Association of Proteomic Patterns With Cardiovascular Health

Writing Sample for PhD Application

(Methods and Results Excerpts)

Jackson (Zhiyuan) Song

Department of Statistics and Data Science

Northwestern University

Email: jackson.song@northwestern.edu

This document is an unpublished draft manuscript prepared solely as a writing sample to demonstrate scientific writing and quantitative analysis skills. Co-author names and protected dataset identifiers have been removed in accordance with confidentiality requirements.

Background

Cardiovascular disease (CVD) remains the leading cause of morbidity and mortality in the United States and worldwide. Maintaining ideal levels of cardiovascular health (CVH), as defined by the American Heart Association's Life's Essential 8 metrics, has been consistently associated with substantially lower lifetime risk of CVD and improved overall health outcomes. Individuals who achieve high CVH by midlife experience fewer cardiometabolic conditions, reduced clinical events, and better long-term health. Yet, CVH generally declines from early adulthood, and fewer than 5% of Americans sustain high CVH into middle age. This deterioration underscores the importance of understanding the biological pathways through which long-term CVH influences cardiovascular aging.

Circulating proteins regulate inflammation, vascular remodeling, metabolic signaling, and immune responses, and therefore represent biologically informative indicators of cardiometabolic health. Prior studies have identified individual proteins associated with CVD risk factors or events, but most have been cross-sectional, relied on a single timepoint of CVH, or were conducted in modestly sized samples. As a result, little is known about how cumulative CVH exposure across adulthood relates to proteomic patterns measured in midlife.

The Coronary Artery Risk Development in Young Adults (CARDIA) study provides a unique opportunity to investigate these questions. CARDIA enrolled 5,115 Black and White young adults in 1985–1986 and has followed participants for more than 35 years with repeated, detailed assessments of lifestyle and cardiometabolic health. Large-scale Olink proteomic profiling was performed in midlife, capturing 5,347 circulating proteins across diverse biological pathways. In addition to long-term CVH measurements, CARDIA includes adjudicated CVD events throughout follow-up, allowing proteomic findings to be interpreted within a clinically meaningful context. This combination of extended CVH data, rich CVD outcome information, and expansive proteomic coverage offers a rare opportunity to examine whether

cumulative CVH is reflected in molecular signatures during midlife. Leveraging this uniquely comprehensive dataset, our study aims to characterize proteomic correlates of CVH and to identify proteins that capture the sustained biological impact of cardiovascular health across adulthood.

Methods

Study Population

This study was conducted within CARDIA cohort, using examination data collected from Year 0 through Year 35. The dataset includes repeated measures of sociodemographic characteristics, clinical risk factors, and CVH metrics across adulthood, along with adjudicated CVD outcomes. Proteomic profiling became available beginning at the Year 15 examination and served as the primary biomarker assessment for the present analyses.

Analytic samples for each analysis were defined based on the availability of complete data for the relevant exposure, outcome, and covariates. Because the required variables differed across analyses—such as the availability of proteomic data, CVH measurements, or CVD follow-up—the sample size varied accordingly. Despite these differences, most analyses included more than 1,000 participants, reflecting the substantial overlap of available CARDIA data elements. Exact analytic sample sizes are reported in the corresponding tables and figures.

Definition of Clinical CVH Score and Cumulative CVH

Clinical CVH score at each examination was assessed using the eight components of LE8: diet quality, physical activity, smoking, sleep health, body mass index, blood pressure, blood glucose, and total cholesterol. Each component is observed at each exam year and scored on

a 0–100 scale based on established LE8 algorithms. The clinical CVH score at each visit was defined as the mean of the eight component scores, yielding a summary CVH value ranging from 0 to 100.

To characterize long-term cardiovascular health across adulthood, we constructed a cumulative CVH measure using an area-under-the-curve (AUC) approach across repeated CVH assessments, it was defined as:

$$\text{Cumulative CVH} \;=\; \frac{1}{t_k - t_0} \int_{t_0}^{t_k} \text{CVH}(t)\, dt,$$

where t_0 and t_k denote the times of the first and last available CVH assessments. The integral was approximated using the trapezoidal rule applied to observed CVH scores at consecutive examinations. This measure reflects sustained cardiovascular health exposure over the full follow-up period.

Proteomic Profiling

Proteomic measurements were generated using the Olink® Explore HT platform, which provides $\log_2$-normalized protein expression (NPX) values based on the manufacturer's standard preprocessing pipeline. Protein assays were first introduced at the Year 15 examination and continued to be collected every five years through Year 30.

Data cleaning steps included restricting measurements to proteins classified with `AssayType = "assay"` and removing technical controls. After filtering, a total of 5,347 proteins remained for analysis. When multiple NPX values were recorded for the same participant–protein pair, values were averaged to yield a single measurement per protein. Protein identifiers followed Olink-provided gene symbols to facilitate downstream biological interpretation.

CVD Outcomes

CVD outcomes were obtained from the adjudicated CARDIA event files and were provided as a composite indicator for the first occurrence of any cardiovascular event. In the analytic cohort, approximately 8.6% of participants experienced an incident CVD event during follow-up.

Participants were followed from the Year 0 examination until death, loss to follow-up, or the end of the most recent adjudication cycle. Individuals with prevalent CVD at baseline were excluded. Deaths not attributed to CVD were treated as censoring events at the recorded date of death. CVD deaths were considered incident CVD events. Time-to-event analyses used time from baseline to either incident CVD or censoring.

Statistical Analysis

Aim 1: Associations Between CVH and Proteomic Profiles

Aim 1 evaluated whether CVH or cumulative CVH in young adulthood were associated with plasma proteomic profiles in midlife. We examined these associations using complementary approaches, including cross-sectional analyses at each proteomic examination, models incorporating cumulative CVH over prespecified life-course intervals, and longitudinal mixed-effects models integrating repeated CVH assessments over time.

Cross-sectional analyses: Cross-sectional associations between CVH and protein levels were assessed at Years 15, 20, 25, and 30. For each examination year and each protein, we

fit a linear regression model of the form:

$$NPX_{ij} = \beta_0 + \beta_1 \,\mathrm{CVH}_i + \beta_2 \,\mathrm{age}_i + \beta_3 \,\mathrm{sex}_i + \beta_4 \,\mathrm{race}_i + \beta_5 \,\mathrm{center}_i + \beta_6 \,\mathrm{education}_i + \varepsilon_{ij},$$

where NPX_{ij} denotes the $\log_2$-normalized protein expression for participant i and protein j. Sex, race, study center, age and education were treated as covariates. Regression models were fit separately for each protein at each examination year, and the primary parameter of interest was β_1, representing concurrent the association between CVH and protein abundance.

All models in Aim 1 were adjusted for age, sex, race, study center, and educational attainment, with all covariates taken from the Year 15 examination. Multiple testing across proteins was controlled using the Benjamini–Hochberg false discovery rate (FDR). Analyses were conducted using complete-case data for all covariates.

Longitudinal analysis: Whereas the cross-sectional analyses characterized associations between CVH and protein levels at individual examinations, the longitudinal models leveraged repeated measurements of both CVH and proteins to evaluate how within-person changes in CVH related to trajectories of protein expression over time. To evaluate longitudinal associations between repeated CVH assessments and protein levels measured from midlife through early older adulthood, we fit linear mixed-effects models for proteins with available measurements at all four proteomic examinations (Years 15, 20, 25, and 30). For each protein, the following model was specified:

$$\begin{aligned} NPX_{it} = \beta_0 + \beta_1 \mathrm{CVH}_{it} + \beta_2 \mathrm{Year}_t + \beta_3 \mathrm{age}_i + \beta_4 \mathrm{sex}_i \\ + \beta_5 \mathrm{race}_i + \beta_6 \mathrm{center}_i + \beta_7 \mathrm{education}_i + b_i + \varepsilon_{it}. \end{aligned}$$

where NPX_{it} denotes the protein expression for participant i at examination t; CVH_{it} is the time-varying CVH score; and Year_t corresponds to the examination year coded as a

continuous time variable (15, 20, 25, 30). A participant-specific random intercept b_i was included to account for within-person correlation across repeated protein measurements. The coefficient of interest was β_1, representing the association between time-varying CVH and longitudinal protein levels.

Cumulative CVH analyses. To evaluate whether long-term CVH from young adulthood was associated with proteomic profiles measured in midlife, we fit regression models relating cumulative CVH to protein levels at Years 15, 20, 25, and 30. For each participant, cumulative CVH from Year 0 to Year 15 was served as the primary exposure of interest. For examinations occurring at Years 20, 25, and 30, models additionally included cumulative CVH from Year 15 to the corresponding examination year to account for more recent CVH. For protein measurements at examination year t, the model was specified as:

$$\begin{aligned} NPX_i = \beta_0 + \beta_1 \text{CVH}_{0\text{–}15,i} + \beta_2 \text{CVH}_{15\text{–}t,i} + \beta_3 \text{age}_i + \beta_4 \text{sex}_i \\ + \beta_5 \text{race}_i + \beta_6 \text{center}_i + \beta_7 \text{education}_i + \varepsilon_i. \end{aligned}$$

where $\text{CVH}_{0\text{–}15,i}$ denotes cumulative CVH from Year 0 to Year 15 and $\text{CVH}_{15\text{–}t,i}$ represents cumulative CVH from Year 15 to examination year t (included only for $t = 20, 25, 30$). The primary parameter of interest was β_1, representing the association between cumulative CVH from young adulthood and midlife protein expression.

Aim 2: Incremental Predictive Value of Proteomic Profiling

Proteomic feature screening. To identify proteomic markers associated with incident cardiovascular events, we first evaluated the marginal association between each protein and time to event using Cox proportional hazards regression. For each protein Z_j, the model was specified as:

$$h_i(t) = h_0(t) \exp\{\beta_j Z_{ij}\},$$

where $h_i(t)$ is the hazard for participant i, Z_{ij} denotes the NPX ($\log_2$-transformed) abundance of protein j, and β_j is the marginal log-hazard ratio. Proteins were ranked by their Wald test p-values.

Although classical SIS recommends selecting the top $n/\log(n)$ features, this cutoff alone may miss biologically meaningful proteins. Therefore, we retained the union of proteins selected by SIS and those meeting an FDR < 0.05 threshold, ensuring that both marginally strong and FDR-significant proteins were included for multivariable modeling.

Penalized Cox regression. Let X_i denote the vector of baseline covariates for participant i (age, sex, race, education, study center, and CVH trajectory), and let $Z_i = (Z_{i1}, \ldots, Z_{ip})^\top$ denote the set of selected proteins from the previous step. Joint modeling and feature selection were performed using a penalized Cox proportional hazards model with least absolute shrinkage and selection operator (LASSO). The multivariable model was:

$$h_i(t) = h_0(t) \exp\left\{ \alpha^\top X_i + \sum_{j=1}^{p} \gamma_j Z_{ij} \right\},$$

where α are unpenalized coefficients for clinical covariates, and γ_j are protein coefficients penalized under an L_1 norm. The LASSO estimator was obtained by:

$$\hat{\gamma} = \arg\min_{\gamma} \left\{ -\ell(\gamma) + \lambda \sum_{j=1}^{p} |\gamma_j| \right\},$$

with λ selected via 10-fold cross-validation based on the partial likelihood. This procedure yielded a sparse set of proteins jointly associated with incident cardiovascular events.

Models for comparison. Two prognostic models were constructed for evaluation. The clinical-only model (M0) included demographic and behavioral covariates:

$$\text{M0:} \quad h_i(t) = h_0(t) \exp\{\alpha^\top X_i\}.$$

The protein-enhanced model (M1) incorporated the LASSO-selected proteins in addition to the clinical covariates:

$$\text{M1:} \quad h_i(t) = h_0(t) \exp\{\alpha^\top X_i + \gamma^\top Z_i\}.$$

Both models were fit using the same complete-case analytic dataset to ensure comparability of predictive performance.

Predictive performance evaluation. Model discrimination was quantified using Harrell's concordance index (C-index), computed from the linear predictors of each model. To assess statistical uncertainty, bootstrap resampling with replacement was performed to obtain bootstrap distributions for the C-index of each model and for the difference in discrimination between the protein-enhanced and baseline models.

For visualization, the linear predictor from the protein-enhanced model (M1) was dichotomized at its median to define high- and low-risk groups. Kaplan–Meier survival curves were constructed for these groups, and differences in survival distributions were assessed using the log-rank test.

Results

Aim 1: Associations Between CVH and Proteomic Profiles

(1) Cross-sectional analyses Across the four examination years, sample sizes for the cross-sectional analyses ranged from **1,145 to 2,802** participants. Large numbers of proteins showed significant associations with same-year CVH after FDR correction: **1,594 proteins at Y15**, **1,158 at Y20**, **2,738 at Y25**, and **2,548 at Y30**. The higher numbers observed at Y25 and Y30 indicated stronger proteomic differentiation of CVH during mid-adulthood.

A total of **955 proteins** were significant at all four timepoints, forming a large and stable set of CVH-associated markers across adulthood. These overlapping proteins also demonstrated consistent directions of association across Y15, Y20, Y25, and Y30, supporting the robustness of the cross-sectional signal.

The effect sizes for CVH–protein associations were modest in magnitude and highly similar across examination years. Median coefficients ranged between **-0.002 and -0.001**, with interquartile ranges concentrated between approximately **-0.006 and 0.001**. Extreme values were rare, and the largest coefficients did not exceed **0.04** in absolute magnitude. As shown in Figure 1, the effect-size distributions were nearly identical across all four exam cycles. Summary statistics for each year are provided in Table 1.

Table 1: Statistics of regression coefficients for cross-sectional CVH–protein associations across examination years.

Year	Min	Q1	Median	Q3	Max	SD
Y15	-0.0454	-0.00838	-0.00171	0.000829	0.0370	0.00447
Y20	-0.0454	-0.00441	-0.00191	0.000469	0.0378	0.00488
Y25	-0.0429	-0.00594	-0.00243	-0.0000752	0.0286	0.00474
Y30	-0.0419	-0.00521	-0.00167	0.000584	0.0276	0.00473

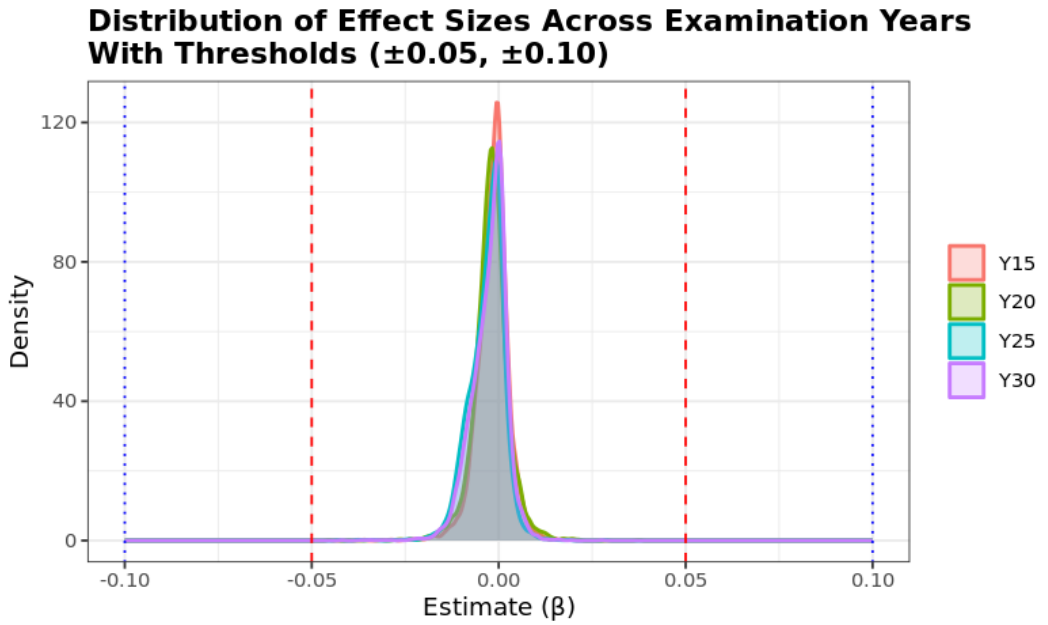

Figure 1: Density distributions of regression coefficients for cross-sectional CVH–protein associations across examination years.

(2) Longitudinal associations. A total of **899** participants contributed repeated proteomic and CVH measurements across four examination years (Y15, Y20, Y25 and Y30), enabling longitudinal evaluation of within-person CVH–proteomics associations.

A large proportion of proteins demonstrated statistically significant longitudinal associations with CVH. In total, **2,364 proteins** met the FDR < 0.05 threshold. Although significant associations were widespread, effect sizes were generally small: the distribution of estimated longitudinal coefficients centered near zero with interquartile ranges of approximately $[-0.003, 0.001]$.

To facilitate interpretation of biological magnitude, we defined effect-size thresholds corresponding to 5% and 10% NPX change per one-unit higher CVH score. On the $\log_2$ NPX scale, these correspond to coefficient magnitudes of $|\beta| \geq 0.070$ and $|\beta| \geq 0.138$, respectively. Among all proteins, **718** exceeded the 5% threshold, and **56** met the stricter 10% threshold. When combining statistical and magnitude criteria, **696** proteins satisfied both FDR < 0.05 and 5% threshold, while **56** proteins remained significant at the 10% threshold.

The volcano plot (Figure 2) illustrates the overall distribution of associations, showing tightly concentrated effect sizes with a subset of proteins exhibiting both strong statistical significance and moderate biological magnitude. Together, these findings indicate that longitudinal CVH dynamics are accompanied by widespread but modest shifts in proteomic profiles, with a smaller set of proteins demonstrating changes of potentially larger biological relevance.

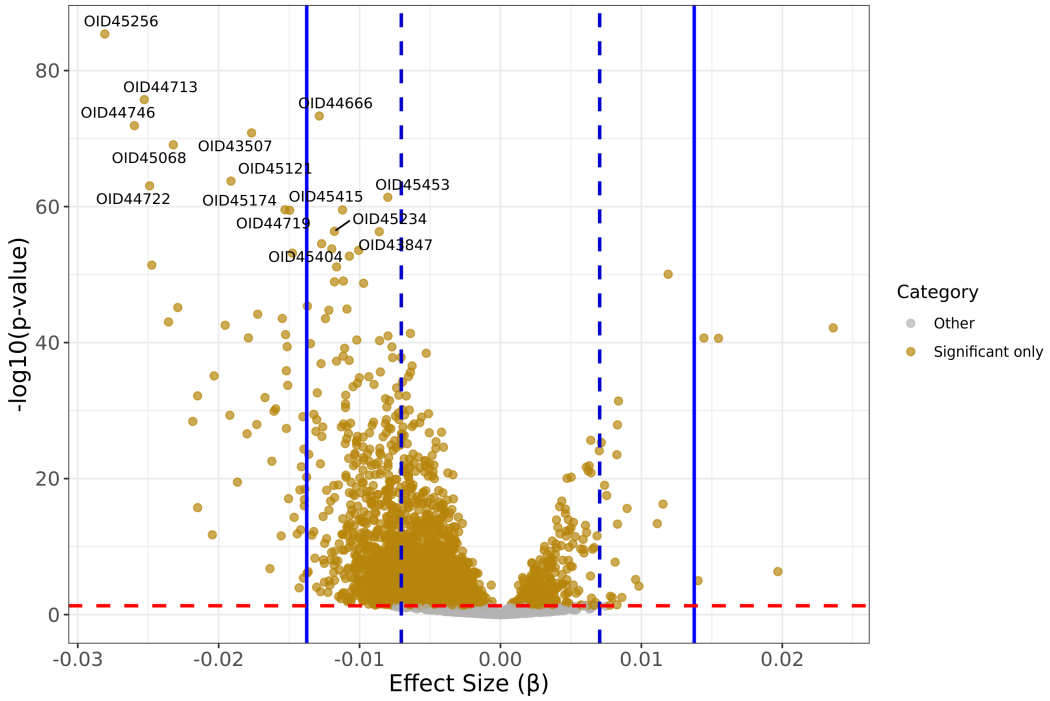

Figure 2: Volcano plot of longitudinal CVH–proteomics associations (FDR and 5%/10% thresholds).

(3) Long-term CVH history and midlife proteomic profiles. Across examinations, analytic sample sizes varied. **At Y15, 1,425** participants had complete $\text{CVH}_{0\text{–}15}$ data, proteomics, and covariates; the corresponding sample sizes were **801 at Y20**, **1,935 at Y25** and **1,834 at Y30**.

Regression analyses showed more recent cumulative CVH were broadly associated with the midlife proteome, identifying **1,350** significant proteins at both Y25 and Y30. In contrast, the early-life CVH component ($\text{CVH}_{0\text{–}15}$) yielded only **16** significant proteins across models.

Among these, a consistent subset of **6** proteins showed significance for both early-life and recent CVH components, indicating a small but robust imprint of CVH accumulated from young adulthood. These overlapping proteins were **OID44346, OID44680, OID44749, OID44891, OID45073,** and **OID45131**.

These findings together indicate that midlife protein expression patterns are shaped largely by CVH levels proximal to the time of assessment, while only a small set of proteins appears to capture long-term CVH exposures from earlier adulthood.

Aim 2: Incremental Predictive Value of Proteomic Markers

(1) Proteomic feature selection. Among the **1,415** participants with complete $CVH_{0–15}$ data, covariates, CVD outcomes and proteomic measurements, a total of **535 proteins** passed the univariate survival screening and were advanced to multivariable modeling. These proteins were entered into the penalized Cox regression along with the full set of baseline covariates ($CVH_{0–15}$ and five demographic variables: age, sex, race, education, and study center). Cross-validated LASSO selected a total of **64 predictors**, comprising **62 proteins** and **2 clinical variables** ($CVH_{0–15}$ and study center).

(2) Improvement in discrimination. The baseline Cox model (M0), incorporating $CVH_{0–15}$ and all five covariates, demonstrated moderate discrimination. Adding the 62 LASSO-selected proteomic markers (M1) substantially improved model performance, increasing the Harrell C-index from **0.793** to **0.916**. The absolute gain in discrimination (Δ**C-index = 0.123**) was consistently observed across 500 bootstrap resamples, indicating robust incremental predictive value.

Bootstrap-derived 95% confidence intervals for M0, M1, and the ΔC-index are summarized in Table 2.

Table 2: Discrimination performance of baseline and protein-enhanced models

Model	C-index	95% CI
Baseline model (M0): CVH + 5 covariates	0.793	[0.755, 0.844]
Protein-enhanced model (M1)	0.916	[0.918, 0.958]
ΔC-index (M1–M0)	0.123	[0.101, 0.183]

(3) Risk stratification. To further illustrate the incremental value of the proteomic markers, participants were stratified into high- and low-risk groups based on the linear predictor from the protein-enhanced model (M1). Kaplan–Meier survival curves demonstrated clear separation between the two groups throughout follow-up, with the high-risk group showing substantially lower survival probabilities.

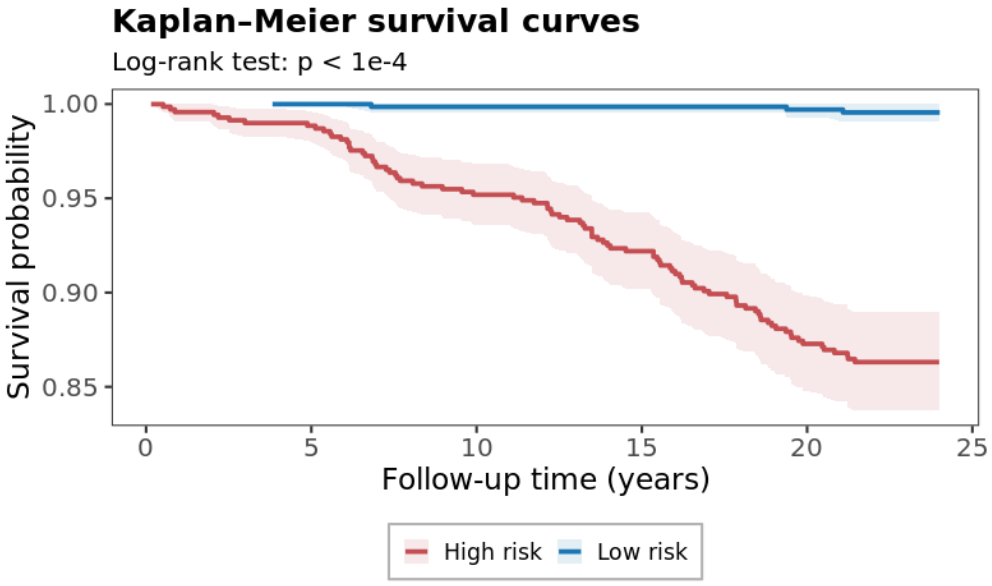

Figure 3: The curves show clear separation across follow-up, and the log-rank test indicated a significant difference in survival distributions ($p < 1 \times 10^{-4}$). Shaded regions represent 95% confidence intervals.