Society for Clinical Trials 31st Annual Meeting

Workshop P11
I am NOT a Statistician, but I Understand What You are Saying

Sunday, May 16, 2010
1:00 PM - 5:00 PM
Laurel CD
WORKSHOP 11 - I am NOT a Statistician, but I Understand What You Are Saying

1. Overall, did the subject context of this workshop meet your expectations and needs?  
   Yes ( )  No ( )  
   If yes, in what way? If no, why not? ____________________________________________  
   ___________________________________________________________________________  

2. Was the content of this workshop of value to you personally or on the Job?  
   Yes ( )  No ( )  

3. Was the content of the workshop:  
   New ( )  New/Review ( )  Review ( )  

4. The level and complexity of this workshop was:  
   Too elementary ( )  Correct ( )  Too advanced ( )  

5. Rate the extent to which this workshop:  
   a. Presented content clearly  
      1 2 3 4 5  
   b. Allowed sufficient time for discussion and audience participation  
      1 2 3 4 5  
   c. Provided useful information  
      1 2 3 4 5  
   d. Utilized appropriate teaching methods, i.e., audiovisual, handouts, lectures  
      1 2 3 4 5  

6. Please rate each workshop faculty member:  

<table>
<thead>
<tr>
<th>Name</th>
<th>Knowledge of Subject</th>
<th>Organization/Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nicole C. Close</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>Anita F. Das</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>Cora MacPherson</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>
1. Are you currently working in a clinical trial? (Yes) (No)

2. What is your job title? __________________________________________________________

3. Do you have any suggested topics for workshops at future meetings? If so, please list below:
   ____________________________________________________________________________
   ____________________________________________________________________________

4. What aspect of the workshop did you like best?
   ____________________________________________________________________________
   ____________________________________________________________________________

5. What aspect of the workshop would you change if this workshop were offered again?
   ____________________________________________________________________________
   ____________________________________________________________________________

6. Additional Comments: _________________________________________________________
   ____________________________________________________________________________
Basic Statistical Concepts

Anita F. Das
AxiStat, Inc.
adas@axistat.com

Overview

- Types of Data
- Descriptive Statistics
- Distributions
- Hypothesis Testing
- Statistical Tests

Types of Data
Types of Data

- **Continuous**
  - Data that can take on potentially infinite number of values (within certain restrictions)
  - Ex. Blood pressure, height, weight

- **Categorical**
  - Data separable into categories that are mutually exclusive
  - Ex. Gender, severity scale (none, mild, moderate, severe), race (Caucasian, African American, Asian, other race)

Categorical Data

- **Dichotomous (special type of unordered categorical data)**
  - Two levels – generally Yes vs No
  - Ex. Gender – instead of male vs. female, can think of this as Male, yes vs. no
  - Continuous or other categorical data can be summarized as a dichotomous variable
  - Ex. Birthweight <1500 grams, yes vs. no

- **Ordinal Data**
  - Ordered categorical data where there is a logical ordering to the categories
  - Ex. Adverse event severity - mild, moderate, severe or size of lesion - small, medium, large
  - For analysis, these may be assigned categories of 1, 2, 3. Numerical order means that 1 is "worse" than 2, but it does not mean that the difference in severity is the same from mild to moderate (1 to 2) and moderate to mild (2 to 3)
Categorical Data

- Unordered Categorical
  - No logical ordering to the categories
  - Ex. Race – Caucasian, African American, Asian, Other
  - Assign categories 1, 2, 3, 4 – the numerical values do not have any meaning

Descriptive Statistics

- Used to describe the main features of a set of data
  - Different from inferential statistics or hypothesis testing
  - Also called estimation
    - Point estimation
- Categorical data – Frequency distribution (n, %), contingency table
- Continuous data – location, dispersion, shape
Frequency Distribution

- Continuous Variable (age) and categorized it into a ordinal variable
- Show n and percent for each category

<table>
<thead>
<tr>
<th>Age Group (years)</th>
<th>Study Population (N=200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>18 (9%)</td>
</tr>
<tr>
<td>35-45</td>
<td>42 (21%)</td>
</tr>
<tr>
<td>46-55</td>
<td>90 (45%)</td>
</tr>
<tr>
<td>56-65</td>
<td>38 (19%)</td>
</tr>
<tr>
<td>&gt;65</td>
<td>12 (6%)</td>
</tr>
</tbody>
</table>

Bar Chart

Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Linezolid (N=200) n (%)</th>
<th>Vancomycin (N=205) n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Cure</td>
<td>170 (85.0)</td>
<td>177 (86.3)</td>
</tr>
<tr>
<td>Clinical Failure</td>
<td>20 (10.0)</td>
<td>25 (12.2%)</td>
</tr>
<tr>
<td>Indeterminate</td>
<td>10 (5.0)</td>
<td>3 (1.5%)</td>
</tr>
</tbody>
</table>
Continuous Data

Location – measures of central tendency
- Mean (average) – sum of all values divided by number of values
  - Most common descriptive statistic
    - Should not be used for data that does not follow a normal distribution (discussed later)
- Median (50th percentile) – the value where 50% of the observations are below and 50% of the observations are above (middle value)
- Mode – most frequently occurring value

Dispersion – measures of spread
- Standard deviation (square root of the variance)
- Range – the difference between the largest and smallest observation
  - Can be presented as the minimum (0%ile) and maximum (100%ile) values
- Percentiles – the value below which a certain percentage of observations fall (25%ile – 75%ile)

Shape - Continuous Data
- Skewness – if mean = median, then skewness = 0
  - Ex. 1, 2, 3, 4, 100 – positive skewness
  - Ex. 1, 1001, 1002, 1003, 1004 – negative skewness
- Kurtosis – measure of “peakedness” of distribution
  - Not presented often in manuscripts, reports but used by statisticians to understand the data
Example of Descriptive Statistics

Probability Distributions

- Inferential statistical analysis is grounded in identifying the type of data being analyzed and the distribution the data follows.
- Data that follows a distribution can be analyzed using parametric methods.
- Data that does not follow a distribution is analyzed using nonparametric methods.
Normal Distribution

- Also called Gaussian distribution or bell-shaped curve
- Corner stone of most methods of estimation and inference
- Most random variables in the general population, for example, birthweight and blood pressure, follow a normal distribution
- Many variables that are not normally distributed can be closely approximated by a normal distribution

Generally refer to a normal distribution as $N(\mu, \sigma^2)$ where $\mu =$ mean and $\sigma^2 =$ variance
- Standard normal is $N(0,1)$
- Standard normal 95% of the area lies between -1.96 and 1.96

Other Distributions - Discrete

- Binomial distribution
  - Dichotomous variables follow a binomial distribution with mean=np and variance=npq
    - n=number of observations, p=probability of event, and q=1-p
  - Has highest variance when p=0.50
- Poisson distribution
  - Used for dichotomous variables when p is very low (rare event)
  - Counts of events (k events in a time interval $T$)
  - Both can be approximated by the Normal distribution (np>=5)
Other Distributions

- Student’s t-distribution
  - A family of distributions indexed by a parameter referred to as the degrees of freedom (df)
  - Always symmetric about 0 for any df
- Exponential Distribution
- Chi-square
  - A family of distributions indexed by df
  - Only takes on positive values and is generally skewed to the right

Estimation

- Assume that properties of underlying distribution of the population from which the data are drawn are known
- Have a sample from the population and want to estimate population parameters
- Ex. Population is normally distributed $N(\mu, \sigma^2)$ (birthweight), what is mean ($x\bar{b}$) and standard deviation of our sample?
Estimation

- Point estimate
  - Descriptive statistics discussed earlier
  - Ex. Mean and standard deviation are point estimates
- Interval estimation
  - Specify a range within which each parameter falls
  - Ex. Confidence interval

Central Limit Theorem

- Let \( x_1, x_2, \ldots, x_n \) be a sample from a population with mean \( \mu \) and variance \( \sigma^2 \). For large \( n \), \( \bar{x} \sim N(\mu, \sigma^2/n) \) even if the underlying distribution of the individual observations in the population is not normal
- What this means – for large \( n \), can almost always use normal distribution even if data are not normally distributed

Interval Estimation

- General formula for a confidence interval
  - 95% CI for \( \mu \) \( (\bar{x} +/1.96\sigma/\sqrt{n}) \)
- Interpretation
  - 95% of intervals that would be constructed taking repeated samples of size \( n \), will contain the parameter \( \mu \)
  - Cannot say that there is a 95% chance that the parameter \( \mu \) will fall with a particular 95% CI
Hypothesis Testing

Hypothesis testing-General Concepts

- One sample inference
  - Ex. Birthweights from women of low socioeconomic status are lower than the national average, where national average is a fixed number
- Two-sample inference
  - Ex. Studied birthweights from women at one hospital and defined two groups of women – low socioeconomic status and average and above socioeconomic status

General Concepts

- Define a null and alternative hypothesis
  - Null H₀: μ₁ = μ₂
  - μ₁ = mean birthweight for women of low socioeconomic status
  - μ₂ = mean birthweight for women of average and above socioeconomic status
General Concepts

- Alternative hypothesis
- Two-sided
  \[ H_1: \mu_1 \neq \mu_2 \]
- One-sided
  \[ H_1: \mu_1 \leq \mu_2 \]

Rarely used

General Concepts

- Four possible events that can occur
  - Accept Ho and Ho is in fact true
  - Accept Ho and H1 is in fact true
  - Reject Ho and Ho is in fact true
  - Reject Ho and H1 is in fact true

General Concepts

<table>
<thead>
<tr>
<th>Ho True</th>
<th>H1 True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept Ho</td>
<td>Got it right!</td>
</tr>
<tr>
<td>Reject Ho</td>
<td>Type I error</td>
</tr>
</tbody>
</table>
General Concepts

- Probability of a type I error
  - Usually denoted by alpha
  - Commonly referred to as the significance level
- Probability of a type II error
  - Usually denoted by beta
  - Power of a test is defined as 1-Beta
- General aim is to make alpha and beta as small as possible

P-Value

- Defined as the alpha level at which we would be indifferent between accepting and rejecting Ho given the sample data at hand
- Is the alpha level at which the given value of the statistic would be on the borderline between the acceptance and rejection region
- The probability of obtaining a result as extreme or more extreme than the actual sample value obtained given that the null hypothesis is true

General Interpretation of P-value

- 0.01<=p<0.05 – results are significant
- 0.001<=p<0.01 – results are highly significant
- P<0.001 – results are very highly significant
- P>0.05 – not statistically significant
- 0.05<=p<0.10 trend towards statistical significance is sometimes noted
Statistical Tests

Assuming two groups
- Group variable is independent variable
- Outcome (variable being tested) is dependent variable

How do you determine what statistical test to use?

Are the groups independent or paired?

What type of data do you have for the dependent variable?
- Continuous or categorical

How is the data distributed?

What is the hypothesis?
Continuous Data

- Clinical trial where women randomized to receive 17P or placebo
- Independent variable is study drug, 17P and placebo
- Independent groups – different women in the 17P group than the placebo group
- Dependent variable (outcome) is birthweight of baby
- Birthweight is a continuous variable
  - Most continuous data is normally distributed

Continuous Data

- Normally distributed
  - Since it follows a distribution, use a parametric method
- Define null and alternative
  - Ho: Mean birthweight in women treated with 17P = mean birthweight in women receiving placebo
  - H1: Mean birthweight in women treated with 17P ≠ mean birthweight in women receiving placebo
- Use t-test or Z-test
  - T-test will equal Z-test when n is 50 or greater
- Report mean and standard deviation at a minimum
  - Median, range, or 25th-75th percentiles often reported as well
  - Report p-value

Continuous Data

- What if not normally distributed
  - Ex. Gestational age at delivery
- Two options
  - Transform the data
    - Use a mathematical transformation such as log(x) – transformed data often normal
    - Difficult to interpret
  - Use a non-parametric analysis method
    - Wilcoxon Rank Sum Test is the non-parametric test analogous to a t-test
    - Report median, range or 25th-75th percentiles
    - Report p-value
Continuous Data

- Paired data
  - Ex. Measure blood pressure prior to receipt of a drug and then after receipt of the drug
  - Each patient serves as own control
  - Follows normal distribution – paired t-test
  - Not normally distributed – Wilcoxon Sign Rank test

Dichotomous Data

- Clinical trial where women randomized to receive 17P or placebo
- Dependent variable (outcome) is percent of women who delivered <37 weeks GA
- Dichotomous data follows a binomial distribution

Dichotomous Data

- Define null and alternative
  - For dichotomous outcome, can look at several different measures
    - Percent delivered preterm in each group
    - Relative risk of preterm delivery
    - Odds ratio of preterm delivery
  - Ho: Percent of women in the 17P group who delivered preterm = percent of women in the placebo group who delivered preterm
  - H1: Percent of women in the 17P group who delivered preterm ≠ percent of women in the placebo group who delivered preterm
Dichotomous Data

- Use chi-square test or z-test based on the normal approximation to the binomial
- Report n, percentages and p-value

Dichotomous Data

- Relative Risk
  - Risk of outcome in one group relative to the comparator group
  - $RR = \frac{p_1}{p_2}$ where $p_1$=percentage of patients with the outcome in one group and $p_2$=percentage of patients with the outcome in the comparator group
  - Tested for statistical significance by calculating a 95% confidence interval
  - If the CI contains 1, the relative risk is not statistically significant

Example of Relative Risk Summary
Categorical Data

- Ordinal and non-ordered categorical data can be tested using the chi-square test
  - Tests that there is a difference between groups
  - Does not indicate where the difference is
- For ordinal data, can also do a test for linearity
  - Ex. Mantel-Haenzel Chi-square test

Continuous Data

- Time to event data
  - Ex. Time to death (classic survival analysis)
  - Time to delivery
- Special statistical methods that allow comparison of time to event data even when not all patients have the event (data are censored)
- Survival analysis techniques
  - Most common is called Kaplan-Meier analysis

Example of Kaplan-Meier
Other Statistical Methods

- Many other statistical methods
  - More than 2 groups
    - Analysis of variance (parametric) for continuous data
    - Kruskal-Wallis (nonparametric) for continuous data
    - Chi-square for categorical data
- Controlling for covariates and modeling
  - Cochrane-Mantel-Haenzel chi-square
  - Linear regression
  - Logistic regression
  - Cox regression

Example Cox Regression

<table>
<thead>
<tr>
<th>Factor</th>
<th>Hazard Ratio</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.11</td>
<td>1.02 to 1.21</td>
<td>0.01</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education Level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of CVD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes Status</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questions and Workshop
Study Design Elements

Nicole C. Close, PhD
President and Principal Biostatistician

- All material adapted from EmpriStat's Boot Camp Training Series

Goals of this Session

- To be introduced to terminology and have it defined
- To discuss the major types of study design
- To learn about sample size, power and implications to the study
- Discuss randomization methodology and techniques
- Implement terms for understanding into practical examples and case studies.

Hypotheses, Objectives and Endpoints
What is your question?

“Each clinical trial must have a primary question. The primary question, as well as secondary questions, should be carefully selected, clearly defined, and stated in advance.”

- Friedman, Furberg and DeMets, 1996

Elements of a Good Hypothesis

- Written as a definite statement not a question
- Based on observations and knowledge
- Be testable
- Predict the anticipated results in clear terminology
- Based on an independent variable (experimental) and a dependent variable (responding, measured/observed)

Hypothesis

A hypothesis can be shown to be supported by the evidence but it can never be proved.

- One of Two Choices:
  - We reject the null hypothesis.
  - We fail to reject the null hypothesis.
Null Hypothesis

- Presumed true until statistical evidence indicates differently
- No difference exists between the two groups for the variable you are comparing

\[ H_0 : \mu_1 = \mu_2 \]

where:

- \( H_0 \) = the null hypothesis
- \( \mu_1 \) = the mean of population 1, and
- \( \mu_2 \) = the mean of population 2.

What is hypothesis testing?

- Probability of observing the obtained data or data more different from the prediction of the null hypothesis, if the null hypothesis is true (significance level)
- How often we would expect to observe our experimental results, or results even more extreme, if we were to take many samples from a population where there was no effect (i.e. we test against our null hypothesis)
- If we find that this happens rarely (5% of the time), we conclude that our results support our experimental prediction — we reject our null hypothesis.

Hypothesis Testing (preview)

- #1: State the hypothesis to be tested.
- #2: Choose the level of significance at which the test will be performed. This is called the size or level of the test. It is the probability of rejecting the null hypothesis when it is true.
- #3: Collect the data and reject the hypothesis or not depending on the observed value of the test statistic.
The questions (objectives)

- Primary:
  - Most interested in answering
  - One that is capable of being answered
  - Basis for your sample size
  - Emphasized in reporting of trial results

- Secondary:
  - Related to your primary question
  - Related to subgroup hypotheses

These define your primary and secondary objectives of the study.

Objectives

- A precise statement of the degree of benefit expected from the intervention, as well as the duration of the benefit

- Clear statements of the time frame of the study (especially in relation to how quickly benefits might occur)

- A definition of the participants for whom the benefit is sought.

Endpoints

- A measurement taken under specific conditions that reflects the response of an individual to treatment.
  - Defined and written in advance
  - Capable of being assessed in all participants
  - Participation generally ends when the endpoint occurs (unless there is a combination of primary endpoints); interested in events subsequent to primary endpoint.
  - Capable of unbiased assessment
  - Ascertained as completely as possible
**Surrogate Endpoints**

- A measure of effect (continuous) of a certain treatment that may correlate with a real endpoint but has no guaranteed relationship.
  - Considerations:
    - Does the endpoint truly reflect the clinical outcome?
    - Can it be assessed accurately and reliably?
    - Is there a large measurement error?
    - Does it require more and expensive equipment?
    - Does it require an invasive procedure?
    - Will the conclusions of the trial be scientifically accepted?

**Surrogate Endpoints**

- Potential advantages:
  - Since they are continuous, sample size can be smaller
  - If sample size is smaller, it may be less expensive as well
  - Likely to occur before the clinical endpoint, shortening the time for the trial conduct
- Potential disadvantage:
  - Effect of the drug on a surrogate endpoint is not necessarily a good indicator of clinical outcome

**Surrogate Endpoints**

- Use in Phase I or II trials
- Phase III; secondary endpoints (else validation should occur)
Composite Endpoints

- Useful if any one endpoint occurs too infrequently for the investigator reasonably to expect a significant difference without using a large number of participants
- Combined events capable of meaningful interpretation
- Only one event per participant will be counted (analysis by participant not event)

Examples

- **Objective**: To determine the efficacy of oral administration of a high and low dose of chenodeoxycholic acid dissolving or reducing the size of gallstones over 2 years in a sample of participants 21-79 years old, as compared with placebo treatment (National Cooperative Gallstone Study Group, 1981)

Examples

- **Endpoint**: Change from baseline in sitting diastolic blood pressure (mmHg) after 12 weeks of treatment.
Summary

- Clearly define your objectives and document
- Formulate your hypotheses to test
- Define and describe your endpoints

Basic Study Designs
Study Designs for Review

- Superiority
- Non-inferiority
- Equivalence
- Cross-over
- Bioequivalence
- Pharmacokinetic
- Pharmacodynamic

Basic Definitions

- Superiority: one is better than the other
- Non-inferiority: one is no worse than the other by a certain amount
- Equivalence: they are approximately the same by a certain amount (in either direction, better or worse)

Confidence Interval Approach
Bioequivalence

- Bioequivalence studies attempt to gain insight on formulation "switchability" (i.e., the ability to substitute one formulation for another without concern of the potential for reduced effectiveness or increased probability of adverse effects).

<table>
<thead>
<tr>
<th>Table 2: Data transformations performed in analysis of the crossover study for the bioequivalence study.</th>
<th>Reference</th>
<th>Reference</th>
<th>Difference</th>
<th>Reference</th>
<th>Reference</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Period 1</td>
<td>Period 2</td>
<td>Difference</td>
<td>Period 1</td>
<td>Period 2</td>
<td>Difference</td>
</tr>
<tr>
<td>1</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>11</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>12</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>13</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>14</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
<tr>
<td>15</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Pharmacokinetics

- How the body affects a specific drug after administration and the way the property of the drug may be affected by the site of administration and the concentration in which the drug is administered and how these affect the absorption rate.

- Population and individual PK/PD studies

- Design includes thought about:
  - Choice of drug input function
  - Choice of route of administration
  - Choice of sampling site
  - Choice of sampling scheme
  - Choice of surrogate effects in PK/PD studies
Pharmacodynamics

- Study of physiological effects of drugs on the body, mechanism of drug action, and the relationship between drug concentration and effect.

Randomization

- “Allocation concealment: A technique used to prevent selection bias by concealing the allocation sequence from those assigning participants to intervention groups, until the moment of assignment. Allocation concealment prevents researchers from (unconsciously or otherwise) influencing which participants are assigned to a given intervention group.”
  -- CONSORT statement
Randomization

- [Allocation sequence: A list of interventions, randomly ordered, used to assign sequentially enrolled participants to intervention groups. Also termed "assignment schedule," "randomization schedule," or "randomization list."]
  -- CONSORT Statement

The Randomization Process

- Randomized clinical trial is the standard by which all trials are judged.
- Randomization is a process where each participant has the same chance of being assigned to either the intervention or the control.
- Neither participant nor investigator should know what the assignment will be before the participant decides to enter the trial.

Randomization Methods

- Common methods with advantages and disadvantages
- Can be assumed that the strategy will allocate participants to two groups, but most methods can be generalized to more than two groups.
Bias

- Allocation is predictable (selection bias)
- Unbalanced groups with respect to risk factors or prognostic covariates
- For large studies the chance of unbalanced groups are negligible.
- Report of the trial should contain brief but clear description of the method employed.

Fixed Allocation Strategies

- Simple randomization
- Blocked randomization
- Stratified randomization
- Also adaptive strategies
  - Urn
  - Play the winner

Simple Randomization

- Toss an unbiased coin each time a participant is eligible for randomization
  - Random digit table (small studies)
  - Random number-producing algorithm (large studies)
- Alternating assignments is not simple randomization
**Advantages/Disadvantages of Simple Randomization**

- Easy to implement
- At the end each group will be in correct proportion, but at any point in the randomization, there could be a substantial imbalance.
  - Not invalid, but awkward
  - Not favored in the field

**Blocked Randomization**

- Participants are randomly assigned with equal probability to Group A or Group B for each block of even size (e.g. 4, 6, 8).
- Order in which the intervention is assigned in each block is random.

**Advantages/Disadvantages of Blocked Randomization**

- Balance between number of participants in each group is guaranteed during the course of randomization.
- If the trial is terminated before enrollment is completed, balance exists between the groups.
- If blocking factor is known and the trial is not double-blinded, assignment for the last person is known before they are randomized
  - Don't use block sizes of 2
  - Vary the block sizes
  - Separate randomization plan; not the protocol!
### Blocked Randomization Example

<table>
<thead>
<tr>
<th>Block Size</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>ABABBA</td>
</tr>
<tr>
<td>8</td>
<td>BAABBABA</td>
</tr>
<tr>
<td>8</td>
<td>BAABABAB</td>
</tr>
<tr>
<td>6</td>
<td>AABABB</td>
</tr>
</tbody>
</table>

### Stratified Randomization

- Measure prognostic factors either before or at the time of randomization.
- If a single factor is used, it is divided into two or more strata.
- If several factors are used, a strata is formed by selecting by one subgroup from each of them.
- The randomization process involves measuring the level of the selected factors for participants, determining which stratum each belongs and performing the randomization within that stratum.
- Within each stratum it can be simple or blocked.

### Stratified Randomization Example

<table>
<thead>
<tr>
<th>AGE</th>
<th>SEX</th>
<th>SMOKING HX</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-49 yr</td>
<td>Male</td>
<td>Current</td>
</tr>
<tr>
<td>50-59 yr</td>
<td>Female</td>
<td>Former</td>
</tr>
<tr>
<td>60-69 yr</td>
<td>Never</td>
<td></td>
</tr>
</tbody>
</table>
Stratified Randomization Example

- $3 \times 2 \times 3 = 18$ strata
- Strata 1: 40-49, Male, Current
- Strata 2: 40-49, Male, Former
- Strata 3: 40-49, Male, Never
- Strata 4: 40-49, Female, Current
- Strata 5: 40-49, Female, Former
- Strata 6: 40-49, Female, Never
- and so on .........

Randomization Review

- Careful attention to the method used.
- Need valid randomization
  - single center studies have an independent statistician not involved in the care of the participant
  - Multicenter studies have a coordinating center conduct randomization
- Assignments distributed:
  - Envelopes, telephone system, web based
- Assigned closest to time of beginning intervention

Envelope Randomization

<table>
<thead>
<tr>
<th>Randomization Sequence for ABC PHARM Protocol Number 000-01</th>
<th>Patient Number</th>
<th>Treatment Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1111</td>
<td>Treatment 1: 50 mg/hr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment 2: 100 mg/hr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment 3: Placebo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment 4: 200 mg/hr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment 5: 250 mg/hr</td>
</tr>
</tbody>
</table>
Double Blind label

Protocol Number: YYYY05
Manufacturer: QQ Pharm, Philadelphia, Pennsylvania, USA
Patient ID: 01
Contents: Amoxicillin or matched placebo tablets divided into 10 blister cards.
Lot Numbers: 0505001 - 0505002
Manufacturing Date: XX/2005
Expiration Date: XX/2007

Scratch-off Label

Central Randomization

- Study personnel calls or faxes central randomization site
  - Usually biostatistician or Data Coordinating Center
- Randomizing personnel verify eligibility and assign randomization code
  - Also work with central drug distribution center
**IVRS and Web Based**

- Study personnel calls or faxes central randomization site
  - Usually biostatistician or Data Coordinating Center
- Randomizing personnel verify eligibility and assign randomization code
  - Also work with central drug distribution center

**Example IWR**

![Example IWR](image1)

**Example IWR**

![Example IWR](image2)
Advantage of IVRS/IWR

- Maintain security and integrity of randomization sequence
  - More easily maintain blind
  - Limit randomization errors
  - Verify subject eligibility criteria
  - Provide inventory control and medication management
  - Create real-time subject tracking reports
  - Emergency code-breaking

Factors Influencing Mechanism of Randomization

- Phase, size, scope and purpose of trial
- Who is doing the randomization and where is it taking place – study coordinator, research pharmacist, field site, clinic
- Blinded or unblinded trial
- Type of medication or treatment – pill vs. IV, surgical procedure
- Point at which randomization occurs
- Cost and complexity

Masking
## Bias

- In any clinical trial, bias is one of the biggest concerns.
- Systematic error: difference between the true value and that actually obtained due to all causes other than sampling variability.
- Conscious or subconscious factors or both.
- To minimize one type of bias, we can mask assigned intervention, assessment, classification and evaluation of the response variables.

## Types of Trials

- Unblinded trials
- Single blind trial
- Double blind trial

## Unblinded Trials

- Open trial
- Participant and investigator know the intervention the participant has been assigned.
  - Exercise studies, eating habits
- Main concern is bias, reporting of symptoms and side effects, data collection.
Single Blind Trial

- Only the investigators are aware of the intervention.
- Knowledge of the intervention may help the investigator care for the patient.
- Participant bias is minimized, but potential of investigator bias in administration of non-study therapy, data collection and data assessment.

Double Blind Trial

- Neither participant nor investigator know the treatment assignment.
- Usually trials of drug efficacy.
- Risk of bias is reduced, any effects of actions theoretically would occur equally in the intervention and control groups.
- Certain functions of the study must be conducted by independent groups (toxicity and benefits).

Double Blind Trial

- More complex and several factors must be evaluated
- Matching medications
- Coding of medications
- Unblinding trials
- Assessment of blindness
Matching Medications

- Visual discrepancies: size, shape, smell, taste, color, sheen and texture
- Drug preparations should be pre-tested if possible.

Coding Medications

- Labeling of individual drug bottles, packets, etc.
- Assign a set of random numbers to active and different set to control.
- Smaller studies, participant should have unique drug code that stays with them for the trial.

Unblinding Trials

- Accidental unblinding
- Laboratory errors
- Monitoring use of study medication prescribed outside of the study.
- Official breaking of the blind
  - Procedures specifically defined
Assessment of Blindness

- Sometimes worthwhile to estimate the degree to which the blind was maintained.
- Patient guess
- Investigator guess
- Degree difference from 50%
- Study adherence versus what they thought they were taking

Sample Size and Power

Sample Size

- All clinical trials should have sufficient statistical power to detect clinical differences of interest between the groups.
- Sample size and power should be considered early in the planning phase.
- Sample size calculations provide only an estimate of the needed size of a study.
Sample Size Information

- Provides Information About:
  - How many subjects should participate in the research?
  - Is the study worth conducting if only n subjects participate?

Power

- The power of a study is its ability to detect a true difference in outcome between the control arm and the intervention arm.
- By definition, a study power set at 80% accepts a likelihood of one in five (that is, 20%) of missing such a real difference.
- Power for large trials is occasionally set at 90% to reduce to 10% the possibility of a so-called “false-negative” result.

Significance Level

- The chosen level of significance sets the likelihood of detecting a treatment effect when no effect exists (leading to a so-called “false-positive” result) and defines the threshold “P value”.
- Results with a P value above the threshold lead to the conclusion that an observed difference may be due to chance alone, while those with a P value below the threshold lead to rejecting chance and concluding that the intervention has a real effect.
**Significance Level**

- The level of significance is most commonly set at 5% (that is, \( p = 0.05 \)) or 1% (\( p = 0.01 \)). This means the investigator is prepared to accept a 5% (or 1%) chance of erroneously reporting a significant effect.

---

**Sample Size**

- Items to review:
  - What is the primary question you want to investigate? (Ex. Effectiveness of a treatment compared to a placebo – difference)
  - How is the question to be answered? How is the data collected and analyzed?
  - Statistician your new best friend!

---

**Sample Size and Power Calculations**

Survival Analysis: Power of 80% and (90%)

<table>
<thead>
<tr>
<th>Test significance level, ( \alpha )</th>
<th>0.050</th>
<th>0.050</th>
<th>0.050</th>
<th>0.050</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or 2 sided test?</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Group 1 proportion, ( p_1 )</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>Group 2 proportion, ( p_2 )</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>Malt ratios: ( q ) = ( \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} )</td>
<td>1.500</td>
<td>1.057</td>
<td>1.000</td>
<td>1.057</td>
</tr>
<tr>
<td>Power (%)</td>
<td>80</td>
<td>80</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>( \alpha ) per group</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Power (%)</td>
<td>80</td>
<td>80</td>
<td>85</td>
<td>85</td>
</tr>
</tbody>
</table>

Material copyrighted: Nicole C. Close, PhD
Sample Size and Power Calculations

<table>
<thead>
<tr>
<th>Survival Analysis: Power of 80% and (90%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two group t-test of equal means (equal n/c)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Test significance level, α</td>
</tr>
<tr>
<td>1 or 2 sided test?</td>
</tr>
<tr>
<td>Group 1 mean, μ₁</td>
</tr>
<tr>
<td>Group 2 mean, μ₂</td>
</tr>
<tr>
<td>Difference in means, μ₁ - μ₂</td>
</tr>
<tr>
<td>Common standard deviation, σ</td>
</tr>
<tr>
<td>Effect size, Δ = μ₁ - μ₂ / σ</td>
</tr>
<tr>
<td>Power (%)</td>
</tr>
<tr>
<td>n per group</td>
</tr>
</tbody>
</table>

Non Adherence

- The compliance adjustment formula is
  adjusted n per arm equals
  \[ N / [(c_1 + c_2 - 1)^2] \]
  where \( c_1 \) and \( c_2 \) are the average compliance rates per arm

  adjusted n = 100/[(0.8+0.8-1)^2] = 280

Documentation

- The sample size calculation should be described in sufficient detail to allow its use in other protocols.
- The power, level of significance and the control and intervention event rates should be clearly documented.
- Information on the scheduled duration of the study, any adjustment for non-compliance and any other issues that formed the basis of the sample size calculation should be included.
**Statistical Power**

- Power is of importance when:
  - We want to have a rationale basis for establishing sample sizes for a study.
  - When we don’t want to be in a position at the end of the study, if we find no difference between our intervention and comparison group and not being able to tell whether there truly isn’t a difference or if we didn’t have a big enough sample size to detect the “true” effect.

**Fundamental Points**

- Refers to the ability of a test statistic to detect a true difference between two or more groups.

- Influenced by several factors:
  - Difference between the intervention and comparison group for a specified outcome variable (effect size)
  - Variance of that outcome variable
  - Size of the sample

**Fundamental Points (cont)**

- We can reduce the possibility of our results coming from chance by eliminating bias in the study design by using techniques such as randomization, blinding, etc.

- However, another factor influences our results, the number of participants studied.
Statistical Power: The Basics

- Intuitively, we assume that the greater the proportion of the population we study, the closer to the truth we get. But how may do we need to study in order to get close to the right answer?
- Power and sample size is used by researchers to determine how many subjects are needed to answer the research question (the null hypothesis).

Example

- Thrombolysis in acute myocardial infarction (AMI).
  - Many years clinicians felt treatment would be of benefit, but successive studies didn’t prove this.
  - Not until the completion of a “mega-trial” that the small, but beneficial effect of thrombolysis was proved.

Example (cont)

- Thrombolysis compared to placebo:
  - Null hypothesis: no difference between the treatments in terms of mortality.
  - Alternative hypothesis: there is a difference between the treatment in terms of mortality.
- To determine if the two groups are the same, we can potentially make two kinds of error: Type I error and Type II error
Example (cont)

- Type I error: we reject the null hypothesis incorrectly.
- Type II error: when we accept the null hypothesis incorrectly.
- Power calculations tell us how many participants are required in order to avoid Type I and Type II errors.

Potential Errors

<table>
<thead>
<tr>
<th>Clinical Trial Result</th>
<th>Actual Truth: Treatment Benefit</th>
<th>Actual Truth: No Treatment Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Benefit</td>
<td>Correct Result</td>
<td>Type I error (alpha error)</td>
</tr>
<tr>
<td>No Treatment Benefit</td>
<td>Type II error (beta error)</td>
<td>Correct Result</td>
</tr>
</tbody>
</table>

What Affects Power?

- Precision and variance of measurements within any sample.
- Magnitude of a clinically significant difference.
- Importance of Type I or Type II errors for the study in question. (how to avoid!)
- Type of statistical test we are performing.
When should SS Calculations be Performed?

- Definitely, before the study, sometimes during and sometimes after.
- In designing the study, we want to make sure:
  - What we do is worthwhile so that we get a correct answer and in the most efficient way.
  - So we can recruit enough participants to give our results adequate power but not too any that we waste time getting more data than we need.
  - When designing we may have to make assumptions about desired effect size and variance within the data.

When should SS Calculations be Performed? (cont)

- Interim power calculations are done when original ones become suspect.
- Avoid premature stopping of a study or to avoid the prolongation of a study. But should only be conducted when stated a priori in the research design.
- Assessing trials of negative results, to make sure the study was not underpowered.

Statistical Tests

- Sample size calculations indicate how the statistical tests used in the study are likely to perform.
- Type of test used affects how the sample size is calculated.
- Parametric tests versus non-parametric tests (need more participants).
<table>
<thead>
<tr>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nicole C. Close, PhD</td>
</tr>
<tr>
<td><a href="mailto:nclose@empiristat.com">nclose@empiristat.com</a></td>
</tr>
<tr>
<td>240-744-0000</td>
</tr>
<tr>
<td><a href="http://www.empiristat.com">www.empiristat.com</a></td>
</tr>
</tbody>
</table>
Statistical Analysis

Presented by
Cora MacPherson, Ph.D.

Statistical Analysis

- Understanding your own
- Evaluating others

SAP: Statistical Analysis Plans

- Why are they important?
- When should they be written?
- What should they include?
- Spell out primary, secondary, and exploratory outcomes
- Details on assumptions used in Sample Size Calculation
- Describe baseline data that will be collected and how it will be compared across groups

Study Populations
- ITT: Intent to Treat
- Safety
- Per protocol
- Subgroups

Method of Analysis
- What statistical test will be used?
- What significance level?
- Will alpha be adjusted?
- Will there be interim analyses?
Multiple Comparisons

- How many hypotheses will you be testing?
- How will this impact the interpretation of the p-values in the study?

Adjustments for Multiple Comparisons

- Bonferroni Method
- Hochberg Procedure
- Hierarchical Testing
- Nominal p-values

Interim Analyses

- Safety
- Futility (Conditional Power)
- Sample Size adjustment
- Efficacy
Safety Data

- P-values can be interesting, but this data does not require strict 0.05 level to be of importance.
- Overall number of AEs vs. Number of Patients experiencing each AE
- Grade, Relationship to Study Drug

Tables, Listings, Figures

- To prevent the possibility of only presenting positive results, mock-ups of the tables, listings, and figures are created before the study results are known.

Study Results

Interpreting the Final Data
General Considerations

- Was the study randomized?
- Blinded?
- Multi-center vs. single center
- Sponsor: drug company vs. NIH

Baseline Assessment

- Usually “Table 1”
- Are groups comparable with respect to important baseline characteristics?

Sample Size

- Too small (power issues)?
- Too large (chance)?
Participant Adherence

- Compliance with treatment
- Drop-out rate

Issues in Data Analysis

- Who was analyzed (Intent-to-Treat)?
- Covariate adjustment
- Outliers
- Missing values
- Subgroup analysis

Interpreting Results

- True finding
- Chance
- Bias

>> Use five criteria to help…
Five Criteria

1. Strength of relationship
2. Consistency
3. Biological Plausibility
4. Dose-Response
5. Temporal

Miscellaneous Issues

- “Rule of Three”
- Adverse Effects
- Publication Bias

Meta-Analysis

- Formal process that uses statistical methods to combine data from many trials into a single analysis
- Hardest part: which trials to include?
- Many negative studies not published & can’t be included
Example: Celebrex
Sept 2000 JAMA publishes CLASS Study
  - Double-blind RCT
  - Specific, objective, pre-specified outcome reviewed by committee
  - Large sample size, multicenter
  - Intent to treat analysis. Subgroup analysis pre-specified.
  - Funded by Pharmacia

June 2002: BMJ Critique
  - Complete information provided to FDA contradicts conclusions of JAMA article.
  - Results actually from first six-months of two separate, longer trials
    - CLASS reported selective positive findings. Next six months of study found celebrex to have high AE.

Conclusion
  - What are the clinical implications?
  - Cost/Benefit
  - “Evidence based medicine”
Progression-Free Survival as a Predictor of Overall Survival in Men With Castrate-Resistant Prostate Cancer

Susan Halabi, Nicholas J. Vogelzang, San-San Ou, Kouros Owzar, Laura Archer, and Eric J. Small

ABSTRACT

Purpose
To explore whether progression-free survival (PFS) or biochemical PFS can be used as a predictor of overall survival (OS) and to investigate the dependence between PFS and OS in men with castrate-resistant prostate cancer.

Patients and Methods
Data from nine Cancer and Leukemia Group B trials that enrolled 1,296 men from 1991 to 2004 were pooled. Men were eligible if they had prostate cancer that had progressed during androgen deprivation therapy and did not receive prior treatment with chemotherapy, immunotherapy, or other nonhormonal therapy. Landmark analyses of PFS at 3 and 6 months from randomization/registration were performed to minimize lead time bias. The proportional hazards model was used to assess the significance effect of PFS rate at 3 and at 6 months in predicting OS. In addition, biochemical progression using the definitions of Prostate-Specific Antigen Working Group (PSAW) Criteria PSAWG1 and PSAWG2 were analyzed as time-dependent covariates in predicting OS.

Results
The median survival time among men who experienced progression at 3 months was 9.2 months (95% CI, 8.0 to 10.0 months) compared with 17.8 months in men who did not experience progression at 3 months (95% CI, 16.2 to 20.4 months; P < .0001). Compared with men who did not progress at 3 and at 6 months, the adjusted hazard ratios for death were 2.0 (95% CI, 1.7 to 2.4; P < .001) and 1.9 (95% CI, 1.6 to 2.4; P < .001) for men who experienced progression at 3 and 6 months, respectively. In addition, biochemical progression at 3 months predicted OS. The association between PFS and OS was 0.30 (95% confidence limits = 0.26, 0.32).

Conclusion
PFS at 3 and 6 months and biochemical progression at 3 months predict OS. These observations require prospective validation.

The United States Food and Drug Administration approved docetaxel for first-line chemotherapy for men with progressive castrate-resistant prostate cancer (CRPC) based on two trials that demonstrated an improvement in overall survival (OS).1,2 The only other United States Food and Drug Administration–approved agents for the treatment of men with CRPC are mitoxantrone and estramustine. This paucity of approved agents for a disease that claims the lives of nearly 29,000 American men each year is striking.3 There are many reasons for this relative lack of agents, but one clearly is the difficulty in designing clinical trials for this group of patients.

A major challenge in designing trials in men with CRPC revolves around defining appropriate end points for these trials. Although OS remains the regulatory gold standard, the relatively long median survival of approximately 20 months slows trial completion. Other challenges in designing trials besides using end points other than survival (ie, radiographic or biochemical response) include the fact that only 30% of the patients have measurable disease,4 bone progression is not an accurate and reliable measure of progression, and prostate-specific antigen (PSA) progression, a biochemical marker, is the most common type of progression. Whereas consensus definitions of progression exist,5,6 it is worthy noting that these are exactly that: criteria derived as a consensus process that have not been prospectively validated. Although improvement in survival and in patient symptoms are sufficient to demonstrate tangible clinical benefit, the former end point requires dauntingly large numbers of patients and the latter end point faces difficulties in terms of data collection and analysis if there is missing information. Therefore, there exists a great unmet need.


definitions of progression exist, it is worthy noting that these are exactly that: criteria derived as a consensus process that have not been prospectively validated. Although improvement in survival and in patient symptoms are sufficient to demonstrate tangible clinical benefit, the former end point requires dauntingly large numbers of patients and the latter end point faces difficulties in terms of data collection and analysis if there is missing information. Therefore, there exists a great unmet need.
for validated intermediate end points in clinical trials in men with CRPC.

In an effort to address this question, we undertook a pooled analysis of 1,296 men with metastatic CRPC who participated in multicenter trials conducted by the Cancer and Leukemia Group B (CALGB). In particular, we sought to (1) evaluate the effect of progression-free survival (PFS) in predicting OS, (2) explore whether biochemical PFS as defined by consensus criteria was in fact a predictor of OS, and (3) investigate the dependence between PFS and OS and biochemical PFS and OS.

PATIENTS AND METHODS

Study Population

Data from 1,296 men with CRPC who were treated on nine CALGB multi-institutional clinical trials from 1991 to 2004 were pooled. Men were eligible if they had prostate cancer that had progressed during androgen deprivation therapy and had not received prior treatment with chemotherapy, immunotherapy, or other nonhormonal therapy. In addition, an Eastern Cooperative Oncology Group performance status of 0 to 2 and adequate hematologic, renal, and hepatic functions were required. Each participant signed an institutional review board–approved informed consent document in accordance with federal and institutional guidelines. Details regarding these trials have been published elsewhere.7-15

End Points

The primary end point was OS, which was defined as the time from randomization/study entry to date of death from any cause. We evaluated the secondary end points of PFS and biochemical progression in predicting OS. PFS was defined uniformly in all CALGB protocols as the time of randomization/study entry to any progression (bone progression defined as two or more new lesions on bone scan, PSA progression using the PSA Working Group consensus criteria of 1999 [PSAWG1],8 or objective progression in lung, liver, nodes, or soft tissue disease) or death, whichever occurred first. PFS rate at 3 and 6 months was defined as a binary variable: a patient experiencing any type of progression at or before 3 months (or 6 months) was considered to have experienced an event. Otherwise, the patient was censored. The protocols required radiologic assessments every 8 to 12 weeks and PSA assessments every 3 to 4 weeks, although the extent of adherence to these requirements is not known.

In addition, we considered two definitions of biochemical progression: the PSAWG13 and PSAWG2.8 For both criteria, biochemical failure depended on whether patients experienced PSA declines of at least 50% or more. If they did, using PSAWG1, the PSA post-treatment measurement should be increased by 50% from nadir and PSA ≥ 5 ng/mL. Otherwise, progression was defined based on a 25% increase from nadir/post-treatment. Conversely, using the PSAWG2, a patient was considered as experiencing biochemical failure if their PSA post-treatment determination increased by 50% and PSA measurement ≥ 2 ng/mL. Both biochemical progression criteria required a second PSA confirmation at least 1 week later.

Data Analysis

Landmark analyses of PFS at 3 and 6 months from randomization/registration were performed to minimize lead time bias.16 We were interested in analyzing the PFS rates at 3 and 6 months for the following two reasons. First, the median time to progression in the CALGB database was approximately 3 months. Second, the two trials demonstrating a survival advantage with docetaxel–based chemotherapy reported that the median time to progression was approximately 6 months in men with CRPC. Ninety-five patients died before 3 months and were excluded from the landmark analysis. The sample of 1,201 men was randomly allocated into training and testing data sets, with a roughly 2:1 ratio: 781 patients (65%) and 420 patients (35%), respectively.

The Kaplan-Meier product-limit method was used to estimate the OS distribution by the PFS rate at 3 and 6 months and by the biochemical progression at 3 months.17 The proportional hazards model was used to assess the significance effect of PFS rate at 3 and at 6 months in predicting OS.18 The estimates of the PFS rates at 3 and 6 months were applied to the testing data set, and the overall misclassification error rates were computed. In addition, biochemical progression (PSAWG1 and PSAWG2) were analyzed at 3 and 6 months and as time-dependent covariates in predicting OS in the proportional hazards model. In these models, PSA progressions were considered as binary variables. Known prognostic variables4,19 including age, race, performance status, Gleason score (was assigned by local pathologists and was not centrally reviewed), hemoglobin, testosterone, PSA, alkaline phosphatase, lactate dehydrogenase (LDH), presence of visceral disease, prior treatment with radiation therapy, and years since diagnosis were included in the multivariable proportional hazards models. LDH, years since diagnosis, alkaline phosphatase, and PSA were modeled using the restricted cubic spline function as they had skewed distributions. Because not all of the protocols considered included docetaxel, the only known agent to prolong survival in CRPC, subgroup analyses of men who had received docetaxel every 3 weeks were performed.

The associations between OS and PFS and between OS and time to PSA progression were investigated using a statistic that estimates Kendall’s τ measure of association for bivariate time to event outcomes subject to censoring.20 The null sampling distribution of the test statistic was approximated using 2,000 permutation replicates. Along with the bootstrap bias and SE, an interval estimate (confidence limit [CI]) of the association parameter was constructed using adjusted bootstrap percentile CIs based on 2,000 permutation replicates.21 S-plus statistical software (version 8.0, Insightful Corp, Seattle, WA) and R were used for data analyses and all statistical tests were two-sided.

RESULTS

Baseline Characteristics

A total of 1,201 men enrolled onto CALGB clinical trials were included in this analysis. Their baseline clinical and laboratory characteristics are presented in Table 1. The median age at diagnosis was 71 years, and approximately 15% of the men were African American. Forty-four percent of men had a Gleason sum of 6 or higher, and 36% had measurable disease. Eighty-nine percent of men had bone metastases. The median LDH, PSA, and alkaline phosphatase levels were 221 ng/mL, 107 U/L, and 148 ng/dL, respectively.

With the exception of LDH, there were no statistical differences in baseline clinical and laboratory variables in the training data set compared with the testing data set. A total of 771 and 419 progression events were observed in the training and testing data sets, respectively. The median survival times were similar between the training and testing data sets and were equal to 13.3 months.

PFS Predicting OS

Progression by PSA is observed as the first progression event in 60% of the patients, bone progression in 18%, measurable disease progression in 7%, and death as the first event in the remaining 15%. The median survival time among men who experienced any PFS at 3 months was 9.2 months (95% CI, 8.0 to 10.0 months) compared with 17.8 months in those men who did not experience any PFS at 3 months (95% CI, 16.2 to 20.4 months; P < .0001). A similar pattern is observed in the testing data set. The median survival times were 8.9 months (95% CI, 7.6 to 10.5 months) and 17.9 months (95% CI, 16.0 to 20.6 months; P < .0001) in men who experienced and did not experience PFS at 3 months, respectively. The Kaplan-Meier product-limit survival curves by PFS at 3 months are presented in Figure 1.
In multivariable analysis, PFS at 3 months predicted OS (Table 2). Compared with men who did not experience disease progression at 3 months, the adjusted hazard ratio (HR) for death of men was 2.0 (95% CI, 1.7 to 2.4; \( P < .001 \)). Other statistically significant factors of OS were age, performance status, race, body mass index, hemoglobin, LDH, PSA, alkaline phosphatase, Gleason score, prior radiotherapy, and years since diagnosis.

In a subgroup analysis of 232 men who received docetaxel every 3 weeks, PFS at 3 months also predicted OS. The median survival times were 7.6 months (95% CI, 6.3 to 9.5 months) and 20.4 months (95% CI, 17.4 to 22.2 months) for men who experienced and did not experience disease progression at 3 months, respectively (\( P < .0001 \)). The adjusted HR for PFS at 3 months was 2.4 (95% CI, 1.3 to 4.2; \( P < .001 \)).

### PFS at 6 Months

The median survival time among men who experienced any type of progression at 6 months was 10.1 months (95% CI, 9.0 to 11.2 months) compared with 19.6 months (95% CI, 16.6 to 21.7 months; \( P < .0001 \)). In the testing data set, the median survival times were 9.3 months (95% CI,
PFS as Significant Predictor of Overall Survival in Men With CRPC

7.6 to 11.6 months) and 19.2 months (95% CI, 16.0 to 23.4 months; \( P < .0001 \); Appendix Fig A1, online only) in men who experienced and did not experience progression at 6 months, respectively.

PFS at 6 months also predicted OS. Compared with men who did not experience progression at 6 months, the adjusted HR for death was 1.9 (95% CI, 1.6 to 2.4; \( P < .0001 \)). The predicted PFS probability at 3 and 6 months was compared with the observed survival probability. The estimates of the PFS rates at 3 and 6 months were applied to the testing data set, and the misclassification error rates were estimated to be 0.27 and 0.25, respectively.

### Table 2. Multivariable Proportional Hazards Model of PFS at 3 Months Predicting Overall Survival Stratified On Study

<table>
<thead>
<tr>
<th>Factor</th>
<th>Hazard Ratio</th>
<th>95% CI</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any progression at 3 months, yes v no</td>
<td>2.00</td>
<td>1.69 to 2.36</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age*</td>
<td>1.13</td>
<td>1.02 to 1.26</td>
<td>.018</td>
</tr>
<tr>
<td>Performance status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 v 0</td>
<td>1.54</td>
<td>1.31 to 1.81</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2 v 0</td>
<td>2.07</td>
<td>1.53 to 2.80</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Presence of visceral disease, yes v no</td>
<td>1.22</td>
<td>0.96 to 1.56</td>
<td>.096</td>
</tr>
<tr>
<td>BMI†</td>
<td>0.87</td>
<td>0.82 to 0.96</td>
<td>.002</td>
</tr>
<tr>
<td>Gleason sum, 8-10 v 2-7</td>
<td>1.19</td>
<td>1.02 to 1.39</td>
<td>.031</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>0.91</td>
<td>0.86 to 0.97</td>
<td>.002</td>
</tr>
<tr>
<td>Testosterone*</td>
<td>1.01</td>
<td>0.99 to 1.02</td>
<td>.548</td>
</tr>
<tr>
<td>Race, African Americans v white and others</td>
<td>0.79</td>
<td>0.63 to 0.98</td>
<td>.035</td>
</tr>
<tr>
<td>Prior radiotherapy, yes v no</td>
<td>1.34</td>
<td>1.13 to 1.59</td>
<td>.001</td>
</tr>
<tr>
<td>Alkaline phosphatase‡</td>
<td>—</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td>Years since diagnosis‡</td>
<td>—</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>PSA‡</td>
<td>—</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>LDH‡</td>
<td>—</td>
<td>&lt;.001</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: PFS, progression-free survival; BMI, body mass index; PSA, prostate-specific antigen; LDH, lactate dehydrogenase.

*Hazard ratio based on 5 units change in the variable.
†Hazard ratio based on 10 units change in the variable.
‡Modeled as a restricted cubic spline.

Furthermore, in men who had received docetaxel, the median survival times were 9.8 months (95% CI, 6.2 to 12.5 months) and 20.3 months (95% CI, 17.7 to 22.7; \( P < .0001 \)) for men who did and did not experience progression at 6 months, respectively. The adjusted HR for OS for PFS at 6 months was 2.6 (95% CI, 1.7 to 4.0; \( P < .001 \)).

Biochemical Progression Predicting OS

The median number of PSA measurements was seven (interquartile range, three to 12). The median survival time among men who experienced biochemical failure progression using PSAWG1 at 3 months was 10.0 months (95% CI, 8.9 to 11.1 months) compared with 15.1 months (95% CI, 13.9 to 16.1 months; \( P < .0001 \); Fig 2) in men who did not experience biochemical failure progression at 3 months. Similar results were observed using PSAWG2 definition at 3 months.

In multivariable analysis, biochemical PFS predicted OS. Compared with men who did not experience biochemical progression at 3 months, the adjusted HR for death of men was 1.5 (95% CI, 1.3 to 1.7; \( P < .001 \)). Furthermore, the HR was 1.44 for biochemical progression using the PSAWG2 definition as a time-dependent covariate (95% CI, 1.28 to 1.62; \( P < .001 \), Table 3) when using the PSAWG2 biochemical progression criteria as a time-dependent covariate.

Similar statistically significant results were observed in men who received docetaxel every 3 weeks. The HRs for death were 1.94 (95% CI,
1.38 to 2.72; \( P < .0001 \) and 1.95 (95% CI, 1.39 to 2.74; \( P < .0001 \)) for biochemical progression using PSAWG1 and PSAWG2 definitions, respectively.

**Dependence Between PFS and OS**

The estimated association parameter for OS and PFS was 0.30 (bootstrap SE = 0.0172, 95% CL = 0.26, 0.32). There is strong evidence suggesting that OS and PFS are statistically associated \( (P < .00001) \). In addition, the estimated association parameter for biochemical PFS and OS was 0.30 (bootstrap SE = 0.017, 95% CL = 0.27, 0.33; \( P < .00001 \)). The results based on a complete-case analysis (uncensored cases) yielded similar results.

**DISCUSSION**

In this pooled analysis of 1,201 men with CRPC, we observed that PFS at 3 months predicts OS. The median survival time was statistically shorter among men who experienced any type of progression at 3 months (9.2 months) compared with men who did not experience progression (17.8 months; \( P < .0001 \)). Furthermore, in multivariable analysis, the HR for death for men who experienced any type of progression was 2.0 compared with men who did not experience progression. Similar results were observed in men who experienced progression by 6 months compared with men who did not experience progression by 6 months. Importantly, these observations were validated in a testing data set.

In this data set, progression by PSA is observed as the first progression event in 60% of the patients, bone progression in 18%, measurable disease progression in 7%, and death as the first event in the remaining 15%. Although PSA progression is by no means the only type of progression in these patients, biochemical progression was experienced first by a majority of patients. Consequently, we further investigated whether biochemical progression at 3 months predicted OS. The median survival times among men who experienced or did not experience biochemical progression at 3 months were statistically different (median, 10.0 ± 15.1 months, respectively).

Furthermore, using the PSAWG definitions of biochemical progression as time-dependent covariates, we have shown that biochemical progression is a statistically significant predictor of OS. The hazard for death for men who had experienced biochemical progression at any time during the trial represented a 50% increase compared with men who did not experience biochemical progression. It is gratifying that at least in this case, a consensus process yielded definitions of progression that could subsequently be validated.

CRPC is heterogeneous disease, and using PFS as a composite end point can be justified in study design to obtain higher event rates and therefore require smaller sample sizes. There are, however, many challenges to using PFS as an end point. The PFS end point requires careful review of its components. The inclusion of one component should be expected to be affected by the therapy, the component should be objective and should be measured accurately and reliably on all patients, it should be clearly defined and prespecified, and, finally, each component should be proven to be an important outcome and demonstrate clinical benefit to the patients.

PFS is usually defined as time to first biochemical, bone, or objective progression, or death. Although biochemical response (30% decline in PSA at 3 months) has been shown in retrospective analysis to be a surrogate of OS, \( ^{22,23} \) prospective validation is still needed. In addition, it is more difficult to establish surrogacy with other components of PFS. There is little agreement on the definition of bone progression. Some investigators assume one new lesion, \( ^{24,25} \) whereas others consider two lesions\( ^{26,27} \) as an indication of disease progression. Furthermore, there have been reports of variability in reading radiographic bone scans. \( ^{28} \) There are no studies that have demonstrated that bone progression predicts OS. Among the challenging factors is the lack of complete data on the individual components of the PFS end point once the patient has experienced progression. Consequently, it is important to collect all types of progression failures so that appropriate statistical methods for analysis can be implemented.

The lack of a standard definition for PFS end point in trials in men with CRPC has hampered the development and approval of novel drugs. Although there is a lack of consensus concerning the definition of progression, these CALGB studies have the advantage of using a standard definition of PFS over 13 years and across nine multicenter trials.

The relationship between PFS and OS and biochemical PFS and OS in men with CRPC remains unclear. However, in the majority of men with CRPC having elevated PSA at the time of study entry, CALGB investigators hypothesized that increasing PSA levels preceded and reflected growing cancer that would shortly be followed by symptomatic progression, progression in bone or soft tissue, or death. Assuming that the prostate cancer model is correct (Fig 3), CALGB investigators postulated that there is a strong dependence between PFS and OS end points. \( ^{29} \)

The results of this large retrospective analysis demonstrated strong evidence of dependence between PFS and OS and biochemical PFS and OS \( (P < .00001) \). These findings are similar to what was reported previously from a single institution. \( ^{30} \) Our estimate, however, is more representative of the patient population (multi-institution) and is more precise (more events observed) than that of Scher et al. \( ^{30} \)

A measure of concordance rather than a linear correlation was used to investigate the association between OS and PFS. It is therefore important to avoid a numerical comparison between the two types of measures. It may be helpful to consider the case of a Gaussian copula with parameter of \( \theta \); a Kendall’s \( \tau \) of 0.3 corresponds to \( \theta = 0.45 \) if the marginals are uniform on \([0,1]\). Although this retrospective analysis suggests that there is strong statistical association between PFS and OS and biochemical progression and OS, the clinical relevance of these associations has been questioned by regulatory bodies and agencies and needs to be prospectively validated.

While these data suggest that a composite PFS end point may serve as an intermediate end point for OS in phase II trials, this analysis has several limitations. First, it is not clear that the experience of men who enrolled on cooperative group trials is generalizable to the CRPC population, and these findings may not necessarily be applicable to noncytotoxic therapies.
PFS as Significant Predictor of Overall Survival in Men With CRPC

Nevertheless, there are some advantages to using these data in that uniform inclusion criteria, uniform monitoring, and uniform definitions of progression were used. A second limitation of the existing data set may be the lack of complete data that are available from each component of the PFS end point, making it difficult to draw inferences about the integrated composite end point. Finally, as noted, PSA progression alone clearly does not capture all progression events.

In summary, PFS and biochemical progression seem to be associated with OS. These data need to be validated prospectively before they can be used routinely as an intermediate end point in phase II trials in CRPC. Because there is no curative treatment for men with CRPC, it is vital to use standard definition of PFS and to standardize the frequency of assessment for all components of the PFS end point. A critical task remains, namely to develop and validate intermediate end points of OS to accelerate drug approval and to improve the survival of the 29,000 men who will die of prostate cancer in 2009.

REFERENCES


AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: Susan Halabi, Nicholas J. Vogelzang, Eric J. Small
Provision of study materials or patients: Nicholas J. Vogelzang
Collection and assembly of data: Eric J. Small
Data analysis and interpretation: Susan Halabi, Nicholas J. Vogelzang, San-San Ou, Kouros Owzar, Laura Archer
Manuscript writing: Susan Halabi, Kouros Owzar, Eric J. Small
Final approval of manuscript: Susan Halabi, Nicholas J. Vogelzang, San-San Ou, Kouros Owzar, Laura Archer, Eric J. Small

Acknowledgment

We thank Chen Jiang for her help in programming in R.
Prevention of Recurrent Preterm Delivery by 17 Alpha-Hydroxyprogesterone Caproate

Paul J. Meis, M.D., Mark Klebanoff, M.D., Elizabeth Thom, Ph.D., Mitchell P. Dombrowski, M.D., Baha Sibai, M.D., Atef H. Moawad, M.D., Catherine Y. Spong, M.D., John C. Hauth, M.D., Menachem Miodovnik, M.D., Michael W. Varner, M.D., Kenneth J. Leveno, M.D., Steve N. Caritis, M.D., Jay D. Iams, M.D., Ronald J. Wapner, M.D., Deborah Conway, M.D., Mary J. O’Sullivan, M.D., Marshall Carpenter, M.D., Brian Mercer, M.D., Susan M. Ramin, M.D., John M. Thorp, M.D., and Alan M. Peaceman, M.D., for the National Institute of Child Health and Human Development Maternal–Fetal Medicine Units Network*

ABSTRACT

BACKGROUND
Women who have had a spontaneous preterm delivery are at greatly increased risk for preterm delivery in subsequent pregnancies. The results of several small trials have suggested that 17 alpha-hydroxyprogesterone caproate (17P) may reduce the risk of preterm delivery.

METHODS
We conducted a double-blind, placebo-controlled trial involving pregnant women with a documented history of spontaneous preterm delivery. Women were enrolled at 19 clinical centers at 16 to 20 weeks of gestation and randomly assigned by a central data center, in a 2:1 ratio, to receive either weekly injections of 250 mg of 17P or weekly injections of an inert oil placebo; injections were continued until delivery or to 36 weeks of gestation. The primary outcome was preterm delivery before 37 weeks of gestation. Analysis was performed according to the intention-to-treat principle.

RESULTS
Base-line characteristics of the 310 women in the progesterone group and the 153 women in the placebo group were similar. Treatment with 17P significantly reduced the risk of delivery at less than 37 weeks of gestation (incidence, 36.3 percent in the progesterone group vs. 54.9 percent in the placebo group; relative risk, 0.66 [95 percent confidence interval, 0.54 to 0.81]), delivery at less than 35 weeks of gestation (incidence, 20.6 percent vs. 30.7 percent; relative risk, 0.67 [95 percent confidence interval, 0.48 to 0.93]), and delivery at less than 32 weeks of gestation (11.4 percent vs. 19.6 percent; relative risk, 0.58 [95 percent confidence interval, 0.37 to 0.91]). Infants of women treated with 17P had significantly lower rates of necrotizing enterocolitis, intraventricular hemorrhage, and need for supplemental oxygen.

CONCLUSIONS
Weekly injections of 17P resulted in a substantial reduction in the rate of recurrent preterm delivery among women who were at particularly high risk for preterm delivery and reduced the likelihood of several complications in their infants.

*Other members of the National Institute of Child Health and Development Maternal–Fetal Medicine Units Network are listed in the Appendix.
PRETERM DELIVERY — THAT IS, DELIVERY before 37 completed weeks of gestation — is the major determinant of infant mortality in developed countries. Preterm delivery is more common in the United States than in many other developed countries and is the factor most responsible for the relatively high infant mortality in this country. The rate of preterm delivery in the United States has increased progressively from 9 percent to 12 percent over the past two decades. Despite many trials of reduced activity, tocolytic therapy, antibiotic therapy, and other strategies for prevention, no effective and reproducible method of preventing preterm delivery has been demonstrated.

One treatment that showed promise in small trials was prophylactic treatment with progesterational compounds. Not all trials reported positive results. One meta-analysis found no evidence of effectiveness of progesterational compounds in the prevention of preterm delivery or the prevention of recurrent miscarriage. Another meta-analysis, restricted to trials of 17 alpha-hydroxyprogesterone caproate (17P), a natural metabolite of progesterone, showed, in composite, a significant reduction in the rate of preterm delivery. We therefore chose this pharmacologic agent as the active drug for our study.

Women who have had a preterm delivery are at especially high risk for preterm delivery in a subsequent pregnancy. We therefore conducted a multicenter trial to test the effectiveness of 17P as compared with placebo in the prevention of recurrent preterm delivery in this group of women.

METHODS

SUBJECTS AND SCREENING
Medical records of women presenting for prenatal care at the 19 participating centers were screened for eligibility to participate in the trial; criteria for eligibility included a history of spontaneous preterm delivery in a previous pregnancy and a current pregnancy between 15 weeks and 20 weeks 3 days of gestation. Reasons for exclusion were multifetal gestation, known fetal anomaly, progesterone or heparin treatment during the current pregnancy, current or planned cervical cerclage, hypertension requiring medication, a seizure disorder, or a plan to deliver elsewhere. An ultrasonographic examination was required between 14 weeks and 20 weeks 6 days of gestation to confirm the duration of gestation and to identify any major fetal anomalies. The duration of gestation at the time of randomization was determined according to a previously described algorithm on the basis of the last menstrual period and the results of ultrasonography.

Candidates for the trial were approached by a research nurse, who explained the study and asked prospective participants to sign a form for the release of medical records to permit the research nurse to obtain a copy of the chart from the previous pregnancy ending in preterm delivery. If the previous preterm delivery was of a liveborn singleton infant between 20 weeks of gestation and 36 weeks 6 days of gestation and was due to spontaneous preterm labor or preterm premature rupture of the fetal membranes, and if no criteria for exclusion were present, the woman was deemed to be eligible for the study. Each eligible woman was then invited to participate and to sign a consent form approved by the local institutional review board.

The trial started in April 1998 but was stopped in February 1999 because the Food and Drug Administration had ordered the pharmaceutical company that supplied the active study drug to shut down and had mandated a total recall of all the company’s drugs, including the study drug, because of poor quality control and documentation. Patient safety was not considered to have been compromised, but the potency of the product that had been supplied was thought to be questionable. At the time the study was stopped, 150 women had been enrolled, but none of the data had been analyzed. The trial was started anew with the study drug and placebo supplied by a company that manages investigation drugs (Eminent Services), and the data that had been collected previously were not included in the analyses.

RANDOMIZATION AND FOLLOW-UP VISITS
Consenting women were given a trial intramuscular injection of the inert oil placebo and asked to return in one week for randomization. If a woman did not return for a randomization visit between 16 weeks and 20 weeks 6 days of gestation, she was not permitted to participate in the trial. Returning eligible patients were then assigned to receive identically appearing active (17P) or placebo (castor oil) injections prepared by a research pharmacy. The women, their caregivers, and research personnel were not informed of the study-group assignment.

The boxes of 17P or placebo were packaged for each center according to a randomization sequence prepared by the George Washington University Bio-
The statistical Coordinating Center. The urn method of randomization, with stratification according to clinical center, was used to create the computer-generated randomization sequence. A 2:1 ratio was used for the assignment of women to 17P or to placebo, because it was known that patients assigned to placebo would be receiving painful injections on a weekly basis with no possibility of direct benefit.

After entering the study, the subjects returned for weekly injections of 17P or placebo given by a study nurse; the injections continued until 36 weeks of gestation or delivery, whichever occurred first. In addition to the weekly visits for study injections, the women received prenatal care at their institutions, as judged appropriate by their caregivers for their known level of risk of preterm delivery.

**Assessment of Outcome**

After delivery, study personnel reviewed all prenatal, delivery, newborn, and postpartum records and documented the date of delivery, birth weight of the infant, and neonatal course, as well as the occurrence of complications of pregnancy. Infants were followed until discharge from the hospital where they were born or, if they were transferred elsewhere, from the hospital to which they were transferred. Preterm delivery was defined as delivery at less than 37 completed weeks (259 days) of gestation, calculated as delineated above.

**Statistical Analysis**

The analysis was performed according to the intention-to-treat principle. Continuous variables were compared with the use of the Wilcoxon rank-sum test, and categorical variables were compared with the use of the chi-square or Fisher’s exact test (the latter when there was an expected value of less than five for any cell). Prolongation of pregnancy was assessed by life-table methods, with the duration considered being that between the time of randomization and the time a woman gave birth, was lost to follow-up, or reached 40 weeks of gestation, whichever came first. Curves for event-free survival were estimated with use of the Kaplan–Meier method, with adjustment to account for differing durations of gestation at entry, and were tested with the log-rank test.

On the basis of data from a previous study by the Maternal–Fetal Medicine Units Network, we estimated that 37 percent of the women in the placebo group would deliver before 37 weeks of gestation. With the use of this estimate, a total sample size of 500 women (334 in the progesterone group and 166 in the placebo group) was deemed to be sufficient for the detection of a reduction of 33 percent in the rate of preterm delivery (from 37 percent to 25 percent), under the assumptions of a type I error (two-sided) of 5 percent and a power of at least 80 percent. Before the study began, it was decided that the independent data and safety monitoring committee would use the group sequential method of Lan and DeMets, with a spending function for the type I error corresponding to the O’Brien–Fleming boundary, for interim monitoring and adjustment of the type I error. At the second interim analysis, conducted when 463 patients had undergone randomization, outcome data were available for 351 patients (70 percent of the planned sample). The boundary (P=0.015) for the test of significance of the primary outcome, preterm delivery, was found to have been crossed, and enrollment in the trial was halted.

**Results**

**Characteristics of the Women**

A total of 2980 women were identified as potentially eligible for the study on the basis of a review of medical records from September 1999 to February 2002. Of these women, 1039 were found to be eligible, and 463 eligible women gave consent for the trial and underwent random assignment to 17P or placebo. The main reasons for ineligibility included lack of documentation of the qualifying preterm delivery (in the cases of 549 women), a gestational age of more than 20 weeks (482 women), and current or planned cervical cerclage (241 women).

The characteristics of the 310 women in the progesterone group and the 153 women in the placebo group are shown in Table 1. The women in the two groups were similar in terms of the mean duration of gestation in the qualifying delivery, the mean duration of gestation at the time of randomization, race or ethnic group, marital status, body-mass index, educational level, smoking status, and substance use during pregnancy. The women in the placebo group had had more previous preterm deliveries (mean, 1.6 vs. 1.4; P=0.007).

**Compliance and Side Effects**

Noncompliance was defined by a gap of 10 days or more between any two injections. According to this definition, 91.5 percent of the women were compliant with all of their injections. There was no dif-

---

*Downloaded from www.nejm.org by Durlin Hickok on June 11, 2003. For personal use only. No other uses without permission. All rights reserved.*
ference in the rate of compliance between the two groups. A total of 231 women (50 percent) reported at least one adverse effect. The most common side effects were local injection-site reactions, including soreness (in 34.2 percent of the women), swelling (in 14.1 percent), itching (in 11.3 percent), and bruising (in 6.7 percent). More women in the progesterone group than in the placebo group had swelling at the injection site (17.2 percent vs. 7.8 percent, P=0.007) or a lump at the injection site (5.5 percent vs. 1.3 percent, P=0.03).

**Table 1. Characteristics of the 463 Women at Randomization.**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Progestrone Group (N=310)</th>
<th>Placebo Group (N=153)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of gestation at the time of qualifying delivery — wk</td>
<td>30.6±4.6</td>
<td>31.3±4.2</td>
</tr>
<tr>
<td>No. of previous preterm deliveries</td>
<td>1.4±0.7</td>
<td>1.6±0.9†</td>
</tr>
<tr>
<td>&gt;1 Previous preterm delivery — no. (%)</td>
<td>86 (27.7)</td>
<td>63 (41.2)</td>
</tr>
<tr>
<td>≥1 Previous term deliveries — no. (%)</td>
<td>153 (49.4)</td>
<td>71 (46.4)</td>
</tr>
<tr>
<td>Duration of gestation at randomization — wk</td>
<td>18.4±1.4</td>
<td>18.4±1.4</td>
</tr>
<tr>
<td>Age — yr</td>
<td>26.0±5.6</td>
<td>26.5±5.4</td>
</tr>
<tr>
<td>Race or ethnic group — no. (%)‡</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic black</td>
<td>183 (59.0)</td>
<td>90 (58.8)</td>
</tr>
<tr>
<td>Non-Hispanic white</td>
<td>79 (25.5)</td>
<td>34 (22.2)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>43 (13.9)</td>
<td>26 (17.0)</td>
</tr>
<tr>
<td>Asian</td>
<td>2 (0.6)</td>
<td>1 (0.7)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (1.0)</td>
<td>2 (1.3)</td>
</tr>
<tr>
<td>Marital status — no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married or living with partner</td>
<td>159 (51.3)</td>
<td>71 (46.4)</td>
</tr>
<tr>
<td>Never married</td>
<td>119 (38.4)</td>
<td>64 (41.8)</td>
</tr>
<tr>
<td>Divorced, widowed, or separated</td>
<td>32 (10.3)</td>
<td>18 (11.8)</td>
</tr>
<tr>
<td>Body-mass index before pregnancy§</td>
<td>26.9±7.9</td>
<td>26.0±7.0</td>
</tr>
<tr>
<td>Yr of education</td>
<td>11.7±2.3</td>
<td>11.9±2.3</td>
</tr>
<tr>
<td>Smoking during pregnancy — no. (%)</td>
<td>70 (22.6)</td>
<td>30 (19.6)</td>
</tr>
<tr>
<td>Alcohol use during pregnancy — no. (%)</td>
<td>27 (8.7)</td>
<td>10 (6.5)</td>
</tr>
<tr>
<td>Substance use during pregnancy — no. (%)</td>
<td>11 (3.5)</td>
<td>4 (2.6)</td>
</tr>
</tbody>
</table>

* Plus–minus values are means ±SD.
† P=0.007.
‡ Race was self-assigned by the women.
§ The body-mass index is the weight in kilograms divided by the square of the height in meters.

Primary Outcome and Preterm Delivery

Outcome data were available for 459 of the 463 women (99.1 percent) (Table 2). The frequency of delivery before 37 weeks of gestation was 36.3 percent in the progesterone group, as compared with 54.9 percent in the placebo group (P<0.001). Delivery before 35 weeks of gestation was also less frequent in the progesterone group (20.6 percent vs. 30.7 percent, P=0.02). There was a 42 percent reduction in the rate of delivery before 32 weeks of gestation in the progesterone group (11.4 percent vs. 19.6 percent, P=0.02). Rates of preterm delivery in the progesterone group did not differ according to the week of gestation at the time of qualifying delivery. Survival analysis showed a significant prolongation of pregnancy with 17P as compared with placebo (P=0.01). Because there was an imbalance between the progesterone and placebo groups with regard to the number of previous preterm deliveries, we performed an analysis with adjustment for this variable. The adjusted relative risk of delivery before 37 weeks of gestation in the 17P group as compared with the placebo group was 0.70 (95 percent confidence interval, 0.57 to 0.85). There were no significant differences between the two groups in the rates of hospital visits for preterm labor, use of tocolytic drugs, corticosteroid use, cesarean delivery, or chorioamnionitis (Table 2).

More than half the women enrolled were black. The reduction in the rate of preterm delivery with 17P among the black women was very similar to that among nonblack women (Table 2).

The effectiveness of 17P in this study suggests that only 5 to 6 women (95 percent confidence interval, 3.6 to 11.1) with a level of risk for preterm delivery similar to that among these women would need to be treated in order to prevent one preterm delivery before 37 weeks of gestation. Similarly, 12 women (95 percent confidence interval, 6.3 to 74.6) with a similar level of risk would need to be treated in order to prevent one delivery before 32 weeks of gestation.

Rates of spontaneous miscarriage between 16 weeks of gestation and 19 weeks 6 days of gestation, and rates of fetal death after 19 weeks 6 days of gestation are shown in Tables 2 and 3. There was a small and nonsignificant increase in the rate of miscarriages and stillbirths in the progesterone group as compared with the placebo group. With one exception, all stillbirths occurred before 24 weeks of gestation.

Outcomes among the Infants

There was a significant reduction in the risk of a birth weight of less than 2500 g in the progesterone group as compared with the placebo group (relative
risk, 0.66; P=0.003) and a nonsignificant reduction in the risk of a birth weight of less than 1500 g (relative risk, 0.62; P=0.08) (Table 3). Treatment with 17P led to significant reductions in the rates of necrotizing enterocolitis (P=0.01), need for supplemental oxygen, and intraventricular hemorrhage of any grade. However, there was no significant difference between groups in the rate of intraventricular hemorrhage of grade 3 to 4 specifically. The rates of infant death, transient tachypnea in the newborn, respiratory distress syndrome, bronchopulmonary dysplasia, need for ventilatory support, retinopathy of prematurity, and patent ductus arteriosus were slightly but not significantly lower in the progesterone group. Of the 17 neonatal deaths, 16 were due to complications of prematurity and 1 to intrapartum hypoxia subsequent to uterine rupture.

Nine of the infants were found to have congenital malformations (2.0 percent in each group). There was no consistent pattern to these defects, and none involved genital organs. One infant of a woman in the progesterone group had torsion of the testicles in utero, with subsequent infarction.

**Discussion**

Treatment with 17P on a weekly basis, beginning at 16 to 20 weeks of gestation and continued to delivery or 36 weeks of gestation, significantly reduced the rate of preterm delivery before 37 weeks, 35 weeks, and 32 weeks of gestation among women at high risk for preterm delivery. The rates of several complications of prematurity were correspondingly decreased among the infants of women assigned to this therapy.

The women enrolled in this study had high rates of preterm delivery, with more than 50 percent of the women who received the placebo injections delivering before 37 weeks of gestation. This high rate of preterm delivery is most likely related to the history of previous preterm deliveries. The earlier in a pregnancy a preterm delivery occurs, the greater the chance of preterm delivery in a subsequent pregnancy. In our study, the mean duration of gestation at the time of the qualifying delivery was 31 weeks, and a third of the women enrolled had had more than one previous preterm delivery. Therefore, the women in this study had particularly high risk. They were also strongly motivated, and compliance was excellent.

Preterm delivery has multiple causes. Some evidence suggests that the causes of early preterm delivery differ from those of later preterm delivery, with earlier preterm deliveries more often being related to infection. Whereas 17P would not be expected to affect an infectious process, in this study, it provided potent protection against early as well as later preterm delivery. The mechanisms of action of 17P in prolonging gestation are not entirely known. The actions of progesterone on the pregnant myometrium include relaxation of myometrial smooth muscle, blocking of the action of oxytocin, and inhibition of the formation of gap junctions. In sheep, goats, and some other mammals, a decrease in plasma progesterone and an increase in circulating estrogen precede the onset of labor. Although no such alteration in the ratio of plasma estrogen to progesterone precedes the onset of labor in primates, there is evidence that local changes in the progesterone level or the ratio of progesterone to estrogen in the placenta, decidua, or fetal membranes may be important in the ini-

<table>
<thead>
<tr>
<th>Table 2. Outcomes of Pregnancy According to Treatment Assignment.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Delivery before 37 wk of gestation</td>
</tr>
<tr>
<td>Spontaneous</td>
</tr>
<tr>
<td>Indicated because of complications</td>
</tr>
<tr>
<td>Black women</td>
</tr>
<tr>
<td>Nonblack women</td>
</tr>
<tr>
<td>Delivery before 35 wk of gestation</td>
</tr>
<tr>
<td>Delivery before 32 wk of gestation</td>
</tr>
<tr>
<td>Miscarriage at &lt;20 wk of gestation</td>
</tr>
<tr>
<td>Hospital visit for preterm labor</td>
</tr>
<tr>
<td>Tocolytic therapy</td>
</tr>
<tr>
<td>Corticosteroids for fetal lung maturity</td>
</tr>
<tr>
<td>Cesarean delivery</td>
</tr>
<tr>
<td>Chorioamnionitis</td>
</tr>
</tbody>
</table>

* Data on hospital visit for preterm labor were missing for 1 woman in the placebo group; data on tocolytic therapy were missing for 2 women in the placebo group; and data on corticosteroids for fetal lung maturity were missing for 14 women in the progesterone group and 1 woman in the placebo group. CI denotes confidence interval, and NA not applicable.
Transient tachypnea was defined by a birth weight of less than 1000 g and a requirement for oxygen therapy (fraction of inspired oxygen, >0.21) for the first 28 days of life. Intraventricular hemorrhage was graded according to the most severe radiologic finding before hospital discharge. Necrotizing enterocolitis was defined by the unequivocal presence of intramural air on abdominal radiography, perforation seen on radiography, or stricture formation after an episode of suspected necrotizing enterocolitis. Infants were recorded as having patent ductus arteriosus if treatment for patent ductus arteriosus was documented in the medical records. Retinopathy was diagnosed by ophthalmologic examination. Proven sepsis was defined by positive cultures of blood, cerebrospinal fluid, or urine on admission to the nursery or (in the absence of positive cultures) clinical evidence of cardiovascular collapse or an unequivocal radiograph confirming the presence of infection in an infant with a clinical diagnosis of sepsis. CI denotes confidence interval, and NA not applicable.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Progesterone Group (N=306)</th>
<th>Placebo Group (N=153)</th>
<th>Relative Risk (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetal death, antepartum or intrapartum</td>
<td>6/306 (2.0)</td>
<td>2/153 (1.3)</td>
<td>1.50 (0.31–7.34)</td>
</tr>
<tr>
<td>Birth weight &lt;2500 g</td>
<td>82/301 (27.2)</td>
<td>62/151 (41.1)</td>
<td>0.66 (0.51–0.87)</td>
</tr>
<tr>
<td>Neonatal death</td>
<td>8/306 (2.6)</td>
<td>9/153 (5.9)</td>
<td>0.44 (0.17–1.13)</td>
</tr>
<tr>
<td>Transient tachypnea</td>
<td>11/305 (3.6)</td>
<td>11/152 (7.2)</td>
<td>0.50 (0.22–1.12)</td>
</tr>
<tr>
<td>Respiratory distress syndrome</td>
<td>29/305 (9.5)</td>
<td>23/152 (15.1)</td>
<td>0.63 (0.38–1.05)</td>
</tr>
<tr>
<td>Bronchopulmonary dysplasia</td>
<td>4/305 (1.3)</td>
<td>5/152 (3.3)</td>
<td>0.40 (0.11–1.46)</td>
</tr>
<tr>
<td>Ventilatory support</td>
<td>26/303 (8.6)</td>
<td>22/151 (14.6)</td>
<td>0.59 (0.35–1.00)</td>
</tr>
<tr>
<td>Supplemental oxygen</td>
<td>45/303 (14.9)</td>
<td>36/151 (23.8)</td>
<td>0.62 (0.42–0.92)</td>
</tr>
<tr>
<td>Intraventricular hemorrhage Grade 3 or 4 Any grade</td>
<td>2/305 (0.7)</td>
<td>0/153</td>
<td>NA</td>
</tr>
<tr>
<td>Necrotizing enterocolitis</td>
<td>0/305</td>
<td>4/152 (2.6)</td>
<td>NA</td>
</tr>
<tr>
<td>Patent ductus arteriosus</td>
<td>7/305 (2.3)</td>
<td>8/151 (5.3)</td>
<td>0.43 (0.16–1.17)</td>
</tr>
<tr>
<td>Retinopathy</td>
<td>5/305 (1.6)</td>
<td>5/152 (3.3)</td>
<td>0.50 (0.15–1.70)</td>
</tr>
<tr>
<td>Proven sepsis</td>
<td>9/305 (3.0)</td>
<td>4/152 (2.6)</td>
<td>1.12 (0.35–3.58)</td>
</tr>
</tbody>
</table>

* Transient tachypnea was defined by a birth weight of less than 1000 g and a requirement for oxygen therapy, mechanical ventilation, or both during the first 24 hours of life in an infant in whom there was no evidence of other causes of respiratory distress. Respiratory distress syndrome was defined by a clinical diagnosis of type I respiratory distress syndrome and a requirement for oxygen therapy for at least 24 hours or by death before 24 hours in an infant who had received such a diagnosis and such therapy. Bronchopulmonary dysplasia was defined by a requirement for oxygen therapy (fraction of inspired oxygen, >0.21) for the first 28 days of life. Intraventricular hemorrhage was graded according to the most severe radiologic finding before hospital discharge. Necrotizing enterocolitis was defined by the unequivocal presence of intramural air on abdominal radiography, perforation seen on radiography, or stricture formation after an episode of suspected necrotizing enterocolitis. Infants were recorded as having patent ductus arteriosus if treatment for patent ductus arteriosus was documented in the medical records. Retinopathy was diagnosed by ophthalmologic examination. Proven sepsis was defined by positive cultures of blood, cerebrospinal fluid, or urine on admission to the nursery or (in the absence of positive cultures) clinical evidence of cardiovascular collapse or an unequivocal radiograph confirming the presence of infection in an infant with a clinical diagnosis of sepsis. CI denotes confidence interval, and NA not applicable.

We chose to use 17P because of reports of its effectiveness in some previous trials. Other studies showed no benefit, including a trial involving women with twin gestations and a trial in women with a low risk of preterm delivery. Most reported trials of other progesterone compounds have not demonstrated effectiveness in reducing the risk of preterm delivery. However, a recently reported trial in which progesterone suppositories were used suggested that this route of administration may be a viable alternative. The risk of preterm delivery was lower among participants in that study than among the women in our study. The entry criteria included a history of delivery before 37 weeks of gestation, cervical cerclage, or a uterine malformation. The women in the placebo group in that trial had a rate of preterm delivery of 28.5 percent as compared with 13.8 percent in the progesterone group. These results lend support to the concept of prophylactic use of progesterone to prevent preterm delivery.

Treatment with 17P also resulted in improved neonatal outcomes. Although the reduction in neonatal mortality in the progesterone group was not significant (relative risk, 0.44; P=0.08), the trial was not designed with sufficient power to address this end point adequately. There were significant reductions in the rates of necrotizing enterocolitis, any intraventricular hemorrhage, and the need for supplemental oxygen in the progesterone group.

17P appeared to be safe. There was no increase in the rate of congenital anomalies in the progesterone group. These results are consistent with surveys of the literature that have indicated an absence of teratogenic effects from the use of 17P during pregnancy.

The results of our trial should be interpreted with caution. Although 17P proved to be effective in preventing preterm delivery in our cohort of women at very high risk, it may not be effective in women with a lower risk of preterm delivery, and most preterm deliveries occur in women with no previous preterm delivery. Therefore, our results may not be generalizable to women whose risk factors for preterm delivery are different from those of the women in this trial. In addition, although 17P significantly reduced the rate of preterm delivery among the women who received it, the rate of preterm delivery in this group remained very high (36.3 percent). Thus, the identification of other causes of preterm delivery and other methods of preventing it remains a pressing need.

Supported by grants (HD27860, HD36801, HD27917, HD22141, HD27861, HD27869, HD27905, HD34208, HD34116, HD21410, HD27915, HD34136, HD34210, HD34122, HD40500, HD40544, HD34116, HD40560, HD40512, and MO1-RR-000080) from the National Institute of Child Health and Human Development.
APPENDIX

Other members of the National Institute of Child Health and Human Development Maternal–Fetal Medicine Units Network are as follows: University of Alabama, Birmingham: A. Northington, D. Rouse; Brown University: H. Silver, J. Tillinghast; Case Western Reserve University: P. Catalano, C. Milluzzi; University of Chicago: P. Jones, M. Lindheimer; University of Cincinnati: N. Elder, T. Siddiqi; Columbia University: M. D’Alton, V. Pemberton; George Washington University Biostatistics Center: A. Das, S. Leindecker; Mage Women’s Hospital: M. Cotroneo, K. Lain; University of Miami: C. Alfonso, S. Beydoun; National Institute of Child Health and Human Development: D. McNellis, S. Pagliaro, A. Willoughby; University of North Carolina, Chapel Hill: K. Dorman, K. Moise; Northwestern University: G. Mallet, M. Socol; Ohio State University: E. Johnson, M. Landon; University of Tennessee: R. Ramsey; University of Texas at San Antonio: O. Langer, S. Nicholson; University of Texas at Houston: M. C. Day, L. Gilstrap; University of Texas Southwestern Medical Center: J. McComblish, G. Wendeel; MCP Hahnemann University: M. DiVito, T. Tolosa; University of Utah: M. Belfort, E. Taggart; Wozz Forest University: E. Mueller-Heubach, M. Swain; Wayne State University: G. Norman, Y. Sorokin.

REFERENCES


Copyright © 2003 Massachusetts Medical Society.
A Trial of 17 Alpha-Hydroxyprogesterone Caproate to Prevent Prematurity in Twins

Dwight J. Rouse, M.D., Steve N. Caritis, M.D., Alan M. Peaceman, M.D., Anthony Sciscione, D.O., Elizabeth A. Thom, Ph.D., Catherine Y. Spong, M.D., Michael Varner, M.D., Fergal Malone, M.D., Jay D. Iams, M.D., Brian M. Mercer, M.D., John Thorp, M.D., Yoram Sorokin, M.D., Marshall Carpenter, M.D., Julie Lo, M.D., Susan Ramin, M.D., Margaret Harper, M.D., and Garland Anderson, M.D., for the National Institute of Child Health and Human Development Maternal–Fetal Medicine Units Network*

ABSTRACT

BACKGROUND
In singleton gestations, 17 alpha-hydroxyprogesterone caproate (17P) has been shown to reduce the rate of recurrent preterm birth. This study was undertaken to evaluate whether 17P would reduce the rate of preterm birth in twin gestations.

METHODS
We performed a randomized, double-blind, placebo-controlled trial in 14 centers. Healthy women with twin gestations were assigned to weekly intramuscular injections of 250 mg of 17P or matching placebo, starting at 16 to 20 weeks of gestation and ending at 35 weeks. The primary study outcome was delivery or fetal death before 35 weeks of gestation.

RESULTS
Six hundred sixty-one women were randomly assigned to treatment. Baseline demographic data were similar in the two study groups. Six women were lost to follow-up; data from 655 were analyzed (325 in the 17P group and 330 in the placebo group). Delivery or fetal death before 35 weeks occurred in 41.5% of pregnancies in the 17P group and 37.3% of those in the placebo group (relative risk, 1.1; 95% confidence interval [CI], 0.9 to 1.3). The rate of the prespecified composite outcome of serious adverse fetal or neonatal events was 20.2% in the 17P group and 18.0% in the placebo group (relative risk, 1.1; 95% CI, 0.9 to 1.5). Side effects of the injections were frequent in both groups, occurring in 65.9% and 64.4% of subjects, respectively (P=0.69), but were generally mild and limited to the injection site.

CONCLUSIONS
Treatment with 17 alpha-hydroxyprogesterone caproate did not reduce the rate of preterm birth in women with twin gestations. (ClinicalTrials.gov number, NCT00099164.)
Preterm Birth Is Responsible for a Substantial Portion of Infant Mortality and Persistent Disability. The problem of preterm birth has proved largely intractable. In 2004, 12.5% of all live-born infants in the United States were delivered preterm — that is, before 37 completed weeks of gestation. In a study published in 2003, weekly injections of 17α-hydroxyprogesterone caproate (17P) were shown to lower the risk of recurrent preterm birth by one third in women who had previously given birth to a preterm infant spontaneously. Although this finding is encouraging, only a minority of women destined to deliver preterm would qualify for 17P on the basis of having had a previous spontaneous preterm delivery.

It is unknown whether 17P can reduce the rate of preterm birth in women with other risk factors. One logical candidate risk factor is twin gestation. Twin gestations are increasingly common, and more than half result in premature births. Between 1980 and 2004, the rate of twin births rose dramatically in the United States (from 18.9 to 32.2 per 1000 live births), as did the absolute number (from 68,339 to 132,219). Consequently, the morbidity and mortality burden of twins is increasingly disproportionate and substantial: almost one in four very-low-birth-weight infants (below 1500 g) born in the United States are twins, as are one in six infants who die in the first month of life. Therefore, we evaluated the efficacy of 17P in preventing preterm birth in women with twin gestations.

Methods

Recruitment

Recruitment to this placebo-controlled, double-blind, randomized clinical trial of 17P for the prevention of preterm birth in twin gestations was undertaken at 14 sites from April 2004 through February 2006. Women carrying twins with a gestational age of at least 16 weeks and no more than 20 weeks 3 days were eligible. The exclusion criteria were serious fetal anomalies, spontaneous death of a fetus after 12 weeks, presumed monoamnionic placenta, suspected twin-to-twin transfusion syndrome, marked ultrasonographic growth discordance (a difference of at least 3 weeks of estimated gestational age between fetuses), planned nonstudy progesterone therapy after 16 weeks, in-place or planned cerclage, major uterine anomaly (e.g., bicornuate uterus), treatment with 10,000 or more units of unfractionated heparin per day, treatment with low-molecular-weight heparin at any dose, and major chronic medical diseases (e.g., insulin-requiring diabetes mellitus or pharmacologically treated hypertension). Twin gestations that were the result of intentional fetal reduction were also excluded.

An ultrasonographic examination was required between 12 weeks and 20 weeks 6 days of gestation to confirm the duration of gestation and to screen for major fetal anomalies. For women who conceived spontaneously, the duration of gestation was determined according to a previously described algorithm on the basis of the last menstrual period and the results of ultrasonography of the larger fetus. For women who conceived by in vitro fertilization, the duration of gestation was calculated on the basis of the date of embryo transfer and the age of the embryos when transferred.

The study was approved by the institutional review boards at each clinical site and at the data-coordinating center. All women gave written informed consent before enrollment in the study.

Protocol

Eligible, consenting women were given a trial intramuscular injection of the placebo and were scheduled for a randomization visit not later than 20 weeks 6 days of gestation. At that visit, women who still met none of the exclusion criteria were assigned to receive identical-appearing injections of active agent (250 mg of 17P) or placebo (castor oil) prepared by a research pharmacy. The simple urn method of randomization with stratification according to clinical center was used by the George Washington University Biostatistical Coordinating Center to create a randomization sequence for each center, and the boxes of 17P and placebo were packaged for each center according to the randomization sequences. The participating women, their caregivers, and the research personnel were unaware of the women's study-group assignments. After entering the study, the women returned for weekly injections through the end of the 34th week of gestation or until delivery, whichever occurred first. At each visit, they underwent systematic assessment for side effects. Otherwise, the women received usual clinical care.

After delivery, study personnel reviewed deliv-
ery, newborn, and postpartum records and documented the date of delivery, the birth weight of the infants, and the neonatal course, as well as the occurrence of complications of pregnancy and obstetrical interventions. The infants were followed until discharge from the hospital of birth, or, if they were transferred, until discharge from the transfer hospital.

**STUDY OUTCOMES**

The primary study outcome was a composite of delivery or fetal death before 35 completed weeks of gestation (245 days). Fetal death includes miscarriage, termination of pregnancy, and stillbirth. Prespecified secondary outcomes included the time from randomization to fetal death or delivery, a composite of serious adverse fetal or neonatal outcomes, and selected individual maternal and neonatal outcomes. In analyses that were not pre-specified, we assessed the proportion of preterm births in each group at different gestational-age thresholds.

The data were analyzed according to the intention-to-treat principle. The unit of analysis was the pregnancy, and if the outcome occurred in either fetus or neonate, the pregnancy was considered to have met the outcome. For example, a pregnancy in which one fetus died at 22 weeks and the other was born alive at 37 weeks would be counted as meeting the primary outcome. However, for fetal or neonatal death, a proportional-odds model, which permits the distinction between the death of one fetus or neonate and the death of both, was also used to compare the two groups.

**STATISTICAL ANALYSIS**

The time to delivery or death of the first fetus was compared between the two groups by means of a proportional-hazards model with left truncation (i.e., with adjustment for gestational age at entry). Similarly, survival curves were plotted with the use of a modified product-limit estimator. Continuous variables were compared with the Wilcoxon rank-sum test, and categorical variables with the chi-square test or Fisher’s exact test, as appropriate.

On the basis of data from two studies performed by the National Institute of Child Health and Human Development Maternal–Fetal Medicine Units Network, we estimated conservatively that in the placebo group, 35% of twins would be delivered before 35 completed weeks of gestation. Thus, a total sample size of 600 was deemed sufficient to detect a 33% reduction in the rate of delivery or death before 35 weeks, under the assumptions of a type I error (two-sided) of 5% and a power of greater than 80%. This sample size also yielded 70% power if it was assumed that there would be no reduction in indicated preterm delivery and that the one-third reduction would apply only to spontaneous preterm deliveries (on the assumption of a 4:1 ratio of spontaneous to indicated deliveries at less than 35 weeks of gestation).

We estimated that a sample size of 600 would provide at least 70% power to detect a 33% reduction in the rate of the composite of the following serious adverse fetal or neonatal outcomes: fetal or neonatal death, respiratory distress syn-
drome, grade 3 or 4 intraventricular hemorrhage, stage 2 or 3 necrotizing enterocolitis, periventricular leukomalacia, bronchopulmonary dysplasia, severe retinopathy of prematurity, or early-onset, culture-proven sepsis. On the basis of our previous data, we expected that at least one component of this outcome would occur in one or both fetuses or neonates in 25% of pregnancies.9

An independent data and safety monitoring committee monitored the trial and reviewed the interim results. Before the study started, the group sequential method of Lan and DeMets with the modified O’Brien–Fleming spending function was chosen for adjustment of the significance level in interim analyses.10 Two interim analyses were performed, and in the final analysis of the primary outcome, two-tailed $P$ values of less than 0.048 were chosen to indicate statistical significance. However, since the adjustment is minimal, 95% confidence intervals are reported. For all other outcomes, a nominal $P$ value of less than 0.05 was considered to indicate significance, and no adjustments were made for multiple comparisons.

**R E S U L T S**

We identified 1526 eligible women, of whom 699 (45.8%) gave consent and 661 (43.3%) were randomly assigned to treatment. Outcome data were available for 655 of these 661 women and for 1310 of 1322 fetuses or infants (Fig. 1). The baseline characteristics of the two study groups were similar (Table 1).

Compliance with the intervention was determined by the proportion of protocol-specified injections (one injection every 7 days from randomization to delivery or to 34 weeks 6 days of gestation, whichever occurred first) that were received. The mean compliance rate was 94.5% in the 17P group and 95.0% in the placebo group ($P=0.97$).

The rate of the primary outcome did not differ significantly between groups. Fetal death or delivery before 35 weeks of gestation occurred in 41.5% of pregnancies in the 17P group (325) and 37.3% of pregnancies in the placebo group (330) (relative risk, 1.1; 95% confidence interval [CI], 0.9 to 1.3). Of those deliveries occurring before 35 weeks of gestation, 72.8% were spontaneous and 27.2% were medically indicated; the proportions of spontaneous and medically indicated deliveries were similar between the groups (Table 2). The mean gestational age at delivery did not differ significantly between groups, nor did the proportion of deliveries occurring before 37 weeks (the standard threshold for preterm birth), 32 weeks, or 28 weeks (Table 2). From the time of randomization, the estimated distributions of time to first miscarriage, fetal death, or delivery for the two groups were very similar (Fig. 2). The rates of selected obstetrical interventions were similar between the groups (Table 2).

In 21 pregnancies, a fetus died; these included 12 pregnancies (18 fetuses) in the 17P group and 9 pregnancies (12 fetuses) in the placebo group (relative risk, 1.4; 95% CI, 0.6 to 3.2). In 37 pregnancies, a fetus or neonate died; these included

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>17P Group (N = 327)</th>
<th>Placebo Group (N = 334)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal age — yr</td>
<td>29.7±7.0</td>
<td>29.6±6.8</td>
</tr>
<tr>
<td>Gestational age at randomization — wk</td>
<td>19.2±1.5</td>
<td>19.2±1.4</td>
</tr>
<tr>
<td>Prepregnancy body-mass index†</td>
<td>26.7±6.5</td>
<td>27.1±7.1</td>
</tr>
<tr>
<td>Race — no. (%)‡</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>75 (22.9)</td>
<td>80 (24.0)</td>
</tr>
<tr>
<td>Hispanic or Latino ethnic background — no. (%)‡</td>
<td>51 (15.6)</td>
<td>54 (16.2)</td>
</tr>
<tr>
<td>Educational level — yr</td>
<td>13.6±2.8</td>
<td>13.6±2.9</td>
</tr>
<tr>
<td>Nulliparous — no. (%)</td>
<td>151 (46.2)</td>
<td>145 (43.4)</td>
</tr>
<tr>
<td>Spontaneous conception — no. (%)</td>
<td>204 (62.4)</td>
<td>226 (67.7)</td>
</tr>
<tr>
<td>Smoking during pregnancy — no. (%)</td>
<td>38 (11.6)</td>
<td>31 (9.3)</td>
</tr>
<tr>
<td>Alcohol use during pregnancy — no. (%)</td>
<td>29 (8.9)</td>
<td>19 (5.7)</td>
</tr>
</tbody>
</table>

* Plus–minus values are means ±SD. $P>0.05$ for all between-group comparisons.† The body-mass index is the weight in kilograms divided by the square of the height in meters.‡ Race and ethnic background were self-reported.
22 pregnancies (34 fetuses or neonates) in the 17P group and 15 pregnancies (22 fetuses or neonates) in the placebo group (relative risk, 1.5; 95% CI, 0.8 to 2.8). Using a proportional-odds model to account for both fetuses or neonates dying versus one dying versus neither dying yielded similar results. The rates of other selected neonatal outcomes, including major congenital malformations, were similar between the groups (Table 3).

The rate of the composite outcome of serious adverse events (fetal or neonatal death, respiratory distress syndrome, grade 3 or 4 intraventricular hemorrhage, periventricular leukomalacia, stage 2 or 3 necrotizing enterocolitis, bronchopulmonary dysplasia, severe retinopathy of prematurity, or early-onset, culture-proven sepsis) was 20.2% in the 17P group and 18.0% in the placebo group (relative risk, 1.1; 95% CI, 0.9 to 1.3).

Side effects of the injections were frequent in both the 17P and the placebo groups, occurring in 65.9% and 64.4% of subjects, respectively (P=0.69), but were generally mild and most often limited to the injection site (Table 2). Three women (two in the 17P group and one in the placebo group) discontinued injections because of side effects; the two women receiving 17P had intense injection-site reactions. In addition, one woman in the 17P group had subjective heart palpitations and presyncope immediately after the first injection. She did not return for any study follow-up and did not provide any outcome information.

### Table 2. Outcomes According to Treatment Group.*

<table>
<thead>
<tr>
<th>Outcome</th>
<th>17P Group (N=325)</th>
<th>Placebo Group (N=330)</th>
<th>Relative Risk (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery or fetal death at &lt;35 wk — no./total no. (%)</td>
<td>135/325 (41.5)</td>
<td>123/330 (37.3)</td>
<td>1.1 (0.9–1.3)</td>
</tr>
<tr>
<td>2 Live births</td>
<td>125/325 (38.5)</td>
<td>115/330 (34.8)</td>
<td>1.1 (0.9–1.4)</td>
</tr>
<tr>
<td>≥1 Fetal death</td>
<td>10/325 (3.1)</td>
<td>8/330 (2.4)</td>
<td>1.3 (0.5–3.2)</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>101/324 (31.2)</td>
<td>86/330 (26.1)</td>
<td>1.2 (0.9–1.5)</td>
</tr>
<tr>
<td>Medically indicated</td>
<td>33/324 (10.2)</td>
<td>37/330 (11.2)</td>
<td>0.9 (0.6–1.4)</td>
</tr>
<tr>
<td>Gestational age at delivery — wk</td>
<td>34.6±3.9</td>
<td>34.9±3.6</td>
<td></td>
</tr>
<tr>
<td>Gestational age at delivery or fetal death — no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;37 Wk</td>
<td>226 (69.5)</td>
<td>232 (70.3)</td>
<td>1.0 (0.9–1.1)</td>
</tr>
<tr>
<td>&lt;32 Wk</td>
<td>55 (16.9)</td>
<td>48 (14.5)</td>
<td>1.2 (0.8–1.7)</td>
</tr>
<tr>
<td>&lt;28 Wk</td>
<td>26 (8.0)</td>
<td>20 (6.1)</td>
<td>1.3 (0.8–2.3)</td>
</tr>
<tr>
<td>Tocolytic therapy — no./total no. (%)</td>
<td>71/324 (21.9)</td>
<td>97/330 (29.4)</td>
<td>0.7 (0.6–1.0)</td>
</tr>
<tr>
<td>Corticosteroid treatment for fetal maturation — no./total no. (%)</td>
<td>80/324 (24.7)</td>
<td>90/330 (27.3)</td>
<td>0.9 (0.7–1.2)</td>
</tr>
<tr>
<td>Cerclage placement — no./total no. (%)</td>
<td>6/324 (1.9)</td>
<td>4/330 (1.2)</td>
<td>1.5 (0.4–7.2)</td>
</tr>
<tr>
<td>Hypertensive disorder — no. (%)</td>
<td>66/320 (20.3)</td>
<td>55 (16.7)</td>
<td>1.2 (0.9–1.7)</td>
</tr>
<tr>
<td>Chorioamnionitis — no./total no. (%)</td>
<td>6/324 (1.9)</td>
<td>6/330 (1.8)</td>
<td>1.0 (0.3–3.1)</td>
</tr>
<tr>
<td>Cesarean delivery — no./total no. (%)</td>
<td>200/324 (61.7)</td>
<td>204/328 (62.2)</td>
<td>1.0 (0.9–1.1)</td>
</tr>
<tr>
<td>Side effects — no./total no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any</td>
<td>211/320 (65.9)</td>
<td>210/326 (64.4)</td>
<td>1.0 (0.9–1.1)</td>
</tr>
<tr>
<td>Injection site</td>
<td>197/320 (61.6)</td>
<td>203/326 (62.3)</td>
<td>1.0 (0.9–1.1)</td>
</tr>
<tr>
<td>Urticaria</td>
<td>11/320 (3.4)</td>
<td>4/326 (1.2)</td>
<td>2.8 (0.9–8.7)</td>
</tr>
<tr>
<td>Nausea</td>
<td>5/320 (1.6)</td>
<td>10/326 (3.1)</td>
<td>0.5 (0.2–1.5)</td>
</tr>
<tr>
<td>Other†</td>
<td>24/320 (7.5)</td>
<td>23/326 (7.1)</td>
<td>1.1 (0.6–1.8)</td>
</tr>
<tr>
<td>Leading to discontinuation of study drug</td>
<td>2/320 (0.6)</td>
<td>1/326 (0.3)</td>
<td>2.0 (0.3–27.5)</td>
</tr>
</tbody>
</table>

* Plus–minus values are means ±SD. Data on complications are based on clinical diagnoses in the medical records.
† The most common other side effects were fatigue, dizziness, and headache.
In this randomized trial conducted among women carrying twins, weekly 250-mg injections of 17P failed to lower the rate of preterm birth, to prolong gestation, or to improve fetal or neonatal outcome.

Our results are generalizable to most women in the United States who are pregnant with twins. The study subjects were drawn from a broad geographic area and were racially and ethnically diverse. Roughly half were nulliparous, and two thirds conceived spontaneously. Although most of the subjects were recruited from academic medical centers, they were not at unusually high risk for preterm birth. On average, the women in this trial delivered at 34.8 weeks, as compared with a national average of 35.2 weeks for women carrying twins. Our results may not be applicable to women carrying twins as a result of intentional fetal reduction from a higher-order multiple gestation, since these women were excluded from the trial.

In previous trials demonstrating a benefit of 17P, the study participants were carrying singletons, and most were at risk for spontaneous preterm birth because of preterm birth in a previous pregnancy. One previous trial conducted in women with twin gestations failed to find a benefit of 17P; however, this trial was statistically underpowered, with only 77 participants, and 17P was initiated much later in gestation than in our study, at a mean of 29 weeks. In a trial reported elsewhere in this issue of the Journal, progestrone treatment reduced the rate of preterm birth among women who were at high risk for preterm birth because of a short cervix. In that trial, which involved mostly women with singleton gestations, both the formulation (micronized progesterone) and the route of administration (vaginal) were different from those in our trial. Concurrently with our twins trial, we enrolled triplets in a companion trial, the results of which are currently being analyzed.

Potential limitations of our study should be noted. Because less than 10% of women in our trial had previously given birth prematurely, it is uncertain whether 17P might be of benefit in twin gestations in which the mother has a history of spontaneous preterm birth. Our choice of using the same dose of 17P (250 mg per week) that was used in our previous trial of 17P in women with singleton gestations might be questioned, because plasma volume is known to be approximately 20% greater in twin than in singleton gestations. Thus, it is possible that a larger dose of 17P might have been efficacious. However, unless there is a threshold effect for 17P, which has not previously been suggested, a lower-than-optimal dosage would have been expected to have an attenuated effect on preterm birth, rather than no effect.

We based our primary outcome on delivery or fetal death before 35 completed weeks of gestation, rather than on rates of adverse fetal or neonatal outcomes. Short- and long-term complications of preterm birth are a direct function of gestational age, and delivery before 35 weeks is a more stringent cutoff point for preterm birth than is the standard definition of delivery before 37 weeks. Moreover, in well-dated pregnancies (as in this trial), the assessment of gestational age at delivery is highly reliable and objective.

In summary, the results of our trial do not support the use of 17P to reduce the risk of preterm birth in twin gestations. Why 17P has been effective in women with singleton gestations and a history of spontaneous preterm birth but was not effective in the present trial in women carrying twins is a question that will be answered only when the mechanisms underlying preterm...
Table 3. Selected Neonatal Outcomes among Live-Born Infants According to Treatment Group.*

<table>
<thead>
<tr>
<th>Outcome</th>
<th>17P Group (N = 632)</th>
<th>Placebo Group (N = 648)</th>
<th>Relative Risk (95% CI)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weight‡</td>
<td>no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;2500 g</td>
<td>377 (60.0)</td>
<td>415 (64.0)</td>
<td>0.9 (0.8–1.0)</td>
</tr>
<tr>
<td>&lt;1500 g</td>
<td>81 (12.9)</td>
<td>64 (9.9)</td>
<td>2.0 (1.0–3.9)</td>
</tr>
<tr>
<td>Major malformation</td>
<td>3 (0.5)</td>
<td>4 (0.6)</td>
<td>0.5 (0.1–2.4)</td>
</tr>
<tr>
<td>5-Minute Apgar score &lt;7</td>
<td>27 (4.3)</td>
<td>33 (5.1)</td>
<td>0.9 (0.5–1.6)</td>
</tr>
<tr>
<td>Patent ductus arteriosus</td>
<td>18 (2.8)</td>
<td>31 (4.8)</td>
<td>0.7 (0.4–1.3)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>8 (1.3)</td>
<td>10 (1.5)</td>
<td>1.0 (0.4–2.7)</td>
</tr>
<tr>
<td>Mechanical ventilation</td>
<td>70 (11.1)</td>
<td>77 (11.9)</td>
<td>1.0 (0.7–1.5)</td>
</tr>
<tr>
<td>Seizures</td>
<td>5 (0.8)</td>
<td>5 (0.8)</td>
<td>1.3 (0.4–5.0)</td>
</tr>
</tbody>
</table>

Components of the composite outcome of serious adverse events§

| Severe retinopathy of prematurity            | 0                   | 0                       | —                     |
| Respiratory distress syndrome               | 96 (15.2)           | 87 (13.4)               | 1.2 (0.8–1.6)         |
| Early-onset, culture-proven sepsis          | 24 (3.8)            | 26 (4.0)                | 1.0 (0.6–1.9)         |
| Stage 2 or 3 necrotizing enterocolitis      | 3 (0.5)             | 4 (0.6)                 | 0.8 (0.1–3.0)         |
| Bronchopulmonary dysplasia                  | 19 (3.0)            | 17 (2.6)                | 1.2 (0.6–2.7)         |
| Grade 3 or 4 intraventricular hemorrhage    | 7 (1.1)             | 6 (0.9)                 | 1.0 (0.3–3.1)         |
| Periventricular leukomalacia                | 5 (0.8)             | 6 (0.9)                 | 0.9 (0.3–2.8)         |

* Data on complications are based on clinical diagnoses in the medical records.
† Relative risks were calculated according to the pregnancy, not the neonate (i.e., if either neonate had the outcome, the pregnancy was credited with the outcome).
‡ In the 17P group, birth weight was recorded for 628 infants.
§ The composite outcome of serious adverse events includes fetal or neonatal deaths as well as the listed individual components.

Birth and the actions of 17P are better understood. Further investigation is warranted to assess whether 17P is effective in other conditions in which the risk of preterm birth is increased.

Supported by grants (HD27869, HD21410, HD40512, HD34136, HD34208, HD40485, HD27915, HD40544, HD40560, HD27917, HD40590, HD34116, HD40545, HD27860, and HD36801) from the National Institute of Child Health and Human Development.

No potential conflict of interest relevant to this article was reported.

We thank Valerija Momirova for protocol and data management and statistical analysis; Allison Northen and Margaret Cotroneo for protocol development and coordination between clinical research centers; Jeffrey S.A. Stringer, M.D., of the University of Alabama at Birmingham for providing thoughtful manuscript review; and Joyce A. Martin, M.P.H., of the National Center for Health Statistics for graciously supplying data on U.S. natality and infant mortality.

APPENDIX


Copyright © 2007 Massachusetts Medical Society. All rights reserved.
REFERENCES


Copyright © 2007 Massachusetts Medical Society.