Workshop Goals

- To give an introduction to several topics arising in the design/analysis or interpretation of clinical trials and other medical studies, including:
  - choice of endpoints
  - surrogate markers
  - Interim monitoring of a trial's results
  - subset analyses
  - measures of association and other ways of describing the value of an intervention
  - patient population and generalizability of results
- Non-mathematical focus, reliance on examples
- Many details avoided due to time constraints, but reference list provided

4 Sessions

- I: Today, 16:00-16:50: Selection of endpoints, surrogate markers
- II: Today, 17:00-17:50: Interim monitoring, multiple analyses, subset analyses
- III: Tomorrow: 9:30-10:30: Guidelines for publishing results, reading the literature, relevance of baseline characteristics, absolute- versus relative-risk reduction, confidence intervals
- IV: Tomorrowy, 10:45-12:00: Applications to clinical practice, including numbers needed to treat.

Endpoints

- We will consider the choice and interpretation of study ‘endpoints’ in the context of a randomized clinical trial (RCT), and thus where treatment groups are being compared

- However, most results apply to the issue of selecting endpoints for other types of biomedical investigation, including: cross-sectional studies, case-control studies, and observational cohort studies

Some Principles in Design of Clinical Trials

- Specify a patient population and hypotheses
- Standardize diagnostic, staging, and follow up procedures
- Use relevant endpoints and pick a sample size that leads to appropriate power
- Attempt to enroll every eligible patient
- Randomly assign study treatment regimens to enrolled patients
- Treat and follow/evaluate every patient according to the study protocol
- Include all randomized patients in analysis
- Use planned analyses to draw conclusions about study hypotheses

Our focus: selection of endpoints

- For simplicity, assume a simple RCT comparing two treatments, A vs. B
  - e.g. A=placebo or standard treatment
    B=new treatment
- Want to know if treatment B is superior to A
  - If so, then B may become standard of care

(in some settings, an equivalence/noninferiority design is more appropriate. However, we focus on a superiority design for simplicity. Guidelines for endpoints similar for both)

Objectives and Endpoints

– The primary objectives of a clinical trial must always be reduced to measurable endpoints and quantifiable hypotheses:

e.g.:

- To reduce the mortality rate
- To reduce the incidence of side effects
- To reduce the level of a measured variable, such as blood pressure

Many elements of judgement arise in defining endpoints, including

- clinical relevance
- the time period patients are evaluated (e.g. 30 day mortality)
- the list of events to be included in the endpoint (e.g, ‘AIDS’ means any of several conditions).
- the procedures for measuring endpoints (e.g., assays & methodologies used)
- The subjectivity/objectivity of the endpoint and possibility that it can be evaluated in a consistent and reproducible way among study subjects

Examples

- Perioperative morbidity requires a definition of the period at risk, the events considered to represent morbidity, and the methods of measurement
- Improved Survival might mean increased median survival, higher five-year survival, or a lower death rate in the first year (protocol needs to be clear about which is meant)

- => important to be as specific as possible before conducting the study

Number of Endpoints

- Trials frequently have a single primary hypothesis and related endpoint. Study design can depart from this principle with a scientific rationale
- Rationale for a single (or sometimes 2) endpoints:
  - Inflated Type I error with multiple statistical tests (if no adjustment). More later…
  - If many primary endpoints, necessary adjustments for Type I error can make it very difficult to detect real differences (i.e., poor power)
- Common to specify 1 or 2 primary endpoints and several secondary endpoints. Understanding is that overall assessment of study results will be based mainly on the primary endpoint

Hard Endpoints

- One canon of design is to prefer “hard” endpoints, that is, endpoints that are well defined and can be measured without observer judgement
  - death is completely objective
  - Events such as recurrent MI, IQ scores, lab test results are somewhat less objective but still considered ‘hard’
- When the endpoints involve judgement, one must be concerned about assessment bias
  - E.g. assessments of pain relief, quality of life
  - blinding can be a critical design strategy

Types of Endpoints

- Measurements: e.g., blood pressure
- Binary or dichotomous: MI within 30 days
- Nominal: The outcome belongs to one of several unordered categories (histology)
- Ordinal: The outcome belongs to one of several ordered categories (e.g., Killip class for severity of angina)
- Counts: Number of Skin Lesions
- Survival Time or Time to Event: The endpoint in many studies of life-threatening disease

Measurements

- Useful when treatment raises or lowers the average value
- Measure of effect is most often the difference in means/medians before and after treatment; sometimes baseline variability is controlled for in other ways
- Statistical methods include t-tests, nonparametric tests (e.g., Wilcoxon test), and linear regression for multivariate analysis, ANOVA

Dichotomies (binary outcome)

- Commonly-used endpoint when the outcome is presence or absence of a condition or event at some fixed time
- Different measures of treatment effect
  - difference in proportions: p_1-p_2
  - relative risk: p_1/p_2
  - odds ratio: $[p_1(1-p_1)] / [p_2/(1-p_2)]$
- Methods for analysis include χ^2 tests, exact tests, and logistic regression
- Converting measurements to dichotomies common, but can lose information
  - e.g., actual blood pressure => 'high' versus 'low'

Nominal or Unordered Categories

- The possible outcomes may not be ordered
  - chronic graft vs host disease, acute GVHD, or no disease
- Most common analytic strategy is to use multiple indicator variables, usually comparing a category to some referent category.

Ordered Categories

- Commonly used for severity or toxicity
  - E.g. 0 (none), 1(mild), …., 5 (lethal)
- Analysis as a measured outcome, e.g., by calculating mean severity score, assumes that the categories are properly scaled
- Methods for analysis of ordered categories are specialized, appropriate for limited circumstances
- Proportional odds models can be useful for multivariate analysis

Counts

- Not commonly encountered in clinical trials
  - Example: number of ‘falls’ among older women being treated for osteoporosis; # seizures
- Multiple occurrences sometimes reduced to a binary outcome (e.g., ‘none’ versus ‘one or more’ falls)
- Effect often measured by relative rate
- Methods such as Poisson regression are available when the counts are of interest

Time to Event Data

- A major advance in methods for comparing treatments for chronic disease (e.g., time to recurrent cancer, time to death, etc.)
- Treatment effect often quantified by the relative hazard rate (relative risk)
- Most common methods of comparative analysis are log-rank test for univariate analysis and Cox's proportional hazards regression
- Usually, some observations are right- censored or sometimes interval censored

Repeated Measures
(a multivariate endpoint)

- Many studies require repeated measurement of outcome variables
  - CD4 count and viral load in HIV studies
- The repeated measurements create options for defining the primary endpoint:
  - The last measurement
  - The average of the last several measurements
  - The trend in the measurement over time

Repeated Measurements

- For outcomes such as bone loss, growth curve analysis can be a powerful analytic approach
- When the last or another particular measurement has special biologic significance, we may use it as the endpoint (e.g. Area Under Curve, slope, last value, change from first to last value)
- Adjusting for the baseline value can reduce variance substantially

Quality of Life
(a subjective endpoint)

- Example: measurements of well-being in a study of a terminal disease; improved cognition in patients with dementia
- Several ‘scales’ have been derived for different diseases
- Due to its subjective nature, sometimes used as a secondary efficacy endpoint

Recent examples where choice of endpoint is complicated and/or leads to controversy

1. Prostatectomy versus Watchful Waiting
   (Holmberg et al, NEJM, 2002, 347:781-9)

2. Low fat versus low-carb diet
   (Samaha et al, NEJM, 2003,348:2074-81)

3. Prostate Cancer prevention trial
   (Thompson et al, NEJM, 2003,349:215-224)

Prostatectomy vs Watchful Waiting

- Population: men with newly-diagnosed early stage prostate cancer (T1b,T1c,T2)
- Treatments: Radical prostatectomy versus 'watchful waiting' (scheduled treatment upon progression)
- Main Endpoint: Death due to Prostate Ca
- Secondary Endpoints: Overall mortality, local progression, metastatic-free survival
- Sample Size (med. follow up): N=695 (6.2 y)

Main Endpoint: Death due to Prostate Ca

- Prostatectomy Group: 16 (n=349):
- Watchful Waiting group 31 (n=349):

p=0.02 RR=.50 (95% CI: .27-.91)

A significant result !!

Prostatectomy vs Watchful Waiting: some additional concerns

- Main endpoint: Was cause of death (prostate ca vs other) accurately diagnosed ?
  - A concern in design of study; several steps taken to ensure uniform diagnosis. But can never be certain.
- Deaths attributable to other causes?
  - Real effects and/or misclassification of cause
- QOL & morbidity associated with each arm?
  - Some undesirable side effects of prostate cancer and surgery

Prostatectomy vs Watchful Waiting:

- Deaths from other causes:
  - Prostatectomy: 37
  - Watchful Waiting: 31
- Overall mortality:
  - Prostatectomy: 53
  - Watchful waiting: 62 p=0.31

Note: can't be sure whether higher number of deaths from other causes in prostatectomy group is due to surgery of misclassification of cause of death. Overall mortality avoids classification error and also gives the 'big picture'

Prostatectomy vs Watchful Waiting

- What is the message to future patients?
  - Prostatectomy not shown to reduce overall mortality
  - Differential morbidity:
    - Pros:Erectile dysfunction, urinary leakage
    - WW: obstructed voiding, fecal leakage
  - Study done before routine use of PSA:

    Would ‘watchful waiting’ results be better if PSA were monitored?

Low Fat versus Low Carb Diet

- Population: severely obese subjects (avg baseline weight = 131 kg)
- Treatments: Low Carb diet versus Low Fat diet for 6 months
- Main Endpoint: Weight loss at 6 months
- Sample Size: 132 (68 Low Fat, 64 Low Carb)

Low Fat versus Low Carb Diet

- Results: 79 of 132 subjects completed the 6-month study. For 29 of the 43 dropouts, a 6-month weight was obtained from (non-study) office visits. For the remaining 14, last available weight was used.

- Average weight loss:
  - Low Fat: 1.9 kg Low Carb: 5.8 kg
    p=0.002

A significant difference !!

Low Fat versus Low Carb Diet: Complications/Concerns

- Is a 6-month study adequate?
- High drop out rate a serious concern. Is use of 'last available weight' a source of bias?
- Endpoint: Is average weight loss the best endpoint? What about percent of subects that achieved a 'substantial' weight loss (eg, >10%)?
- Magnitude of effect: Despite statistical significance of difference, is a 1.9 kg loss vs 5.8 loss 'clinically significant' in patients whose average baseline weight was 131 kg?

Prostate Cancer prevention trial

- Population: men >55 with normal digital rectal exam and PSA<3 ng/ml
- Treatments: Finasteride (5 mg/day) vs Placebo for 7 years
- Main Endpoint: Prevalence of prostate cancer during 7 year study
- Evaluation: routine evaluation of PSA with biopsies recommended for high PSA values
- Sample Size: N=18,882 (9423 Finas. vs 9459 placebo)

Prostate Cancer prevention trial: complications

- Finasteride reduces PSA, and thus a different 'PSA trigger' for biopsy used in finasteride and placebo groups
- Finasteride impacts prostate volume
- Not all subjects recommended for prostate biopsy agreed to have one
- Persons with higher PSA's biopsied,so resulting fraction of patients found with cancer probably does not reflect 'prevalence'

Prostate Cancer prevention trial: main results

- Among men with available data at final analysis, frequency of prostate cancer was:
  - 18.4% in finasteride group (803/4368)
  - 24.4% in placebo group (1147/4692)
    - » $P < 0.001$
- However, more (280) finasteride patients with higher grade (Gleason 7-10) tumors than placebo patients (237)
- Could finasteride lower overall prostate cancer incidence but increase incidence of high-grade prostate cancers?
- Implications for patient management?

Lessons Learned

- Prostatectomy vs Watchful Waiting:
  - Good design, though arguably main endpoint should be overall mortality
  - Lack of a significant difference in overall mortality and differential side effects make implications unclear
  - Changing definition of ‘watchful waiting’ further complicates things
- Low Fat versus Low Carb Diets
  - Durations need to be sufficiently long to assess sustainability
  - Critical to get final weights for all patients
  - Magnitude of difference needs to be considered
- Finasteride to prevent prostate cancer
  - Good design but unexpected result. Finasteride effect on prostate volume could have impacted overall prevalence and high-grade prevalence.
  - Not clear where to go next !

Surrogate Endpoints

- A surrogate endpoint is one measured in place of the biologically or clinically definitive endpoint
  - osteoporosis study: bone mineral density as a surrogate for fractured bone resulting from a fall
  - HIV study: change in HIV viral load as a surrogate for clinical progression
- Main advantage is cost, time, or ease of measurement

- Critical issue is validity in assessing treatment's effect on clinical outcome based on its effect on the surrogate

- e.g. if a drug increases bone mineral density, will it decrease risk of bone fractures?

Surrogate are Disease-Specific

Disease	Clin Endpoint	Surrogate
HIV	AIDS or death……	CD4 Count
Cancer	Progression……	CR or PR
Prostate CA	Progression……	PSA Level
CVD	Stroke……..	BP
	MI……	Lipid Level
Glaucoma	Vision Loss……	Interocular Pressure
Osteoporosis	Fractures……	Bone Min. Density

Desirable Properties of Surrogates

- Measured simply without invasive procedures
- Part of, or close to, the causal pathway
- Yields the same inference about treatment benefit as the definitive endpoint (what does this mean???)

Trials Using Surrogates Can Mislead

- CAST (Cardiac Arrhythmia Suppression Trial)
  - encainide and flecanide reduced arrhythmias (surrogates), suggesting that they would be beneficial in reducing mortality
  - However, they actually were later shown to increased sudden death
- Milrinone improved hemodynamic parameters in CHF but led to increased long-term morbidity and mortality
- Flouride therapy for osteoporosis increased bone mass but led to a higher incidence of fractures

Conceptual Framework for Assessing a Potential Surrogate

- Y=clinical outcome (eg, AIDS progression)
- X=treatment group (eg, AZT or placebo))
- Z=possible surrogate (eg, change in CD4 count)

- Suppose that data show that :
  - X is associated with Y (AZT reduces AIDS prog)
  - X is associated with Z (AZT improves CD4 count)
  - Z is associated with Y (higher CD4 associated with lower risk of AIDS progression)

- Does it follow that Z (CD4) is a valid surrogate for Y (AIDS progression) ?? NO !

Conceptual Framework for Assessing a Potential Surrogate (continued)

- Required condition:

  $P(Y \mid X,Z) = P(Y \mid Z)$

In words: effect of X (treatment) on Y (clinical outcome) is entirely a result of its effect on Z. Thus, once we take account of Z, taking account of X doesn't add any information

- In practice, evaluation of surrogates is difficult
  - Need outcome data for both Y and Z for treated and untreated subjects
  - Technical evaluation can involve complex statistical methods, especially when surrogate is a marker such as CD4, PSA, BMD that can be measured repeatedly over time

Evaluating a Potential Surrogate

- Example: Mortality for patients with AIDS
  - X: AZT versus Placebo
  - Z: changes over time in CD4 cell count
  - Y: death
- AZT improves CD4 count and reduces risk of death. Yet beneficial effect of AZT on CD4 explains only a small amount of its benefit on survival.
- CD4 not a reliable surrogate for death.

Reference: Wulfsohn & Tsiatis, Biometrics, 1985.

Surrogate Markers: Lessons

- Difficult to fully assess whether a marker is a complete surrogate in the sense defined previously
- Often unrealistic to expect a single marker to be a 'complete' surrogate
- If marker lies on one of several 'causal pathways' to clinical outcome, then it's possible that:
  - A treatment with a beneficial clinical effect can have little/no effect on surrogate
  - A treatment with minimal clinical benefit can have large effect on surrogate
    - Be cautious in use and interpretation of surrogates !

Summary

- Selection of ‘endpoints’ is critical to a clinical trial or other biomedical study
- Endpoint must on one had be clinically relevant yet also be well-defined and evaulable
- Need to interpret results for the study in context of endpoint (and others)
- Surrogate endpoints/markers can be very useful, but we need to be mindful of their limitations regarding inferences about treatment effects on clinical outcomes