A. SPECIFIC AIMS

The Biostatistics Core (Core C) will bridge the transition from the DIAN database stored and managed by the Informatics Core (Core H) to the analyses of the longitudinal data within, between, and among the various data domains, the latter of which serves as the primary responsibility of the Core. Specifically the Biostatistics Core will: **(1)** Oversee the statistical quality control of data for the study and produce appropriately de-identified and statistically analyzable datasets for distribution/analysis; **(2)** Lead the statistical data analyses and collaborate in report preparation for all Cores and projects, and consult on the design of all projects and on the application of appropriate statistical and methodological techniques; **(3)** Develop and implement appropriate statistical models for longitudinal changes in potential markers and use these models to test the statistical hypotheses about the preclinical changes of AD and about the temporal difference on preclinical changes of AD among various markers, and **(4)** Serve as an advisory group for other researchers interested in using the DIAN database for additional analyses.

B. SIGNIFICANCE

The scientific significance of establishing an international network for identification, evaluation, and follow-up of families with early onset dominantly inherited Alzheimer's disease is adequately addressed in the overview of this application and in Clinical Core's (Core B) application described elsewhere. The activities of the Biostatistics Core (Core C) are designed to enhance the research objectives of DIAN by serving the DIAN investigators with a smooth transition from the database to statistical analyses, providing appropriate statistical analysis resources to each of the Cores and projects, and developing the necessary longitudinal statistical models to test the preclinical hypotheses of DIAN.

The major hypotheses in DIAN conjectured a period of preclinical (presymptomatic) AD in individuals who are destined to develop early onset dementia (gene carriers) that can be detected by changes in biological fluids and in neuroimaging correlates in comparison with individuals who will not develop early onset dementia (noncarriers) and a temporal difference in preclinical changes among the CSF molecular biomarkers, neuroimaging markers, and cognitive markers for the gene carriers. The methodological significance of these hypotheses is the fact that, in addition to the standard longitudinal statistical models such as the general and generalized linear mixed models (Diggle et al. 2002) and the Cox's proportional hazards models (Cox 1972), DIAN requires novel longitudinal statistical methodologies to adequately test these hypotheses. The hallmark of these novel approaches should be the modeling of the rates of progression on repeatedly measured disease markers during the preclinical period and their association with the risk of subsequent development of AD.

A statistical conceptualization on the existence of a preclinical stage of AD for gene carriers can be made through acceleration on the rate of longitudinal change for disease markers at some time point prior to the onset, which subsequently leads to an accelerated risk of developing AD. This conceptualization is consistent with the clinical and biological evidence of AD which indicated a preclinical stage of the disease during which no clinical disease symptoms were present but neuropathological changes of AD, notably senile plaques and neurofibrillary tangles were accumulated (Tomlinson et al. 1968, Troncoso et al. 1996, Hulette et al. 1998, Haroutunian et al. 1998, 2000, Schmitt et al. 2000, Knopman et al. 2003). The time point when the longitudinal rate of change accelerates on disease markers likely will coincide with the time point that pathological changes begin during the preclinical period. Statistically, the conceptualization of a changepoint in the longitudinal rate of change leads to the rather novel longitudinal statistical changepoint models (Xiong et al. 2003, Ji, Xiong & Grundman 2003, Hall et al. 2000) which have not been fully developed in the statistical literature.

C. PRELIMINARY DATA

The leader of the Biostatistics Core, Dr. Chengjie Xiong, has a tracking record of extensive publications on statistical methodologies and their medical and biological applications, especially in the areas of longitudinal statistical applications in AD research (please see the attached Biosketch). Since 2001, Dr. Xiong has primarily worked for the statistical designs and analyses of major research projects conducted at the Washington University (WU) ADRC and two closely associated program projects entitled *Healthy Aging and Senile Dementia* (HASD, P01AG03991) and *Antecedent Biomarkers for AD: Adult Children Study* (ACS, P01AG026276). Dr. Xiong has demonstrated potential leadership skills by serving as the leader of the Biostatistics component for the program projects *ACS*. The extensive interactions with clinicians/investigators have helped Dr. Xiong understand the clinical, behavioral, cognitive, and biomedical correlates of very mild and mild AD and healthy aging, resulting in many joint publications on AD with WU ADRC investigators. Dr. Xiong also served as an associate editor of the *Journal of Alzheimer's Disease* in 2002 and the Associate Editor of Biostatistics for the journal *Alzheimer's Disease & Associated Disorders* from 2005 to 2007. In 2006, Dr. Xiong was granted a K25 award from NIA to assess the quality of statistical applications, especially the longitudinal statistical applications, in publications on AD. During his K25 training period, Dr. Xiong has worked to further obtain leadership skills by assembling a very diverse research team for his K25 project and other ongoing research projects. He led the research effort and interacted well with established team members such as Dr. John Morris from WU ADRC, Dr. Lenore Launer from the National Institute on Aging, and Dr. Gerald van Belle from University of Washington in Seattle. These leadership skills and interactions have resulted in 2 manuscripts for his K25 project so far (one published by *Neuroepidemiology* and the other under review). Because his K25 will finish on June 30, 2008, it will allow him time to assume the responsibility of DIAN Biostatistics Core. Dr. Xiong's unique training as a statistician with expertise in longitudinal statistical applications in AD, as well as his training in and commitment to AD, make him an ideal leader for the Core.

Mr. Scott Fague will serve as the database manager of DIAN Biostatistics Core.

Dr. John Rice is a well known genetic epidemiologist with extensive experience in family and genetic studies. Dr. Rice will serve as a consultant to the Biostatistics Core as needed (please see his support letter).

The Biostatistics Core personnel are experienced in standard longitudinal statistical models and have long been interested in the statistical models of changepoints, especially for the longitudinal studies. Realizing the nonlinear feature of cognitive decline over time at different stages of AD, Xiong et al. (2003) developed a bi-linear random coefficient model linked at the changepoints of adjacent dementia severity as staged by the Clinical Dementia Rating (CDR) (Morris, 1993) to describe and test the longitudinal cognitive progression from one stage of dementia to the next stage of impairment. This type of longitudinal piecewise linear models are natural extensions of the more general linear random coefficients model of Laird & Ware (1982) and have been further described in Xiong et al. (2007, please see Appendix). Ji, Xiong & Grundman (2003) further proposed another longitudinal statistical model to detect an unknown changepoint in the longitudinal progression from nondemented aging to AD. Other statistical changepoint models associated with survival distributions and accelerated life testing as well as stochastic orderings have also been developed by the Biostatistics Core personnel (Xiong et al. 1999, 2000, 2002). These statistical models contain the key elements that can be used to further develop the longitudinal statistical changepoint models needed to address the preclinical hypotheses of AD in DIAN. Although the previous work in the longitudinal changepoint models either assumed known changepoints (i.e., the conversion times of CDR (Xiong et al. 2003)) or was restricted to more balanced longitudinal designs (Ji, Xiong & Grundman 2003), they nonetheless point to the Biostatistics Core personnel's interest and ability to develop novel longitudinal statistical models and serve as the foundation for our future effort to fully extend these statistical methodologies to address the preclinical hypotheses of AD in DIAN.

D. METHODS

The Biostatistics Core has played a major role in the statistical design of DIAN and reviewed all Cores that are proposed for funding as part of the DIAN. In these reviews, particular attention was paid as to whether the project could be strengthened by design changes and whether the proposed sample sizes and the statistical methodologies were appropriate. These reviews were further augmented with analyses on current age and their parents' age at onset (AAO) for potential DIAN participants based on data obtained from a survey by the Core to all seven participating sites. The Biostatistics Core will provide another review just prior to the implementation of DIAN for additional possible changes dictated by our accumulating knowledge. The collaborations between the Biostatistics Core and other Cores will then be carried forward during the data gathering phase of DIAN. Periodic review of the accumulating data will be made to safeguard the statistical quality of the data and to ascertain whether they conform to the expectations of the original statistical design. These collaborations will be nurtured during the entire funding period of DIAN with ongoing interactions and data analyses, leading to quality scientific publications. Overall, the Biostatistics Core will facilitate the investigators with a seamless transition from the database stored and managed by the Informatics Core (Core H) to the appropriate statistical analyses as well as lead the data analyses activities by providing and developing longitudinal statistical models as dictated by the individual scientific hypotheses of DIAN.

D1. *Statistical Quality Control of DIAN Data*

The undertaking of the multicentre DIAN project implies the need to work with a heterogeneous research team and yet at the same time attain a common goal by following a homogeneous methodology. This demands an additional effort on the statistical data quality control. Whereas the Informatics Core and the ADCS Coordinating Center (the latter for clinical/cognitive data only) will implement standard database quality control procedures, the Biostatistics Core will oversee the *statistical* quality control of entire data before formal statistical data analyses are conducted. These *statistical* quality control procedures are crucial for choosing the appropriate statistical models and for safeguarding the validity of statistical inferences from these analyses.

1. Outlier detection and comparison among sites

Although there has been considerable debate in statistical literature regarding what to do with extreme or influential data points, detected outliers are often candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimates, inflated error rates, and incorrect statistical inferences (Osborne & Overbay 2004, Zimmerman 1994). It is therefore important to identify and examine them prior to modeling and analyses of DIAN data. On a regular basis (i.e., once every quarter), the Biostatistics Core will follow a two-step procedure to implement the statistical assessment of outliers:

First Step (outlier detection): The first component is the selection of outliers via appropriate statistical algorithms which the Biostatistics Core will develop during the DIAN funding period. These algorithms will be based on the standard approaches to the detection of outliers in exploratory data analyses (Barnett & Lewis 1985, Tukey 1977). Because of the large number of disease markers to be assessed in DIAN, both univariate and multivariate methods will be used in the detection of outliers (Rousseeuw & Van Zomeren 2000; Garrett 1989). Because of the longitudinal follow-up on DIAN subjects, we will also examine the potential outliers with respect to the rate of longitudinal changes based on longitudinal statistical models (Diggle et al. 2002).

Second Step (data acceptability review): When the statistical outliers are detected, the Biostatistics Core will collaborate with the Informatics Core to contact the individual study sites to request for the human reviews and examinations of data designated as statistical outliers. This is a manual stage of quality assurance in terms of data acceptability into the database. This decision relies upon known factors that might have influenced the sites' data. The outcome of these procedures will be the identification of whether the detected statistical outliers are valid or invalid observations.

The amount of outliers will also be compared among the study sites so that potential problems in data collection can be detected early in individual study sites.

It is very important to note that statistical outliers selected by the above procedures do not automatically become invalid or bad data. A large deviation from the norm does not necessarily imply that reported data are faulty or not usable. Outliers become unacceptable or not usable only when they are judged (by the study sites) as being too excessive, unreasonable, and not well-founded to accept them as a part of the disease marker data. When outliers are confirmed valid observations for disease markers, they carry important information about the disease progression, and therefore appropriate statistical analyses (e.g., more robust methods) will be applied to the entire data set including these valid outliers. An investigation will be made concerning the sensitivity of the results of analysis to explore the influence of outliers. One analysis with the outliers and at least another eliminating/reducing the outliers will be performed and the results compared.

2. Missing data detection and comparison among sites

Missing data represent a potential source of bias in DIAN. The Missing Data Mechanism (MDM) can have important implications on whether the subsequent statistical analyses produce valid statistical inferences (Little & Rubin 1987). Hence, every effort will be undertaken to fulfill all the requirements of the DIAN protocol concerning the collection and management of data which has been fully described in the Clinical Core and Informatics Core applications. In reality, however, there will almost always be missing data. Statistical assessments of missing data prior to the final analyses are therefore important to understand the MDM and safeguard the validity of the subsequent statistical analyses. On a regular basis (i.e., once every quarter), the Biostatistics Core will follow a two-step procedure to implement the statistical assessments of missing data:

First Step (missing data detection): The first step will identify subjects with missing data from the DIAN database. Because of the large number of disease markers to be assessed in DIAN, missing data patterns will be assessed (Little & Rubin 1987) at each planned assessment visit. Because of the longitudinal follow-up on DIAN subjects, the Biostatistics Core will pay special attention to these subjects who are likely to be the early dropouts (i.e., those who have missing data on consecutive visits).

Second Step (missing data review and coding): after identifying subjects with missing data, the Biostatistics Core data manager will collaborate with the Informatics Core to contact the individual study sites to request for the human reviews and examinations of the missing data. The outcome of the review will result in the coding of the missing data based on the reasons why participants cannot complete testing and assessments. The Biostatistics Core will also collaborate with the Informatics Core to create the missing data coding form prior to the start of data collection.

The amount of missing data will also be compared among the study sites on a regular basis (once a quarter) so that potential problems in data collection can be detected early for individual study sites.

Although no universally applicable methods of handling missing values can be recommended, we will assess the missing data patterns and perform statistical tests on missing data mechanism of DIAN (Diggle et al. 2007, Diggle 1989; Ridout 1991; Cochran 1977; Barnard 1963). We will also conduct analyses concerning the sensitivity of the statistical inferences to the methods of handling missing values (Little & Rubin 1987) by comparing the results based on different approaches, especially if the number of missing values is substantial.

D2. ***Creation/Distribution/Sharing of Analysis Data Sets***

Research projects using DIAN data can be developed by any DIAN investigators from any individual study sites. The Informatics Core application has demonstrated how DIAN investigators will be able to access the entire relational database and generate scientific hypotheses to be formally tested. Once an investigator decides to have the data for formal analysis and/or for publication, the investigator submits a request to the Publication Committee of DIAN (chaired by Dr. Richard Mayeux), which approves all requests to ensure that there is no unwitting duplication or competition. The Committee approval also would encourage the investigator to consult with the Biostatistics Core for the analysis dataset and the formal analyses. After the Biostatistics Core receive such requests, the Core data manager reviews the requests and assures that the requestor has consulted with each Core leader and received the permission from the Core leader whose data will be requested. Whereas DIAN database will be stored and managed by the Informatics Core, the transition from the database to statistically analyzable individual data sets and further to final statistical analysis requires appropriate merging of data from various domains, appropriate creation of analysis variables (i.e., family pedigree identification, censoring status), and finally the distribution of final analysis data sets. Along with the investigators, the Biostatistics Core data manager will first help formulate the necessary inclusion and exclusion criteria (i.e., age, mutations, cognitive status) for the specific sample to be used to address the investigators' hypotheses. Then the Biostatistics core data manager will appropriately merge data from different domains from the DIAN database, de-identify subjects in compliance with HIPPA, create all necessary analyses variables that need to be derived based on the existing database, and finally convert data sets into SAS format or other desirable formats and distribute the data set for analyses. After the statistical analyses, the final analysis data sets including these derived variables during the analyses will be sent back to the Informatics Core for further archiving and releasing to the public once the manuscript reporting them has been accepted for publication.

D3. ***Statistical Data Analyses and Consultation***

1. Process of statistical analyses/consultations/reporting

SAS will be used for the statistical data analyses conducted by the Biostatistics Core, although other valid statistical packages such as SPlus and STAT will also be used for specialized statistical analyses. The selection of SAS is based on the fact that it is the standard statistical software in clinical and pharmaceutical research and also achieves several additional goals. First it uses software/data entry procedures that are well supported by both the vendor and a large international user base. Similar procedures are currently used by the Biostatistics Core of WU ADRC and by the WU Division of Biostatistics (where the DIAN Biostatistics Core will be located) for both intramural and multi-institutional studies where the Division serves as the data coordinating center. Second, because the biostatistics staff and many of our collaborating investigators already rely on SAS for the majority of data management and statistical analysis processes, direct maintenance of data in SAS datasets simplifies the entire system. Such datasets are directly transferred to the Wubios Computing Resource within the Division of Biostatistics for subsequent merging with other data and for offsite archival storage. Significant savings of staff training are also achieved by ensuring that all staff are well versed in the SAS system. Because SAS functions on a wide variety of hardware/software platforms, it affords the project an opportunity to execute on the most cost-efficient platform. Finally, The WU Division of Biostatistics already has in place security and backup procedures in order to safeguard the integrity of the study data sets and protect participants' privacy. Because SAS is particularly powerful in reading different types of computer files, the conversion of data sets from the Informatics Core to SAS can be readily achieved in a secure process.

All DIAN investigators will participate in data analysis activities; however, the Biostatistics Core personnel have primary responsibility and will play a leading role in this area. Data analyses may be initiated and requested by any of the DIAN investigators. The Biostatistics Core staffs prepares and distribute the final analysis data sets, and provide advice and consultation to the investigators about appropriate data analysis strategies and statistical models as well as the adequate implementation of these models for the investigators who wish to execute their own analyses. More importantly, the Biostatistics Core will help DIAN investigators in interpreting the outputs from standard statistical packages. Investigators who do their own analyses are always encouraged to consult with Biostatistics Core staff or to have Core staff conduct particular analyses. In all cases, the Biostatistics Core will lead the activities of statistical data analyses and safeguard the validity of the statistical inferences. The final step of statistical data analyses is reporting the statistical analyses in scientific publications. For these analyses done by the Biostatistics Core, the Core will be responsible for writing the section of statistical analyses of the manuscripts. If investigators choose to do their own analyses, the Core will provide statistical review of the resulting manuscripts before they are submitted for publications. Finally, the director of the Biostatistics Core serves on the Data Sharing & Publications Committee of DIAN.

2. Standard longitudinal statistical methods

The unique analytic features of DIAN are two major correlational structures that need to be incorporated into any statistical models of DIAN data: the first is the within-subject correlational structure over the repeated measurements of disease markers for the same individuals over time, and the other is the within-family correlational structure for the disease markers and AAO from individuals within the same families.

The entire statistical analyses of DIAN can be conceptualized as containing two major parts: the standard analyses which are readily implemented in SAS as well as in other valid statistical packages, and the more novel statistical models targeting at the specific scientific hypotheses about the preclinical changes of AD that need to be further developed and implemented during the funding period. For the first part, a wide variety of standard statistical procedures will be used in analyses of the expected data from DIAN. In some cases, simple descriptive statistics of the distributions of various disease markers within defined clinical and genetic groups or various correlational analyses will suffice. In others, the relevant scientific questions are better addressed with general linear mixed models and generalized linear models (Diggle et al. 2002, Xiong et al. 2007—please see Appendix). General and generalized linear mixed models are the most frequently used statistical methodologies to analyze longitudinal data. These models have the capacity to recognize both the within-family correlation by introducing a random effect of family and the correlational structure from the repeated measurements of the disease markers for the same participants over time in DIAN. General linear mixed models are built on either explicit parametric models of the covariance structure of repeated measures over time whose validity can be checked against the available data or, where possible, use methods of inference that are robust to misspecification of the covariance structure. In many cases, especially when the sample size is relatively small or moderate and there are many covariate variables, a parametric structure must be imposed on the covariance matrix of repeated measurements over time. Many different types of covariance structures have been used in the general linear mixed models, but there are essentially two traditional ways to build a structure into a covariance matrix: one uses serial correlation models and the other uses random effects. Weighted least squares estimation and the maximum likelihood (ML) or restricted maximum likelihood methods (REML) through the EM algorithm (Patterson &Thompson, 1971; Cullis & McGilchrist, 1990; Verbyla & Cullis, 1990; Tunnicliffe-Wilson, 1989; Dempster, Laird, & Rubin, 1977) are used to estimate the mean response and the covariance parameters. In a semiparametric approach, a robust estimator to the variance of estimated mean response (the sandwich variance estimator, Liang & Zeger, 1986) can be used to make inferences about the mean response, as it is not sensitive to the choice of the weight matrix in the weighted least squares estimation (Liang & Zeger, 1986; Diggle et al. 2002).

Another possible way of introducing a correlational structure on repeated measurements within the same individuals as well as the correlational structure of measurements among different individuals but within the same family is through the two-stage random effects models. When participants are sampled from participating sites of DIAN, various aspects of their behavior may show stochastic variation on the measurements of disease markers between individuals. The simplest example of this is when the general level of the response profile varies between participants, that is, some participants are intrinsically high responders, others low responders. The two-stage random effect model (Diggle 1988; Laird & Ware, 1982; Vonesh & Carter, 1992) allows the individual response profile or 'growth curve' for each participant within each family at the first stage, which can also account for the likely family effect as well as the possible correlations for measurements of individuals within the same families by the introduction of a random effect of family. The second stage of the two-stage random effects model introduces the between-subjects variation of the subject-specific effects and the population of the subject-specific effects. The entire process leads to the development of the general linear mixed models. The maximum likelihood estimates, the restricted maximum likelihood estimates, and the method-of-moment estimators can be used to estimate the regression. In addition, the general linear mixed models not only provide the Best Linear Unbiased Estimator (BLUE) (Graybill 1976) for any estimable contrast of the regression parameters but also estimate the subject-specific effects arising in multiple populations through the Best Linear Unbiased Predictor (BLUP) (Harville1977). Nonlinear mixed effects models will also be used to model longitudinal data of DIAN when necessary (Lindstrom & Bates 1990).

The generalized linear mixed models for longitudinal data extend the techniques of general linear models. They are suited specifically for nonlinear models with binary or discrete data, such as logistic regression, in which the mean response is linked to the explanatory variables through a nonlinear link function (Liang &

Zeger, 1986; Zeger & Liang, 1986). The marginal models permit separate modeling of the regression of the response on explanatory variables and the association among repeated observations of the response for each participant. They are appropriate when inferences about the population averages are the focus of DIAN. The techniques of Generalized Estimating Equations (GEE) can be used to estimate the regression parameters (Liang & Zeger, 1986; Pretince 1988; Zhao & Prentice 1990; Liang, Zeger, & Qaquish 1992; Fitzmaurice, Laird, & Rotnitsky 1993). The approach of random effects models in the set-up of generalized linear mixed model allows heterogeneity among subjects in a subset of the entire set of the regression parameters. Two general approaches of the estimation can be used. One is to find the marginal means and covariance of the response vector and then apply the technique of GEE (Zeger & Qaqish 1988; Goldstein 1991; Breslow & Clayton 1993). The other is the likelihood approach (Anderson & Aitken 1985; Hinde 1982) or the Penalized Quasi-Likelihood (PQL) approach (McGilchrist & Aisbett 1991; Breslow & Clayton 1993). Another generalized linear model is the transition model for which the conditional distribution of the response at a time given the history depends only on the prior observations with a specified order through a Markov chain. Full maximum likelihood estimation can be used to fit the Gaussian autoregressive models (Tsay 1984), and the conditional maximum likelihood estimation can be used to fit logistic and log-linear models (Zeger, Liang, & Self 1985; Zeger & Qaqish 1988).

One of the most important endpoints in DIAN is the age at onset (AAO) of AD, which is traditionally analyzed by the techniques of survival analyses (Kalbfleisch & Prentice 1980). A basic feature of such survival data is that we almost never observe the clinical outcome (i.e., AAO) in all subjects, making the unobserved AAO statistically censored. The regression method introduced by Cox (1972) will be used to investigate the effects of several variables on survival at the same time. Cox's proportional hazards model is a semiparametric approach—no particular type of distribution is assumed for the survival data, but a strong assumption is made about the effects of differences. Regression diagnostic procedures will be used to assess the assumption of proportional hazards (Schoenfeld 1982; Therneau, Grambsch, & Fleming 1990), and tests on the assumption of proportional hazards will also be introduced through the incorporation of time-dependent covariates (Therneau & Grambsch 2000). Extensions to the Cox proportional hazards model are the analysis of residuals, time-dependent coefficient, multiple/correlated observations, time-dependent strata, and estimation of underlying hazard function (Therneau & Grambsch 2000; Fleming & Harrington, 1991; Anderson & Gill, 1982). Because multiple AAOs from members of the same family tend to be correlated, Cox models with frailty effects (Therneau & Grambsch 2000) will be used to analyze such clustered AAO data. Classical ways to fit frailty models are likelihood based. The shared frailty model provides an appropriate way to describe the within family dependence of AAO data. Likelihood methods to fit shared frailty models include: EM-algorithm (Klein 1992), penalized partial likelihood (Therneau and Grambsch 2000; McGilchrist 1993), Bayesian analysis (Ducrocq and Casella 1996). Different frailty distributions (i.e., Gamma, Gaussian, log-normal) will be fitted and inferential results compared. In recent papers more complex frailty models have been studied. Within the context of DIAN, we will apply frailty models with a random family effect and a random gene mutation effect. To fit such frailty models, the likelihood based methods mentioned above will be adapted to cover this extra complexity in the data: EM algorithm (Vaida and Xu 2000; Cortinas and Burzykowski 2005), penalized partial likelihood (Ripatti and Palmgren 2000), and Bayesian approaches (Legrand et al. 2005).

Because both repeatedly measured disease markers (CSF molecular markers, neuroimaging markers, and cognitive markers) and the AAO are collected through the longitudinal follow-up of DIAN subjects, we will also use the state-of-the-art statistical joint modeling methodologies developed from the standard general linear mixed models (Diggle at al. 2002) for longitudinal data on disease markers and the standard proportional hazards models for the AAO data (Therneau & Grambsch, 2000). More specifically, AAO will be jointly modeled with the repeatedly measured disease markers through a common set of latent random variables to derive the association between the longitudinal rates of changes on these markers and the hazards of developing AD. These common latent random variables include the longitudinal rate of change for disease markers assessed in DIAN. Maximum likelihood based approach (Tsiatis & Davidian 2004, Wang & Taylor 2001, Henderson et al. 2000, Xu & Zeger 2001) and an alternative conditional score approach (Stefanski & Carroll 1987, Tsiatis & Davidian 2001), as well as a semiparametric likelihood approach (Song et al 2002a) by relaxing on the normality assumption on shared latent variables will be used. These analyses will include the potential confounding covariates (i.e., gender, APOE4, baseline age, education), as well as their possible interactions among themselves and with the rate of longitudinal change from the disease markers. The inclusion of these potential covariates and interactions serves two major purposes. First, it will allow us to test the effects of these confounding factors on the risk of developing AD. Second, it will enable us to test the risk effect of longitudinal rates of change for disease markers with appropriate adjustment for the effects of the confounding factors. We will also jointly model a multivariate longitudinal process including multiple disease markers along with the AAO. This type of multivariate joint models has been recently developed and implemented by Song et al. (2002b). All these proposed analyses will be implemented in SAS macros along with PROC MIXED/SAS (Littell et al. 1996) and PROC PHREG/SAS (SAS 1990).

In the use of each of these statistical models, a variety of technical assumptions are required. Diagnostic techniques to check the appropriateness of these assumptions for a given dataset will be implemented. Particularly for the primary analyses, we will assure ourselves that the reported results are not simply artifacts of a particular method of analysis. Repeating the analyses with a variety of analytic techniques and comparing the results is one approach to insure against such mistakes and will also be used here. These sensitivity analyses will be especially important in the statistical analyses of outliers and missing data (please see Section D1). Graphical techniques will be used to understand multivariate relationships. The Wubios Computing Resource has a variety of hardware and software to support graphics generated with SAS/Graph, SPlus and various graphics editors. Both PC and Linux based systems allow the generation of dynamic graphics for multivariate data displays.

It should be emphasized that the model we have implemented in the Biostatistics Core is that these analyses are not undertaken in isolation by the Core investigators but rather represent a collaborative undertaking with the clinical investigators interacting at all stages of the analyses. We recognize that DIAN will measure a large number of potential variables for analysis and testing multiple scientific hypotheses, thus creating the possibility that differences will be “significant” by chance alone. Because of the parallels between comparisons being examined in DIAN and studies previously conducted in sporadic AD studies, we can focus the proposed analyses on those relationships that have been previously demonstrated in AD literature. This replication of prior results provides the best protection against the problems sometimes associated with multiple comparisons. Where a large number of comparisons are still required we will use appropriate techniques (e.g., the global F-test or sometime approximate F-test in general and generalized linear mixed models, and the global Chi-squared test in a Cox’s proportional hazards model) or simple Bonferroni corrections for these multiple comparisons in order to control the Type I error rate.

3. Development of novel longitudinal statistical models addressing the preclinical hypotheses of DIAN

The test of preclinical hypotheses of DIAN require some novel longitudinal statistical methodologies. The first primary hypothesis in DIAN states that there is a period of preclinical (presymptomatic) AD in individuals who are destined to develop early onset dementia (gene carriers) that can be detected by changes in biological fluids and in neuroimaging correlates in comparison with individuals who will not develop early onset dementia (noncarriers). We conceptualize the existence of a preclinical stage of AD through acceleration on the rate of longitudinal change of disease markers at some time point during the preclinical period, which subsequently leads to an accelerated risk of developing AD.

Statistically, the conceptualization of a changepoint in the longitudinal rate of change for disease markers for gene carriers leads to the longitudinal statistical changepoint models. These models will be developed and statistically implemented in SAS during the DIAN funding period. Our approach will be based on our previous publications on the statistical models of changepoints, especially in the longitudinal studies (Xiong et al. 2003, Xiong et al. 2007, Ji, Xiong & Grundman 2003). More specifically, we use $Y_i(t_j)$ to represent a disease maker measurement for the *i*-th subject in the group of gene carriers at age t_j before AAO. Here we propose a changepoint τ to model the acceleration on the rate of change for the marker, i.e., the rate of change for the disease marker is β_{1j} before τ, whereas the rate of change is $\beta_{1j}+\beta_{2j}$ after τ. Denote the acceleration indicator $A(t)=0$ before τ, and $A(t)=1$ after τ. Let e_{ij} be the within-subject random errors. Then the changepoint model becomes $Y_i(t_j)=\beta_{0j}+\beta_{1j}t_j+\beta_{2j}t_jA(t_j)+e_{ij}$. The standard general linear mixed model further assumes a 3-dimentional normal distribution for the vector $(\beta_{0j},\beta_{1j},\beta_{2j})$ with the mean $(\beta_0,\beta_1,\beta_2)$ and an unstructured covariance matrix. If τ is assumed known, this model can be readily implemented in SAS through the maximum likelihood approach. Let $(\hat{\beta}_{0\tau},\hat{\beta}_{1\tau},\hat{\beta}_{2\tau},\hat{\theta}_\tau)$ be the maximum likelihood estimator of $(\beta_0,\beta_1,\beta_2,\theta)$ when the changepoint is assumed at τ, where θ is the set of variance/covariance parameters. We apply the general statistical procedure we developed for the test and detection of a changepoint in Xiong & Milliken (2000) and Xiong & EL Barmi (2002). First, the hypothesis on the existence of a changepoint τ is tested by $LL = \max_\tau [2PL(\hat{\beta}_{0\tau}, \hat{\beta}_{1\tau}, \hat{\beta}_{2\tau}, \hat{\theta}_\tau) - 2PL_0(\hat{\beta}_0, \hat{\beta}_1, \hat{\theta})]$, where τ takes all possible values of the changepoint, and $PL(\hat{\beta}_{0\tau}, \hat{\beta}_{1\tau}, \hat{\beta}_{2\tau}, \hat{\theta}_\tau)$ and $PL_0(\hat{\beta}_0, \hat{\beta}_1, \hat{\theta})$ are the maximum log likelihood when the rate acceleration is at τ and when there is no rate acceleration (i.e., $\beta_2 = 0$), respectively. The distribution of *LL* under null hypothesis of no changepoint will have no close analytic form but can be assessed by the method of bootstrapping (Efron & Tibshirani 1994, Beran 1988, Hall 1992) through a computer-intensive large process of resampling, which then facilitates a statistical test on the existence of a changepoint on longitudinal rate. Similar to what was proposed in Xiong & Milliken (2000) and Xiong & EL Barmi (2002), if the existence of a rate acceleration is statistically confirmed, we propose to estimate the true changepoint τ for the gene carriers by $\hat{\tau}$ such that $PL(\hat{\beta}_{0\hat{\tau}}, \hat{\beta}_{1\hat{\tau}}, \hat{\beta}_{2\hat{\tau}}, \hat{\theta}_{\hat{\tau}}) = \max_\tau PL(\hat{\beta}_{0\tau}, \hat{\beta}_{1\tau}, \hat{\beta}_{2\tau}, \hat{\theta}_\tau)$. For these who will not develop early onset dementia (noncarriers), on the other hand, we anticipate a statistical confirmation that no rate acceleration exists based on the above proposed analytic procedures.

We will also develop longitudinal statistical models that allow the joint modeling of AAO data and the repeatedly measured disease markers incorporating a changepoint for the rate acceleration. These models will be based on the current work of joint modeling (Tsiatis & Davidian 2004), but the addition of an unknown changepoint creates many new analytic challenges that need to be addressed. Our plan is to develop these novel longitudinal models and implement them during the DIAN funding period before the final analyses.

The second primary hypothesis states a temporal difference on the preclinical changes among different disease markers for gene carriers. More specifically, it is hypothesized that the sequence of preclinical changes initially will involve CSF amyloid-beta-42 (Aβ_{42}) associated with the process of production and clearance, followed by evidence through amyloid imaging for cerebral deposition of Aβ_{42}, followed by cerebral metabolic activity (functional imaging), and finally by regional atrophy (structural imaging). Statistically, this hypothesis can be conceptualized by the temporal ordering among different disease markers on the changepoints when the longitudinal rate of change accelerates. To test this hypothesis, the changepoint models discussed above will be applied to each disease marker as identified in the hypothesis, and estimates to marker-specific changepoints (i.e., acceleration points) of rate will be obtained for gene carriers. Again there will be no close form for the joint distribution of estimated changepoints among these markers. To assess whether the changepoints are the same across different disease markers, computer-intensive resampling methods (Efron & Tibshirani 1994) will again be used to assess the joint distribution of the estimated changepoints which then subsequently leads to 95% confidence interval estimates to the differences on the changepoints between different disease markers.

The third hypothesis of DIAN states that the phenotype of symptomatic early-onset familial AD, including its clinical course, is similar to that of late-onset “sporadic” AD. The statistical testing of this hypothesis involves the comparisons on the longitudinal course between these with early onset familial AD from DIAN and those with “sporadic” AD which are not included in DIAN. We will choose adequate samples from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and from WU ADRC longitudinal database for the purpose of comparison. More specifically, because it has been well established that the baseline severity of dementia is associated with the rate of subsequent longitudinal changes (Storandt et al. 2002), we will identify all sporadic AD cases from ADNI and WU ADRC whose baseline CDR matches with these cases in DIAN. Because WU ADRC is a participating site of ADNI, we will make sure that cases from ADRC are not repeated twice in the comparison group of sporadic AD. Because ADNI and DIAN follow very similar study protocols, essentially all disease markers from both studies will be longitudinally compared. When the combined samples from both ADNI and WU ADRC are used as the comparison group of sporadic AD, the statistical comparisons will be done only on the disease markers commonly and longitudinally assessed by DIAN, ADNI and WU ADRC. The statistical comparison will be carried out by the standard general linear mixed models, especially the random intercept and random slope models of Ware and Laird (1982) and these from Xiong et al. (2007—please see Appendix). Family will again be treated as a random effect in these models. The subject-specific rate of longitudinal change on disease markers will be modeled as a function of type of AD (i.e., early-onset familial AD vs. sporadic AD) and baseline CDR, as well as other covariates such as education and APOE4 genotype. The mean rate of longitudinal change on disease markers between these with early-onset familial AD and those with sporadic AD will be compared within each CDR group after adjusting for the effect of potential covariates by computing a 95% confidence interval (CI) for the difference. These analyses will be implemented by PROC MIXED/SAS (Littell et al. 1996).

D4. Power analyses

The scientific hypotheses of DIAN are based on the clinicopathologic evidence from AD literature that points to the existence of a preclinical stage of the disease and further motivated by our preliminary analyses on a subset of the longitudinal WU ADRC database. We identified a total of 37 subjects from the WU ADRC database who were clinically nondemented through life but nonetheless with substantial AD neuropathology at brain autopsy which led to the neuropathological diagnosis of AD. These subjects were confirmed clinically nondemented based on their expiration Clinical Dementia Rating (CDR) (Morris 1993) of 0 which was obtained by telephone interviews with a reliable collateral source conducted within weeks after the death. These subjects were therefore truly at the preclinical stage of AD before death. The mean age at the last clinical visit is 83.33 y with a standard deviation (SD) of 7.83y. The number of annual clinical/cognitive follow-ups ranged from 2 to 19, and the mean number of years from the last clinical visit to death is 2.12 y (SD=3.25). The careful examinations of the longitudinal cognitive data indicated that the average growth curve pattern on four well established composite cognitive factor scores (Rubin et al. 1998, Kanne et al. 1998) before death was nonlinear, with acceleration on the rate of decline approximately 7.5 years before death. A general linear mixed effects model was then fitted to the data to estimate the mean rate of cognitive decline of these subjects both before and within 7.5 y of death. The estimated slope for the primary factor (Rubin et al. 1998) 7.5 years before death is -0.0328 per year, whereas that within 7.5 years of death is -0.0503, resulting in an acceleration of 0.0165 per year. The standard error for the estimated acceleration on the rate of decline is 0.0159 per year. Based on these statistics, a total sample of 192 presymptomatic subjects (with 96 each for gene carriers and noncarriers) provides at least 80% power to detect an acceleration of 0.0396 per year on the rate of cognitive decline (based on the primary factor of Rubin et al. 1998) for the gene carriers in comparison to the noncarriers who are expected to experience no acceleration (i.e., mean acceleration=0) over the study period. This sample size computation is based on a t-test at a significance level of 5%. Similar effect sizes based on other cognitive factor scores (i.e., the temporal factor, the frontal factor, and the parietal factor of Kanne et al. 1998) can also be detected with this sample size. These projected effect sizes and the power are further confirmed with a large simulation through longitudinal changepoint models (Xiong et al. 2007—please see Appendix) which were run over multiple choices of possible changepoints. Because these projected effect sizes are much smaller than those reported between CDR 0 and incipient AD (Storandt et al. 2002), DIAN is reasonably powered for testing/comparing the preclinical cognitive changes. Further, for the presymptomatic gene carriers alone, a sample of size 96 provides at least 80% power to detect an acceleration of as small as 0.028 per year on the rate of cognitive decline (based on the primary factor of Rubin et al. 1998). This sample size computation is based on a paired sample t-test at a significance level of 5%.

We further assessed the statistical power on the test of longitudinal atrophy rates and CSF biomarker changes through similar power analyses. Godbolt et al. (2005) and Fox et al. (2000) reported accelerated rate of volumetric changes in subjects with familial AD. Data from our own WU ADRC further suggested that the mean whole brain atrophy rate is -0.45% per year for subjects with normal aging and -0.98% per year for subjects with AD (Fotenos et al. 2005). Assuming a standard deviation of 1.1% per year for the whole brain atrophy rate (Fox et al. 2000, Fotenos et al 2005) and a Pearson correlation of 0.5 between the atrophy rates before and after the acceleration point, a sample of 192 presymptomatic subjects provides at least 80% power to detect an acceleration of 0.46% per year on the whole brain atrophy rate for the gene carriers in comparison to the noncarriers who are assumed to experience no acceleration (i.e., mean acceleration=0). Because this projected effect size is smaller than that reported between CDR 0 and early stage AD (Fotenos et al. 2005), DIAN is reasonably powered for testing/comparing the preclinical atrophy changes. For the presymptomatic gene carriers alone, we estimate that a sample of size 96 provides at least 80% power to detect an acceleration of as small as 0.32% per year on the rate of atrophy change. Longitudinal rates of CSF biomarker changes were rarely reported in the literature, and we were unable to obtain the degree of variability (i.e., SD) on them (which highlights the need of DIAN). Assuming a Pearson correlation of 0.5 between the rate of change in CSF $A\beta_{42}$ before and after the acceleration point, a sample of 192 presymptomatic subjects provides at least 80% power to detect an acceleration of 0.41 SDs per year on the rate of $A\beta_{42}$ change for the gene carriers in comparison to the noncarriers who are assumed to experience no acceleration.

D5. *Interactions/Collaborations with other Cores*

The entire DIAN Biostatistics Core personnel are also members of the Biostatistics Core of WU ADRC who meet weekly to discuss center wide progress and biostatistical applications/issues among research projects with representatives of the WU ADRC Clinical, Psychometric, and Administrative Cores who are also the members of relevant Cores of DIAN. This weekly meeting will be extended to discuss new research proposals and data requests/analyses associated with DIAN as well as report on work in progress. DIAN Biostatistics

Core staff will also regularly attend the weekly clinical core staff meetings and DIAN investigators meetings to present the current status of the Core. In addition, the Biostatistics Core staff will attend the weekly WU ADRC research seminars which are coordinated by the WU ADRC Education Core. The Biostatistics Core director will serve in the Steering Committee of DIAN to oversee the biostatistical applications for the entire project.

Core A: Administrative Core (JC Morris)

The Biostatistics Core will provide descriptive statistical analyses of DIAN data for DIAN Investigators Meetings and Steering Committee Meetings concerning DIAN goals and progress.

Core B: Clinical Core (RJ Bateman)

The ongoing extensive collaboration and interactions between WU ADRC Biostatistics Core and Clinical Core (which include the entire staff of DIAN Biostatistics Core and Clinical Core, respectively) are evident (please see the joint publications from Dr. Chengjie Xiong' biosketch) and will be carried forward to DIAN. The Biostatistics Core staff will meet weekly with Clinical Core representatives to discuss DIAN progress and biostatistical applications/issues for research projects using DIAN clinical and psychometric data. The Core staff will lead in the analyses of the data from the Clinical Core which addresses the main theme of DIAN---examining the preclinical (presymptomatic) changes in individuals who are destined to develop early onset dementia (gene carriers) in comparison to individuals who will not develop early onset dementia (noncarriers).

Core D: Neuropathology (NJ Cairns)

Although the number of expected brain autopsies in DIAN might be limited, the Biostatistics Core will help provide descriptive statistics including diagnostic/quantitative information for all autopsies and the reporting.

Core E: Biomarker (AM Fagan)

Extensive collaboration and interaction have already been ongoing among the Biostatistics Core and Biomarkers Core staff for many projects within WU ADRC. These collaborations are evidenced by several joint publications between staff from these two Cores (please see the joint publications from the Core directors, Dr. Xiong and Dr. Fagan's biosketches). These existing collaborations will be continued in DIAN.

Core F: Genetics (AM Goate)

The Biostatistics Core will lead in the effort of data analyses associated with genetic data. Different gene mutations will be examined as to whether they have different longitudinal growth patterns both before and after the disease onset. The effect of APOE4 genotype will also be examined. Because DIAN is a family study, the Biostatistics Core will collaborate with the Genetic Core to propose appropriate statistical methods (i.e., those allowing the correlational structure for disease marker measures from the same family).

Core G: Imaging (MA Mintun)

The Biostatistics Core have already interacted well with the Imaging Core in several research projects within WU ADRC. We will continue these interactions in DIAN. Due to the rather large number of regions of interest (ROI), the primary hypotheses on imaging variables will be tested with an *a priori* variable set (Please see Imaging Core (Core G) application). We will also explore the use of factor analyses techniques to seek composite factor scores that capture most of the variation in preclinical atrophy, metabolism and Aβ plaque measures over multiple ROIs for the tests of primary hypotheses of DIAN.

Core H: Informatics (D Marcus)

We have already outlined in this application the extensive interactions and collaborations with the Informatics Core about the statistical quality control of DIAN data, more specifically in the detection and handling of statistical outliers for multimodal disease markers as well as in the assessment of missing data (please see Section D1), and in the creation/distribution of final analysis data sets (please see Section D2). The Biostatistics Core's interaction with Informatics Core will be on a regular basis so that a smooth transition can be bridged from the DIAN database to the analyses of the longitudinal data within, between, and among the various data domains. The Core leaders (Dr. Xiong & Dr. Marcus) have already met several times to specifically discuss the interactions on database and statistical analyses issues for DIAN. This high degree of interactions will be carried forward. In fact, both Core leaders will meet monthly to discuss interactions between two Cores and other outstanding issues. The two Cores will jointly provide DIAN Steering Committee regular reports on the progress of DIAN (enrollment, demographics, etc, from the Informatics Core) and the results from regular statistical quality control of DIAN data (i.e., outliers, missing data).

D5. *Advising Other Researchers Using the DIAN Database for Additional Analyses*

The Biostatistics Core will also be the first contact for other researchers interested in using the DIAN database for additional analyses, and serve as an advisory group for these investigators in their data set preparation and statistical analyses.

E. Human Subjects

Since the Biostatistics Core (Core C) does not directly interact with subjects, the descriptions of the characteristics of the participants may be found in the Cores which actually recruit and gather data on subjects. Specifically, please refer to Core B: Clinical for the DIAN-wide Human Subjects summary. The primary human studies concern is that of the confidentiality of the information contained in the database maintained by the Informatics Core (Core H) which will be accessed by the Biostatistics Core staff on a regular basis. All staff within the Division of Biostatistics sign confidentiality agreements. The data analyses will be carried out by the Wubios Computing Resource which is used exclusively for the research agenda of the Division and its collaborators. Neither students nor general faculty and staff have access to the system. All access to the shared use systems are controlled by passwords and restrictions based on the physical location of the workstation accessing the information. The ethernet segment within the Division is physically contained within space under our control and is isolated from outside of the Division by a firewall. The university ethernet backbone is centrally administered and is designed to maintain a high level of security. Video surveillance cameras record access to facilities within the Division. The computer room where the servers are located is always locked.

The Biostatistics Core vigorously enforces the use of coded ID numbers for exchange of data between Cores and projects and among Cores. Names, or other personal identifiers, are never contained in reports generated for use in final data analysis. While the DIAN database does contain names and other personal identifiers in order to support the logistics of the operation of the DIAN and statistical quality control of data, access to this information is strictly limited. An additional level of password security is required for direct information on shared use systems. The information is never stored on disk or in plain text, but rather is always encrypted. Web-based access to such information is restricted based on both passwords and IP access restrictions. A secure-web-server is used where appropriate to provide an additional level of protection. The use of web-based reports will substantially reduce the number of printed reports distributed among the staff with identifiers plainly visible.

F. VERTEBRATE ANIMALS - None

G. SELECT AGENT RESEARCH - None

H. LITERATURE CITED.

Anderson DA, Aitkin M. Variance component models with binary response: interviewer variability. Journal of the Royal Statistical Society, 1985; B 47:203-210.

Anderson PK, Gill RD. Cox's regression model for counting processes: a large sample study. Annals of Stat. 1982; 10:1100-1120.

Barnard GA. Contribution to the Discussion of Professor Bartlett's paper. Journal of the Royal Statistical Society 1963; B 25:294.

Barnett V, Lewis T. Outliers in Statistical Data (John Wiley & Sons, 2d ed., New York, NY, 1985).

Beran R. Prepivoting test statistics: a bootstrap view of asymptotic refinements. Journal of the American Statistical Association 1988;83:687-697.

Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 1993;88:9-25.

Cochran WG. Sampling Techniques. New York: John Wiley, 1977.

Cortinas Abrahantes J, Burzykowski T. A version of the EM algorithm for proportional hazards model with random effects. Biometrical Journal 2005; 47:847-862.

Cox DR. Regression models and life tables (with Discussion). Journal of the Royal Statistical Society 1972;B,74:187-200.

Cullis BR, McGilchrist CA. A model for the analysis of growth data from designed experiments. Biometrics 1990; 46:131-142.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). Journal of the Royal Statistical Society, B, 1977;39:1-38.

Diggle PJ. An approach to the analyses of repeated measures. Biometrics 1988;44:959-971.

Diggle PJ. Testing for random dropouts in repeated measurement data. Biometrics 1989;45:1255-1258.

Diggle PJ, Farewell D, Henderson R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. Appl. Statist. 2007;56(5):499-550.

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL, Analysis of Longitudinal Data, 2nd ed. New York: Oxford University Press, 2002.

Ducrocq V, Casella G. A Bayesian analysis of mixed survival models, Genetics Selection Evolution 1996;28:505-529.

Efron B, Tibshirani R. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1994.

Fitzmaurice GM, Laird NM, Rotnitsky AG. Regression models for discrete longitudinal response (with Discussion). Statistical Science 1993;8:284-309.

Fleming TR, Harrington DP. Counting Processes and Survival Analysis. Wiley, New York, 1991.

Fotenos AF, Snyder AZ, Girton LE, et al. Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology 2005;64:1032-1039.

Fox NC, Cousens S, Scahill R, Harvey RJ, Rossor MN. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer's disease: power calculations and estimates of sample size to detect treatment effects [see comments]. Arch Neurol 2000;57:339-344.

Garrett RG. The chi-square plot: A tool for multivariate outlier recognition. Journal of Geochemical Exploration. 1989; 32:319-341.

Godbolt AK, Cipolotti L, Anderson VM, Archer H, Janssen JC, Price S, Rossor MN, Fox NC. A decade pf pre-diagnostic assessment in a case of familial Alzheimer's disease: tracking progression from asymptomatic to MCI and dementia. Neurocase 2005;11:56-64.

Goldstein H. Nonlinear likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 1991;72:320-340.

Graybill, F.A. Theory and Application of the Linear Model; California: Wadsworth & Brooks/Cole Advanced Books & Software, 1976.

Hall P. The Bootstrap and Edgeworth Expansion, New York, Springer-Verlag, 1992.

Hall CB, Lipton RB, Sliwinski M, et al. A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease, Statistics in Medicine 2000;19:1555-1566.

Harville DA. Maximum Likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 1977;72:320-340.

Henderson R, Diggle P, Dobson A. Joint modeling of longitudinal measurements and event time data. Biostatistics 2000;4:465-480.

Haroutunian V, Perl DP, Purohit DP, Marin D, Khalid K, Lantz M, Davis KL, Mohs RC. Regional distribution of neuritic plaques in the nondemented elderly and subjects with very mild Alzheimer disease. Arch Neurol 1998; 55:1185-1191.

Haroutunian V, Purohit DP, Perl DP, Marin D, Khan K, Lantz M, Davis KL, Mohs RC. Neurofibrillary tangles in nondemented elderly subjects and mild Alzheimer disease. Arch Neurol 1999; 56:713-718.

Hinde J. Compound Poisson regression models. GLIM 82: Proceedings of the International Conference on Generalized Linear Models. Ed. Gilchrist R. Berlin: Springer, 1982.

Hulette CM, Welsh-Bohmer KA, Murray MG, Saunders AM, Mash DC, McIntyre NJ. Neuropathological and neuropsychological changes in "normal" aging: Evidence for preclinical Alzheimer disease in cognitively normal individuals. J Neuropathol Exp Neurol 1998; 57:1168-1174.

Ji M, Xiong C, Grundman M. Use of mixed effects models to evaluate change points during the course of cognitive decline to AD. Journal of the Alzheimer's Disease 2003; 5,5: 375-382.

Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York: John Wiley, 1980.

Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 1992;48:795-806.

Knopman DS, Parisi JE, Salviati A, Floriach-Robert M, Boeve BF, Ivnik RJ, Smith GE, Dickson DW, Johnson KA, Petersen LE, McDonald WC, Braak H, Petersen RC. Neuropathology of cognitively normal elderly. J Neuropathol Exp Neurol 2003; 62:1087-1095.

Laird NM, Ware JH. Random-effects models for longitudinal data, Biometrics 1982; 38: 963-974.

Legrand C, Ducrocq V, Janssen P, Sylvester R, Duchateau L. A Bayesian approach to jointly estimate center and treatment by center heterogeneity in a proportional hazards model. Statistics in Medicine 2005; 24:3789-3804.

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73:13-22.

Liang K-Y, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data (with Discussion). Journal of the Royal Statistical Society 1992; B,54:3-40.

Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. Biometrics

1990;46:673-687.

Littell R, Milliken GA, Stroup W, et al. SAS System for Mixed Models. Cary NC: SAS Institute Inc., 1996.

Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley,1987.

McGilchrist C. REML estimation for survival models with frailty. Biometrics 1993; 49:221-225.

McGilchrist CA, Aisbett CW. Restricted BLUP for mixed linear models. Biometrical Journal 1991;33:131-141.

Morris JC. The clinical dementia rating (CDR): current version and scoring rules, Neurology 1993; 43: 2412-2414.

Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation 2004; 9(6) online at http://pareonline.net/getvn.asp?v=9&n=6

Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika 1971; 58:545-54.

Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988; 44:1033-1048.

Ridout M. Testing for random dropouts in repeated measurement data. Biometrics 1991;47:1617-1621.

Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood, Biometrics 2000; 56:1016-1022.

Rousseeuw PJ, Van Zomeren BC. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association. 2000; 85(411):633-651.

SAS Institute, Inc. SAS/STAT User's Guide, (Version 6), Cary, NC, 1990.

Schmitt FA, Davis DG, Wekstein DR, Smith CD, Ashford JW, Markesbery WR. Preclinical AD revisited. Neuropathology of cognitively normal older adults. Neurology 2000; 55:370-376.

Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika 1982;69:239-241.

Song X, Tsiatis AA, Davidian M. A semiparametric likelihood approach to joint modeling of longitudinal covariates and time-to-event data. Biometrics 2002a; 58:742-753.

Song X, Tsiatis AA, Davidian M. An estimator for the proportional hazards model with multiple covariates measured with error. Biostatistics 2002b; 3:511-528.

Stefanski LA, Carroll RJ. Conditional scores and optimal scores in generalized linear measurement error models. Biometrika 1987; 74:703-716.

Storandt M, Grant EA, Miller JP, Morris JC. Rates of progression in mild cognitive impairment and early Alzheimer disease. Neurology 2002;59:1034-1041.

Therneau TM, Grambsch PM. Modeling Survival Data: extending the Cox model. New York: Springer, 2000.

Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals and survival models. Biometrika 1990;77:147-160.

Tomlinson BE, Blessed G, Roth M. Observations on the brains of non-demented old people. J Neurol Sci 1968; 7:331-356.

Troncoso JC, Martin LJ, Dal Forno G, Kawas CH. Neuropathology in controls and demented subjects from the Baltimore Longitudinal Study of Aging. Neurobiol Aging 1996; 17:365-371.

Tsay R. Regression models with time series errors. Journal of the American Statistical Association 1984; 79:118-124.

Tsiatis AA, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika 2001; 88:447-458.

Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica, 2004; 14:809-834.

Tukey, JW. Exploratory Data Analysis. Addison-Wesley, Reading, MA. 1977.

Tunnicliffe-Wilson G. On the use of marginal likelihood in time series model estimation. Journal of the Royal Statistical Society B 1989; 51:15-27.

Vaida F, Xu R. Proportional hazards model with random effects, Statistics in Medicine 2000;19:3309-3324.

Verbyla AO, Cullis BR. Modeling in repeated measures experiments. Applied Statistics 1990;39:341-56.

Vonesh EF, Carter RL. Mixed effect nonlinear regression for unbalanced repeated measures. Biometrics. 1992;48:1-18.

Wang Y, Taylor, JMC. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. J. Amer. Statist. Assoc. 2001;96:895-905.

Xiong C, El Barmi H. On detecting change in likelihood ratio ordering. Journal of Nonparametric Statistics 2002; 14, 5: 555-568.

Xiong C, Miller JP, Morris JC. Testing correlation of cognitive decline at adjacent stages of dementia. Journal

of the Alzheimer's Disease 2003; 5,5:409-418.
Xiong C, Milliken GA. Step-stress life-testing with random stress-change times for exponential data. IEEE Transactions on Reliability 1999; 48, 2: 141-148.
Xiong C, Milliken GA. Changepoints in stochastic ordering. Communications in Statistics: Theory and Methods 2000; 29, 2: 381-400.
Xiong C, Zhu K, Yu K, JP Miller. Statistical Modeling in Biomedical Research: Longitudinal Data Analysis. Epidemiology and Medical Statistics, (ed. C.R. Rao, J. Philip Miller, D.C. Rao, in press, 2007)
Xu J, Zeger SL. Joint analysis of longitudinal data comprising repeated measures and times to events. Appl. Statist. 2001;50:375-387.
Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121-130.
Zeger SL, Liang KY, Self SG. The analysis of binary longitudinal data with time-dependent covariates. Biometrika 1985;72:31-38.
Zeger SL, Qaqish B. Markov regression models for time series: a quasi-likelihood approach. Biometrics 1988;44:1019-1031.
Zhao LP, Prentice RL. Correlated binary regression using a generalized quadratic model. Biometrika 1990; 77:642-648.
Zimmerman DW. A note on the influence of outliers on parametric and nonparametric tests. Journal of general Psychology, 1994;121(4):391-401.

I. MULTIPLE PI LEADERSHIP PLAN – Not applicable

J. CONSORTIUM/CONTRACTUAL ARRANGEMENTS - None

K. RESOURCE SHARING – The DIAN resource sharing plan can be found in Core A: Administration.

L. CONSULTANTS - Dr. John Rice is a well known genetic epidemiologist with extensive experience in family and genetic studies. Dr. Rice will serve as a consultant to the Biostatistics Core as needed (please see his support letter).