part of a multi-phased effort seeking to utilize samples collected by ADNI to qualify multiplex panels in both plasma and cerebrospinal fluid (CSF) to diagnose patients with Alzheimer's Disease (AD) and monitor disease progression.

Biomarker tools for early diagnosis and disease progression in Alzheimer's disease (AD) remain key issues in AD drug development. Identification and validation of cost-effective, non-invasive methods to identify early AD and to monitor treatment effects in mild-moderate AD patients could revolutionize current clinical trial practice. Treatment prior to the onset of dementia may also ensure intervention occurs before irreversible neuropathology.

The aim of the project is to determine the ability of a multiplex plasma based immunoassay panel to discriminate among disease states and to monitor disease progression over a one year period. The multiplex panel is based upon luminex immunoassay technology and a 151 analyte panel has been developed by Rules Based Medicine (RBM) to measure a range of inflammatory, metabolic, lipid and other disease relevant endpoints. Prior studies using an older version of the RBM panel (an 89 analyte version) suggested some analytes on the panel differed between AD and controls. The panel has been expanded to include analytes from a recent article describing plasma based biomarkers of AD.

The analyses described in this statistical analysis plan should be regarded as exploratory and meant for hypothesis and model generation, rather than for hypothesis confirmation and model validation. Findings will need to be confirmed and expanded upon in subsequent studies using other, independent data sets.

2 Study Design and Objectives

2.1 Study Design

Samples from baseline and one year ADNI plasma sample set will also be assessed (N=229 Controls, 192 AD, 398 for amnestic mild cognitive impairment (MCI)). Of the 398 MCI subjects, 153 subjects have progressed to dementia as of March 2010. This statistical analysis plan addresses the analysis of data from a subset of these samples.

Previously, a small pilot study was run using University of Pennsylvania banked plasma samples from AD (N~98) and control (N~72) subjects. Data from the pilot study have already been analyzed **(Hu et al., in preparation)**. Assessing the utility of the expanded RBM panel will incorporate some of the findings from the pilot study.

2.2 Study Objectives

- To determine whether baseline levels or changes from baseline levels for individual analytes are associated with patient demographics (age, gender) or disease status.
- To determine whether baseline levels for a combination of analytes derived from either a biological pre-selection based method and/or from a statistically based/machine learning language approach will provide a panel with distinctly different profiles for the ADNI normal controls (NC), MCI or AD.
- To determine whether baseline levels for a combination of analytes derived from either a biological pre-selection based method and/or from a statistically based/machine learning approach will provide a panel that discriminates pre-demented subjects who will progress to dementia in one year and/or two years.
- To determine whether change from baseline levels for a combination of analytes (derived as above) predict cognitive decline in AD or correlate with disease progression.

3 Univariate Analysis

Univariate analyses will be performed first. The results of the univariate analyses may be used to inform and select analytes to be used in the pathway analyses and multivariate predictive model-building. Results from the univariate and multivariate sets of analyses will be compared for overlap and a final panel selected based on optimal overlap.

3.1 Classification Endpoints

Clinical diagnosis at time of enrollment/collection will be used to classify AD, MCI and control groups. Clinical diagnosis of amnestic MCI followed by diagnosis of AD will be used to classify pre-demented progressors.

3.2 Data Quality Control (QC)

Up to 190 analytes may be measured in the plasma updated RBM panel. Plasma data will be analyzed separately and compared for each analyte dependent upon sample availability. The data will be prepared for all analysis as follows:

- Review of the quality control samples data for each run to determine the variability characteristics of the spiked plasma (or serum) QC samples. Characteristics examined for the LOW, MEDIUM and HIGH QC samples for each biomarker will include mean, standard deviation (SD) and the percent coefficient of variation (%CV) for each analyte to determine not only the variability at each concentration but whether or not there is a major change in variability across the concentration range for each analyte.
- Analytes with more than 10% missing data will not be analyzed further. Missing data are generally indicated by "QNS" (quantity not sufficient for analysis) by RBM.
- Analytes with more than 10% recorded as "LOW" will not be included in the multivariate analysis. These analytes will be assessed to compare the proportion of measurable samples in each disease status category. If proportions differ substantially among disease status categories for some analytes, alternative approaches may be explored for incorporating such analytes in the multivariate analyses described below.
- For expression values preceded by a "<" or ">" modifier, the numeric portion of the value will be used for all subsequent data preparation and analyses.
- For each analyte, the distribution of measured values within each diagnostic group will be examined. If the distributions are not normal, the team will seek appropriate transformations (e.g., Box-Cox transformations **(Box and Cox, 1964)** so the transformed markers approximate normality. All subsequent data preparation and analyses will then be conducted on the transformed values.
- Analytes with less than 10% missing/"LOW" values will have the non-numeric values imputed as follows:
    - Values recorded as "LOW" will be imputed to LLD/2
    - Missing values will be imputed to be the mean of the non-missing values for that analyte.
    - Samples with imputed values for more than 25% of the analytes will be excluded from the analysis
- Multidimensional scaling and/or Mahalanobis distances will be used to detect sample outliers and misclassified subjects.
- For univariate analysis, outliers that are more than 5 STD from mean will be assigned the value of the nearest non-outlier point. For outliers apparent in

multivariate reviews, outliers will be imputed using a nearest neighbor or other appropriate algorithm.

The imputation and outlier definition strategy defined above is only one of many possible strategies that could be used. If resources permit a limited number of alternative strategies may be used to assess the robustness of the analytical conclusions obtained using the strategy defined above.

As part of data QC, patient, visit, and sample identifiers will be checked for uniqueness and logical consistency. Graphical displays will be used to check for systematic patterns related to batch, run date, sample quality measures, and QC sample characteristics.

Cleaning, outlier detection, and distribution displays of all samples will be performed prior to merging phenotype data with the biomarker data. Misclassification assessment will be performed prior to statistical analysis.

3.3 General approach

Analysis of variance (ANOVA) and analysis of covariance (ANOCOVA) models will be used to compare mean analyte levels among groups of interest. These ANOVA/ANOCOVA models will include the diagnosis/disease status group and other covariates including age, gender and apoE4 genotype/status, as well as possible interactions among these factors. The interactive effect between group and other covariates will be tested. Depending on the outcome of these tests, the differences between groups will be tested either by the main effect of diagnosis or the effect of diagnosis at a fixed level of other covariates (i.e., apoE4 status) or through the adjusted least square means.

A major analytic concern in these tests is the control of overall type I error rate due to the relatively large number of CSF and plasma proteins tested in this aim. The team will address this concern using false discovery rate (FDR) methodology.

3.4 Hypotheses to Be Tested

The following univariate hypotheses will be addressed for each analyte:

HO1i: Analyte i is not associated with age [age treated as a continuous variable]
HO2i: Analyte i is not associated with gender
HO3i: Analyte i is not associated with ApoE status

HO4i: Analyte i is not associated with disease status or change in disease status (adjusted for age, gender, and/or ApoE status as necessary)

An initial set of analyses will look at whether the mean baseline level of each individual marker differs among disease groups (normal, MCI, AD) via an ANOVA or ANCOVA and t-test analysis. "Disease status" will be based on the clinical calls recorded in the ADNI database. Additional analyses may be conducted using disease status defined using one or more alternative definitions based on cognitive and/or functional tests.

False discovery (FDR) corrections will be applied to p-values and will be reported along with raw p-values. When adjusting for and evaluating the impact of multiple tests, a distinction will be made between:

- a set of proteins defined a priori as being of particular biological interest based on review of pathway annotations (see Table 1)
- the remaining assayed proteins.

FDR corrections based on the Benjamini-Hochberg method will be calculated separately for the two sets.

A second set of analyses will be performed using data only from MCI subjects. ANOVA/ANCOVAs similar to the above will be run to assess whether mean baseline levels of the analytes differ among MCI non-converters and converters.

A third set of analyses will be run to determine whether change from baseline analyte levels at one year are associated with change in disease status.

A fourth set of analyses will be run to determine whether any of the analytes correlate with significant changes in Clinical Dementia Rating Scale-Sum of the Boxes (CDR-SB) or Auditory-Verbal Learning Test (AVLT).

A fifth set of analyses will determine whether levels of any of the analytes are associated with low CSF abeta/high tau, high amyloid brain burden and significant brain atrophy.

Analyses to examine relationships between analyte levels and use of acetyl cholinesterase inhibitors or other medications by subjects may also be performed.

4 Pathway Analysis of Biomarkers

Although statistical machine learning-based approaches can generate a short list of discriminatory proteins, such analyses reveal little about biological relevance. In addition to machine learning approaches, the current proposal will use a systems biology approach to better understand pathway relationships between identified proteins. The Project Team will use pathway mining tools, such as those offered by Ingenuity and Pathway studio, to find the functional connections between the markers from plasma samples. This will provide direct evidence to support key hypotheses. To further increase the biological

relevance of the protein markers in the predictive models, biomarkers will be selected based on their presence in distinct biological pathways.

In addition empirical characterizations of marker data such as pair-wise correlations or higher-order relations (e.g. principal components analysis (PCA)) will be used. This analysis will derive an initial short list that will then be analyzed using multivariate and machine learning language approaches.

5 Multiple Marker Models

Multivariate statistical methods and multiple machine learning approaches will be used to identify the optimal combination of a group of proteins to predict disease status. The problem of classification and prediction has received a great deal of attention in mining "-omics" data. In the case of this project, the task will be to classify and predict the diagnostic category of a sample on the basis of protein quantitative profiles. The main type of statistical problem is the identification of "marker" genes that characterize the difference between diagnosis groups (e.g. AD, MCI) – the so called "variable/feature selection" problem. One challenge is to find the optimal combination of uncorrelated proteins. This factor not only is very important to improve prediction accuracy but also contributes to the merits of a good classifier: the simplicity and insight gained into the predictive structure of the data.

In all multivariate model building, feature selection will be done using data only from the training set. Feature selection based on a completely independent data set is not feasible for this project due to sample size and the fact that this is the first study to use this version of the RBM panel.

Multiple marker analysis will be used to build relationships to the disease groups. The candidate models include: logistic regression, linear discriminant analysis, nearest shrunken centroid, random forests, support vector machines and partial least squares. The technique of **Xiong et al. (2004)** may be applied to search for the linear combination of informative proteins that optimally discriminates between the diagnostic groups. Models generated by the various methods will be compared and the "best" model will be chosen based on model fit, robustness, and parsimony considerations.

Models will be fit with two sets of covariates, 1) assay results only and 2) assays results plus additional patient information including gender, age, and ApoE4 allele status Other biomarkers such as amyloid PIB load, hippocampal atrophy, baseline mini-mental state examination (MMSE), and/or baseline Alzheimer's Disease Assessment Scale-Cognitive Subscale 11 (ADAS-cog 11) may also be used. For a specific model, differences in performance between models fit using the two classes of predictors variables should be characterized to understand the predictive ability of the assays beyond that of routine clinical information on the patients. If possible, formal inference should be made regarding the statistical significance of including the assay variables above and beyond that of the clinical data. Analysis will focus on the following:

- good characterizations of error rates; poor fitting models should not be interpreted.
- any feature selection routines should be extensively cross-validated **(see Ambroise and McLachlan, 2002)**
- measures of marker importance should be biased towards those that use uncertainty (e.g. logistic regression slope tests) as opposed to those that do not (e.g. random forest variable importance, etc).

The multivariate results will be compared to the single marker analysis and (especially) biological relevance.

In addition to the modeling efforts outlined above and described in detail in sections 5.1 and 5.2, an additional set of analyses will be based upon confirming a predictive model developed from the University of Pennsylvania pilot dataset. In brief, a predictive model using pilot data for the 24 analytes shown in Table 1 was identified as providing good discrimination between AD vs Control.

Table 1. Analytes in Predictive Model from UPenn Pilot Study

Analytes	Analyte	Analyte
Alpha-1 Microglobulin ↑	FAS ↑	Resistin ↑
Angiopoietin 2 ↑	HGF ↑	Stem Cell Factor ↑
Apolipoprotein E ↓	IGF-BP2 ↑	Tenascin C ↑
Beta2-Microglobulin ↑	IL-10 ↑	Thrombomodulin ↑
B Lymophocyte Chemoattractant ↑	NT-proBNP ↑	TIMP-1 ↑
Cortisol ↑	Osteopontin ↑	VCAM-1 ↑
E-Selectin ↓	Pancreatic Polypeptide ↑	VEGF ↑
FABP ↑	PAPP-A ↑	Von Willebrand Fact ↑

The model based on pilot data for these 24 analytes will be tested using the ADNI dataset for its predictive abilities to discriminate AD from Control, MCI from Control and MCI rapid progressors to dementia (within 2 years) vs slow or stable MCI The algorithm for the pilot study prediction model will be provided by the model developers.

5.1 Analyte Filtering

Several approaches to filtering and feature selection may be examined. Results of the univariate analyses described above may be used to define a starting set of markers for the analysis. Results of the pathway analysis may also be used to define a starting set. In addition, pre-filtering of markers in an unsupervised fashion prior to building models based on empirical measures may also be applied.

5.2 Model Building Approach

For each type of model, predictive model building will be based on an iterative resampling approach.

For each of the K resampling iterations, the steps will include:

- Splitting the data into training and test sets
- Applying an unsupervised filter on the predictors based on data in the training set only.
- Building and tuning the predictive model on the current training set
- Predicting the current test set
- Calculating and saving the performance (classification accuracy, Kappa)
- End resampling iteration
- Assess performance of the model over the K sets of performance metrics

In the above algorithm, the resampling schemes can include cross-validation, the bootstrap and repeated training/test set splits (see Appendix). Methods for unsupervised feature selection can include filters on variance of individual predictors, high pair-wise predictor correlations, etc.

In addition to the iterative resampling approach above, an additional set of predictive models may be built based on the training-test 60-40 split data sets and 10-fold cross-validation set defined by the ADNI Biostatics Core. This set of models will permit easier comparisons with other modeling efforts performed using the ADNI Biostat core-defined approach.

6 Power Calculations

The sample size for this project and resulting analyses are based upon and limited by the availability of samples from both the pilot and ADNI samples. Additional post-hoc analysis will be completed based upon variability characteristics of the current study to understand power requirements for subsequent analysis of future datasets, in discussion with the Project Team.

7 References

Ambroise C. and McLachlan G.J. (2002), 2002, Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. U.S.A., 99, 6562-6. Epub 2002 Apr 30.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. J. Royal Stat. Soc. B26, 211-246.

Hu, W.T., Chen-Plotkin, A., Shaw, L.M., Arnold, S., Grossman, M., Clark, C.M., Leight, S., McCluskey, L., Ellman, L., Karlawish, J., Hurtig, H.I., Siderowf, A., Lee, V.M-Y., Trojanowski, J.Q., Swensen, F., Pickering, E., Kuhn, M., Soares, H.D. A study to assess the diagnostic accuracy of a plasma-based multiplexed immunoassay panel to identify Alzheimer's disease. In preparation.

Xiong, C., McKeel, D.W., Jr., Miller, J.P. and Morris, J.C. (2004). Combining correlated diagnostic tests: application to neuropathologic diagnosis of Alzheimer's disease. Med. Decis. Making 24, 659-69.

8 Appendix I

In order to make sure that there is balance among the pre-specified cross-validation groups, we present a table of the 10 partitions vs. the baseline-year1 status of ADNI subjects. No info from month 6 or other visits were used.

	AD-AD	AD-MCI	AD-NA	MCI-AD	MCI-MCI	MCI-NA	MCI-NL	NL-MCI	NL-NA	NL-NL
a	16	0	3	7	29	5	0	0	0	21
b	16	1	2	10	25	3	1	0	1	23
c	16	0	1	6	32	2	1	0	1	21
d	17	0	3	6	29	4	1	0	1	21
e	15	0	5	5	28	8	0	1	2	21
f	15	0	2	4	29	3	3	2	1	20
g	17	0	5	6	30	2	0	0	2	21
h	18	0	2	6	27	4	1	0	3	19
i	16	1	3	6	31	3	0	0	2	23
j	15	0	4	8	27	3	2	0	3	20

The following table restricts to the primary groups of interest.

	AD-AD	MCI-AD	MCI-MCI	NL-NL
a	16	7	29	21
b	16	10	25	23
c	16	6	32	21
d	17	6	29	21
e	15	5	28	21
f	15	4	29	20
g	17	6	30	21
h	18	6	27	19
i	16	6	31	23
j	15	8	27	20