

Biost 524:
Design of Medical Studies

.....

Lecture 2:
Screening Studies

Susanne May, Ph.D. / Scott S. Emerson, M.D., Ph.D.
Associate Professor / Professor of Biostatistics
University of Washington

April 4, 2011

1

© 2011 Scott S. Emerson, M.D., Ph.D.

Summary from last class

.....

- Typical chronology
- Types of studies
 - Cause and effect
- Efficacy vs Effectiveness
- Scientific Hypotheses

2

Summary from last class

.....

- Typical chronology
- Types of studies
 - Cause and effect
- Efficacy vs Effectiveness
- Scientific Hypotheses

- Where do we want to be at the end?
- What design / data do we need to get there?
- If in doubt push for the most objective approach.
- If in doubt push for the “high road”.

3

Screening Studies

.....

Phases of Clinical Trials

Where am I going?

- Establishing the medical value of a new treatment proceeds through a series of investigations
- Goal:
 - understand the importance of ensuring safety / ethics
 - understand the value of staged accumulation of evidence.

4

Ultimate Goal

.....

- Indications for treatment
 - Disease
 - Population
 - Treatment strategy
 - Outcomes
- Evidence based medicine
 - Patients
 - Intervention
 - Comparators
 - Outcomes

5

Phases of Investigation

.....

- Preclinical
 - Epidemiology including risk factors
 - Basic science: Physiologic mechanisms
 - Animal experiments: Toxicology
- Clinical
 - Phase I: Initial safety / dose finding
 - Phase II: Preliminary efficacy / further safety
 - Phase III: Confirmatory efficacy / effectiveness
- Approval of indication based on total evidence to date
 - Evidence based medicine
 - (Phase IV: Post-marketing surveillance, REMS)



6

Phase III Clinical Trials

.....

- Confirmatory to establish efficacy / effectiveness
 - Goals:
 - Quantify treatment's effect on disease process
 - Incidence of major adverse effects
 - Therapeutic index
 - Modify clinical practice (obtain regulatory approval)
 - Methods
 - Relatively large number of participants from true target population (almost)
 - Clinically relevant outcome

7

Phase III Clinical Trials: Settings

.....

- Phase III: Common scenarios
 - Establish efficacy / effectiveness of new treatment
 - superiority over no intervention
 - superiority over existing treatment
 - Establish equivalence with current treatment
 - Two-sided equivalence: bioequivalence
 - establish response not markedly higher or lower
 - One-sided equivalence: noninferiority
 - establish treatment not markedly worse
 - perhaps superior on secondary endpoint
 - Establish harm of existing treatment

8

Why Emphasize Confirmatory Trials?

.....

“When you go looking for something specific, your chances of finding it are very bad, because of all the things in the world, you’re only looking for one of them.

“When you go looking for anything at all, your chances of finding it are very good, because of all the things in the world, you’re sure to find some of them.”

- Darryl Zero in “The Zero Effect”

9

Why Emphasize Confirmatory Trials?

.....

“When you go looking for something specific, your chances of finding [a spurious association by chance] are very bad, because of all the things in the world, you’re only looking for one of them.

“When you go looking for anything at all, your chances of finding [a spurious association by chance] are very good, because of all the things in the world, you’re sure to find some of them.”

10

Spurious associations – example

.....

- Random high bias
 - Fleming (2010, Ann Intern Med, 153, 400)
 - Exploratory analyses
 - Misleading *P values*
 - Overestimate of effect sizes

11

Spurious associations – example

.....

- Fleming (2010)
 - Nursery, 22 infants
 - 2 boys
 - 20 girls
- P-value: 0.0001

12

Spurious associations – example

.....

- Performing second study based on promising results
(not part of the Fleming article)
 - Maternity unit 1, 22 infants (2 boys, 20 girls)
 - Maternity unit 2, 22 infants
 - 11 boys
 - 11 girls

13

Spurious associations – example

.....

- Combining study data
 - Maternity unit 1, 22 infants (2 boys, 20 girls)
 - Maternity unit 2, 22 infants (11 boys, 11 girls)

Combined (13 vs 31)
P-value: 0.0096

Need to address the fact that these were two studies!
Even then, carrying the problem forward!

14

Multiple Comparisons in Biomedicine

.....

- Observational studies
 - Observe many outcomes
 - Observe many exposures
 - Consequently: Many apparent associations
- Interventional experiments
 - Rigorous science
 - Well defined methods and outcomes
 - Exploratory science (“Drug discovery”)
 - Modification of methods
 - Multiple endpoints
 - Restriction to subgroups

15

Mathematical Basis

.....

- The multiple comparison problem is traced to a well known fact of probability

$$\Pr(A \text{ or } B) \geq \Pr(A)$$

$$\Pr(A \text{ or } B) \geq \Pr(B)$$

16

Statistics and Game Theory

.....

- Multiple comparison issues
 - Type I error for each endpoint – subgroup combination
 - In absence of treatment effect, will still decide a benefit exists with probability, say, .025
- Multiple endpoints and subgroups increase the chance of deciding an ineffective treatment should be adopted
 - This problem exists with either frequentist or Bayesian criteria for evidence
 - The actual inflation of the type I error depends
 - the number of multiple comparisons, and
 - the correlation between the endpoints

17

Ex: Level 0.05 per Decision

.....

- Experiment-wise Error Rate

| Number Compared | Worst Case | Correlation | | | | |
|-----------------|------------|-------------|------|------|------|------|
| | | 0.00 | 0.30 | 0.50 | 0.75 | 0.90 |
| 1 | .050 | .050 | .050 | .050 | .050 | .050 |
| 2 | .100 | .098 | .095 | .090 | .081 | .070 |
| 3 | .150 | .143 | .137 | .126 | .104 | .084 |
| 5 | .250 | .226 | .208 | .184 | .138 | .101 |
| 10 | .500 | .401 | .353 | .284 | .193 | .127 |
| 20 | 1.000 | .642 | .540 | .420 | .258 | .154 |
| 50 | 1.000 | .923 | .806 | .624 | .353 | .193 |

18

Multiple comparisons

.....

From <www.causeweb.org>



"We test thousands of new treatments each year, so to avoid multiple testing issues we always do a validation experiment to confirm our positive results".

19

Multiple comparisons

.....

From <www.causeweb.org>



How often do those work out?

20

Multiple comparisons

.....

From <www.causeweb.org>



About 5% of the time!

21

Real-life Examples

.....

- Effects of arrhythmias post MI on survival
 - Observational studies: arrhythmias ⇔ high risk for death
 - CAST: anti-arrhythmics have higher mortality
- Effects of beta-carotene on lung CA and survival
 - Observational studies: high dietary beta carotene has lower cancer incidence and longer survival
 - CARET: beta carotene supplementation in smokers leads to higher lung CA incidence and lower survival
- Effects of hormone therapy on cardiac events
 - Observational studies: HT has lower cardiac morbidity and mortality
 - WHI: HT in post menopausal women leads to higher cardiac mortality

22

Need for Exploratory Science

.....

- Before we can do a large scale, confirmatory Phase III trial, we must have
 - A hypothesized treatment indication to confirm
 - Disease
 - Patient population
 - Treatment strategy
 - Outcome
 - Comfort with the safety / ethics of human experimentation
- In “drug discovery”, in particular, we will not have much experience with the intervention

23

Phase II Clinical Trials: Screening

.....

- Preliminary evidence of efficacy
 - Goals:
 - Screening for any evidence of treatment efficacy
 - Incidence of major adverse effects
 - Decide if worth studying in larger samples
 - Gain information about best chance to establish efficacy
 - » Choose population, treatment, outcomes

24

Phases of Investigation

.....

- Preclinical
 - Epidemiology including risk factors
 - Basic science: Physiologic mechanisms
 - Animal experiments: Toxicology
- Clinical
 - Phase I: Initial safety / dose finding
 - Phase II: Preliminary efficacy / further safety
 - Phase III: Confirmatory efficacy / effectiveness
- Approval of indication based on total evidence to date
 - Evidence based medicine
 - (Phase IV: Post-marketing surveillance, REMS)



25

Medical Studies as Diagnostic Tests

.....

- Clinical testing of a new treatment or preventive agent is analogous to using laboratory or clinical tests to diagnose a disease
 - Goal is to find a procedure that identifies truly beneficial interventions
- Not surprisingly, the issues that arise when screening for disease apply to clinical trials
 - Predictive value of a positive test is best when prevalence is high
 - Use screening trials to increase prevalence of beneficial treatments

26

Ex: Syphilis and VDRL

.....

- Conditional error based on diagnosis
 - False Positives: Pr (Pos among Healthy)
 - “Specificity” is 1 – False Positive rate
 - False Negatives: Pr (Neg among Diseased)
 - “Sensitivity” is 1 – False Negative rate
- Conditional error based on test result
 - Positive Predictive Value: Pr (Disease among Pos)
 - Negative Predictive Value: Pr (Healthy among Neg)

27

Definitions: Sens, Spec, PPV, NPV

.....

| | | Truth | | |
|------|----------|----------|---------|-----|
| | | Diseased | Healthy | |
| Test | Positive | a | b | a+b |
| | Negative | c | d | c+d |
| | | a+c | b+d | |

- False Positives: Pr (Pos among Healthy) = $b / (b+d)$
 - “Specificity” is 1 – False Positive rate = $d / (b+d)$
- False Negatives: Pr (Neg among Diseased) = $c / (a+c)$
 - “Sensitivity” is 1 – False Negative rate = $a / (a+c)$
- Pos Pred Value: Pr (Disease among Pos) = $a / (a+b)$
- Neg Pred Value: Pr (Healthy among Neg) = $d / (c+d)$

28

Diagnostic Testing

- We most often characterize the sensitivity and specificity of a diagnostic test by conditioning on disease status
 - Sensitivity of test: Positivity in diseased
 - Sample a group of subjects with the disease
 - Estimate the proportion who have a positive test result: $\Pr(+ | D)$
 - Specificity of test: Negativity in healthy
 - Sample a group of healthy subjects
 - Estimate the proportion who have a negative test result: $\Pr(- | H)$

29

Predictive Values

- We are actually interested in the diagnostic utility of the test, which conditions on test result
 - Predictive value of a positive test: Probability of disease when test is positive
 - $\Pr(D | +)$
 - Predictive value of a negative test: Probability of health when test is negative
 - $\Pr(H | -)$

30

Computing Predictive Values

- Bayes' Rule

$$\Pr(D | +) = \frac{\Pr(+ | D)\Pr(D)}{\Pr(+ | D)\Pr(D) + \Pr(+ | H)\Pr(H)}$$

$$\Pr(H | -) = \frac{\Pr(- | H)\Pr(H)}{\Pr(- | H)\Pr(H) + \Pr(- | D)\Pr(D)}$$

31

PV+: Relationship to Prevalence

- Need to know sensitivity, specificity, AND prevalence of disease

$$\Pr(D | +) = \frac{\Pr(+ | D)\Pr(D)}{\Pr(+ | D)\Pr(D) + \Pr(+ | H)\Pr(H)}$$

$$PVP = \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev)}$$

32

PV-: Relationship to Prevalence

.....

- Need to know sensitivity, specificity, AND prevalence of disease

$$\Pr(H | -) = \frac{\Pr(- | H)\Pr(H)}{\Pr(- | H)\Pr(H) + \Pr(- | D)\Pr(D)}$$

$$PVN = \frac{Spec \times (1 - Prev)}{Spec \times (1 - Prev) + (1 - Sens) \times Prev}$$

33

Ex: Syphilis and VDRL

.....

- Typical study: Sample by disease
 - Sensitivity of test: Probability of positive in diseased
 - 90% of subjects with syphilis test positive
 - (Actually depends on duration of infection)
 - Specificity of test: Probability of negative in healthy
 - 98% of subjects without syphilis test negative
 - (Actually depends on age and prevalence of certain other diseases)

34

Ex: PV+, PV- at STD Clinic

.....

- Ex: 1000 patients at STD clinic
 - Prevalence of syphilis 30%
 - PV+: 95% with positive VDRL have syphilis

| | | Syphilis | | Tot |
|-------|-----|----------|-----|------|
| | | Yes | No | |
| VDRL | Pos | 270 | 14 | 284 |
| | Neg | 30 | 686 | 716 |
| Total | | 300 | 700 | 1000 |

35

Ex: PV+, PV- in Marriage License

.....

- Ex: Screening for marriage license
 - Prevalence of syphilis 2%
 - PV+: 48% with positive VDRL have syphilis

| | | Syphilis | | Tot |
|-------|-----|----------|-----|------|
| | | Yes | No | |
| VDRL | Pos | 18 | 20 | 38 |
| | Neg | 2 | 960 | 962 |
| Total | | 20 | 980 | 1000 |

36

Diagnostic Tests

.....

- Questions?

37

Medical Studies as Diagnostic Tests

.....

- Frequentist testing as a diagnostic test
 - P value: Probability of observing positive (statistically significant) test in absence of true treatment effect
 - Level of significance is 1 - specificity
 - Common choice: $\alpha = .025$ means specificity is 97.5%
 - Statistical power: Probability of observing positive test in presence of true treatment effect
 - Power is sensitivity
 - Common choice: 80% sensitivity
 - (not usually recommended by me for Phase III studies)

38

Predictive Value of Medical Studies

.....

- Prevalence is the percentage of effective treatments among all tested treatments
 - How many of our ideas are “good”
- Positive predictive value is the probability that a statistically significant trial indicates a truly useful treatment

39

Preliminary Studies in Screening

.....

- In cancer less than 5% of treatments studied in clinical trials are adopted
 - NCI drug development program 1970 - 1985
 - 350,000 unique chemical structures studied
 - 83 pass preclinical and phase I testing
 - 24 pass phase II tests for biological activity

40

Preliminary Studies in Screening

- Two general approaches to studying new treatments
- Scenario 1:
 - Study every treatment in a large definitive experiment
 - Only do Phase III studies
 - Level of significance 0.025, high power
 - (Ignore, for now, the safety / ethics of this)
- Scenario 2:
 - Perform small screening trials, with confirmatory trials of promising treatments passing early tests
 - Phase II studies
 - Level of significance, power (sample size) to be determined
 - Confirmatory
 - Level of significance 0.025, high power

41

Scenario 1: Only Phase III

- Only large trials using 1,000,000 subjects
 - 10% of drugs being investigated truly work
 - Level of significance .025, .025, or 0.05
 - Sample size / power
 - 979 subjects, $\alpha=0.025$, 97.5% power \rightarrow 1,021 RCT
 - 500 subjects, $\alpha=0.025$, 80.0% power \rightarrow 2,000 RCT
 - 394 subjects, $\alpha=0.05$, 80.0% power \rightarrow 2,538 RCT
 - Results
 - N= 979: 99 effective / 23 ineffective (PV+ = .81)
 - N= 500: 160 effective / 45 ineffective (PV+ = .78)
 - N= 394: 202 effective / 114 ineffective (PV+ = .64)

42

Scenario 2a: Screening Phase II

- Use 700,000 subjects in Phase II studies
 - 10% of drugs being investigated truly work
 - Level of significance .025
 - Sample size / power
 - 100 subjects provide 24% power \rightarrow 7,000 RCT
 - Results
 - N= 100: 168 effective / 158 ineffective (PV+ = .52)

43

Scenario 2a: Confirmatory Phase III

- Use 300,000 subjects in confirmatory Phase III studies
 - 52% of drugs being investigated truly work
 - Level of significance .025
 - Sample size / power
 - 921 subjects provide 96.7% power \rightarrow 326 RCT
 - Results
 - N= 921: 162 effective / 4 ineffective (PV+ = .98)

44

Scenario 2a: Comparison

.....

- Scenario 1: Only large trials
 - Test 1 - 3K treatments (10% eff)
 - Power 0.975, $\alpha = 0.025$: 99 eff; 23 ineff (PV+ = .81)
 - Power 0.80 , $\alpha = 0.025$: 160 eff ; 45 ineff (PV+ = .78)
 - Power 0.80 , $\alpha = 0.05$: 202 eff ; 114 ineff (PV+ = .64)
- Scenario 2a: Use of pilot studies (70% N)
 - Screen 7,000 treatments (10% eff, $\alpha = 0.025$)
 - Power 0.240: 168 effective / 158 ineffective (PV+ = .52)
 - Test 326 treatments (52% eff, $\alpha = 0.025$)
 - Power 0.967: 162 effective / 4 ineffective (PV+ = .98)

45

Scenario 2b: Screening Phase II

.....

- Use 700,000 subjects in Phase II studies
 - 10% of drugs being investigated truly work
 - Level of significance .10
 - Sample size / power
 - 342 subjects provide 85% power → 2,047 RCT
 - Results
 - N= 342: 173 effective / 184 ineffective (PV+ = .49)

46

Scenario 2b: Confirmatory Phase III

.....

- Use 300,000 subjects in confirmatory Phase III studies
 - 49% of drugs being investigated truly work
 - Level of significance .025
 - Sample size / power
 - 839 subjects provide 95% power → 357 RCT
 - Results
 - N= 839: 165 effective / 5 ineffective (PV+ = .97)

47

Scenario 2b: Comparison

.....

- Scenario 1: Only large trials
 - Test 1 - 3K treatments (10% eff)
 - Power 0.975, $\alpha = 0.025$: 99 eff; 23 ineff (PV+ = .81)
 - Power 0.80 , $\alpha = 0.025$: 160 eff ; 45 ineff (PV+ = .78)
 - Power 0.80 , $\alpha = 0.05$: 202 eff ; 114 ineff (PV+ = .64)
- Scenario 2b: Use of pilot studies (70% N)
 - Screen 2,047 treatments (10% eff, $\alpha = 0.10$)
 - Power 0.850: 173 effective / 184 ineffective (PV+ = .49)
 - Test 357 treatments (49% eff, $\alpha = 0.025$)
 - Power 0.953: 165 effective / 5 ineffective (PV+ = .97)

48

Burden of Larger Studies?

.....

- Sequential sampling
 - Aggressive early stopping for futility
 - Greatest efficiency
 - Conservative early stopping for efficacy
 - Burden of proof, other endpoints

49

Screening Phase II: Bottom Line

.....

- Pilot studies increase the predictive value of a positive study while using the same number of subjects.
 - Screening parameters can be optimized
 - Proportion of subjects in Phase II vs Phase III
 - Type I error at Phase II
 - Power at Phase II
- But need to consider the number of RCT and the prevalence of effective treatments
 - Will we have same prevalence of “good” ideas when we have 1,000 RCT vs 7,000 RCT?

50

Phase II Clinical Trials: Methods

.....

- As typically implemented, the screening role of Phase II RCT is somewhat attenuated
 - Disease definition may be restrictive due to desires for
 - Efficacy: to demonstrate proof of concept
 - Safety / ethics: Caution with the unknown treatment in extremely serious disease (e.g., cancer)
 - Participants may not be true target population
 - Heavier trial burden in early trials
 - Unknown impact of concomitant disease
 - Outcome often a surrogate
 - Reduce costs / duration of RCT
 - Plausibility of effect on clinical outcome

51

Phase II Clinical Trials: Results

.....

- The design of confirmatory Phase III trials often uses exploratory analyses from Phase II trials
 - More knowledge is now available
 - Lessened safety / ethical concerns
 - Less need for heavy trial burden
 - Exploratory subgroup analyses
 - More / less severe disease
 - Restrictions based on concomitant disease
 - Modifications of treatment intensity, ancillary treatments
 - Try for optimal definition of clinical outcome, summary measures, statistical tests

52

Seamless Phase II / III Clinical Trials

.....

- Much recent interest in moving quickly from Phase II to Phase III studies
 - Plan a Phase III study
 - Incorporate early analysis of data to assess whether complete study
 - Early analysis may be based on different endpoint
 - Comments
 - Major gain is related to timeline of accrual
 - Only feasible if no changes based on early results
 - Same eligibility, treatment, measurement of outcomes
 - Blurs role of screening and confirmation
 - Should early phase data be included in analysis?
 - » Valid only if no changes to protocol

53

Phase II/III Screening Trials

.....

- Questions?

54

Phases of Investigation

.....

- Preclinical
 - Epidemiology including risk factors
 - Basic science: Physiologic mechanisms
 - Animal experiments: Toxicology
 - Clinical
 - Phase I: Initial safety / dose finding
 - Phase II: Preliminary efficacy / further safety
 - Phase III: Confirmatory efficacy / effectiveness
 - Approval of indication based on total evidence to date
 - Evidence based medicine
 - (Phase IV: Post-marketing surveillance, REMS)
- 

55

Phase I Clinical Trials

.....

- Initial safety / dose finding in humans
 - Goals:
 - Pharmacokinetics / pharmacodynamics
 - Pharmacokinetics: Absorption, excretion
 - Pharmacodynamics: Dose response of activity
 - Incidence of major adverse effects
 - Decide whether it is ethical to continue testing in humans
 - Methods
 - Participants often not true target population
 - Sometimes dose escalation

56

Preclinical Experiments

.....

- Prior to “first in human testing”
 - Laboratory evidence of plausibility
 - Effects on cell lines
 - Animal evidence of
 - Efficacy, mechanism of action
 - Pharmacokinetics
 - Pharmacodynamics
 - Toxicology / teratology

57

Phase IV Clinical Trials

.....

- Therapeutic: Post-marketing surveillance
 - Monitor for rare serious events
 - No matter how large a Phase III trial is performed, far larger, more diverse populations will ultimately be exposed to the treatment
 - Risk Evaluation and Mitigation Strategies (REMS)
- Prevention: Effectiveness
- Diagnostic tests: Outcomes
 - Does disease screening lead to improved survival or quality of life?

58