Data Analysis
Continuous Variables

James Ware, PhD
Harvard School of Public Health

Data Analysis

The next three segments discuss methods for the analysis of clinical trial data

We discuss the three most important types of endpoints:

- Continuous Variables
- Binary Outcomes
- Time to Event Data

Data Analysis

We discuss both univariate and multivariable methods

The goals of the analysis:

- Estimate measures of association
- Test for a treatment effect
- Multivariable analysis

The specific statistical methods depend on whether the endpoint is continuous, binary, or time-to-event

Tests for Association By Endpoint Type

Outcome Variable	Univariate	Multivariable
Continuous	t test Mann-Whitney	Multiple Linear Regression
Binary	Fisher's Exact test Chi-Square test	Logistic Regression
Time-to-Event	Log-rank test	Proportional Hazards Regression

Treatment of Lead-Exposed Children Trial*

TLC was a randomized trial of oral chelation therapy with succimer in children exposed to environmental lead. The goal was to prevent neurologic injury.

We enrolled 780 children. The primary endpoint was a measure of general intelligence (the Wechsler Scales of Intelligence) 3 years after randomization

The plan was to compare mean IQ scores in the succimer and placebo groups

*NEJM 2001; 344:1421-26

Univariate Analysis of Measured Outcomes

If the sample means are approximately normally distributed, we can use the two-sample t test to test for equality of means

If not, use a non-parametric test, usually the Wilcoxon (Mann-Whitney) rank test

Intelligence test scores are normed to be approximately normally distributed

Univariate Analysis in the TLC Trial

	Treatment Group	
	Succimer	Placebo
WSPPI-R*		
No. Tested	368	377
Score	80.6 ± 13.1	80.6 ± 13.3

t = .02 (P = 0.98)

*Wechsler Scales of Preschool and Primary Intelligence-Revised

Multiple Linear Regression Analysis

Multivariable analysis is the general term for models that describe the effect of several variables, including treatment group, on study endpoints

These other variables may be called prognostic factors, predictor variables, risk factors, or independent variables

Why Study Prognostic Factors?

To correct estimates of treatment groups for imbalances in baseline characteristics

To improve the precision of comparisons

To investigate effect modification

To define risk for individual patients

To guide the design of future studies

To aid in understanding mechanisms of disease and to generate new leads

Multiple Regression Analysis

For measured outcomes, the most common approach to multivariable analysis is multiple linear regression

Sometimes called analysis of covariance, especially if the baseline value of the endpoint is in the model

Multiple Regression Analysis

Let y_i be the measured outcome for the i^{th} patient

Also, let X_{ij} be the value of the j^{th} covariate for the i^{th} patient

Then, we fit models of the form:

$$y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + e_i$$

α is the intercept,

β_i is the regression coefficient for X_{ij}

Multiple Regression in TLC

"We used multiple linear regression to adjust the mean differences in WPPSI-R score for a set of baseline covariates chosen in advance."

Clinical center,

Body-surface area,

Baseline blood lead level,

Language [Spanish or English],

Caregiver's IQ,

Child's baseline IQ

The TLC Regression Model

$$\text{WPPSI-R Score} = \alpha + \beta_1 \text{Clinic}_1 + \ldots \beta_m \text{Clinic}_{m-1} + \beta_{m+1} \text{BSA}_1 + \beta_{m+2} \text{BSA}_2 + \beta_{m+3} \text{PbLevel} + \beta_{m+4} \text{Spanish} + \beta_{m+5} \text{CaregiverIQ} + \beta_{m+6} \text{BaselineMDI}$$

Clinic, BSA, PBLevel, and Spanish are indicator variables for categorical covariates

CaregiverIQ and BaselineMDI are continuous variables

Multiple Regression in TLC

"After adjustment, the mean full-scale IQ score on the WPPSI-R for children given succimer was 1.1 points (95 percent confidence interval for the difference, –2.6 to 0.5) lower than that for children given placebo"

NEJM, Volume 344:1421-1426

Univariate SE = 0.97, Adjusted SE = 0.79

Properties of ML Regression

The validity of MLR depends on the assumption that the relationships are (approximately) linear

Individual values must be "well behaved". Extreme values can cause trouble

The estimated regression coefficients are approximately normally distributed even if the individual values are not normally distributed. This follows from the central limit theorem