Applied Statistics for Translational Researchers: Design and Analysis of Clinical Trials and Animal Studies

Laurel Beckett, PhD

University of California, Davis

11 January 2017

Overview of talk

- Clinical trials: from animals to clinic to community
- A brief history
- Elements of the trial
- Study design
- Data analysis: plan ahead
- Study reporting: plan this ahead, too
- Ethical issues

Clinical trials: an introduction

A *clinical trial* is an experiment testing medical treatments in human subjects.

- An *experiment*: means that key features - what treatment, for how long, who participates, what outcome measured - are determined by the investigator.
- *Medical treatments*: the goal is to cure, relieve symptoms of, delay progression of, or prevent a disease or medical problem.
- *Human subjects*: the focus is on disease in humans, not in animals, and not in tissue culture, though these may precede studies in humans.

Early history of clinical trials

Experimental intervention studies in human subjects have a long history, some of it troubling.

- Nutrition: Biblical story of Daniel.
- Scurvy: James Lind, lemon juice for scurvy, 1742; British Navy did not adopt limes until 1804 but Captain James Cook used them on 3 voyages 1768-1779.
- Blood letting: Helmont 1662 (used lots to decide who got bled); Hamilton 1816 (alternated among surgeons; death rate 10x as high with blood letting.
- Beri-beri: Fletcher 1907, alternated types of rice.
- Diphtheria: Bingel 1918: used antitoxin, alternated with control serum of similar appearance.
- Syphilis: Tuskegee study.

Are experiments in humans ethical?

This topic is covered in required research training. I will try to add some statistical perspective. Let's start with equipoise. Required conditions:

- We don't know the best treatment.
- We think outcomes could be improved over current practice if we found out.
- This study is designed solidly to answer the question and move us out of equipoise.

Implications for statisticians

- Competence: Be sure you have the knowledge, skills, and resources, and you understand the study.
- Study design: Adequate sample size, minimal bias, early stopping if appropriate.
- Balance of harm and benefit: Who will be included/ excluded? What benefits/risks to patient, family, society?
- Formal review of plan should include statistics.
- Protection of participants: informed consent, protection of data, including from analysts as much as possible.
- Data and safety monitoring: for compliance, adverse events, accrual, possibly efficacy.
- Avoid conflicts of interest or misconduct.
- Statistical associations have developed ethical guidelines.

Some questions on ethics and statisticians

- What steps can you take to ensure privacy and confidentiality?
- What is a conflict of interest for a statistician? Wanting a study funded to support a student? Having given a paid talk at a drug company? UCD having a patent pending for a clinical device?
- Most experimental cancer and Alzheimer's treatments have turned out to be failures in that they never lead past Phase I or Phase II trials. How do you address informed consent for patients in this case?
- What are the ethical issues in using patient genotype for subgroup analysis during clinical trials? What about families? Privacy?

Modern clinical trial framework: quite recent

- The core statistical principles for experiments were formalized in the 1930's (Fisher, agricultural research).
- Beginning in 1950's, applied to clinical trials; dramatic changes, supporting evidence-based medical practice.
- A major example: the Salk vaccine trials.
- Standardized process evolved in cancer research, adapted elsewhere.
- Regulatory process: FDA oversight for approval of drugs and devices (European counterpart.)
- Modifications sometimes: e.g. AIDS trials, Ebola.

Clinical trial framework: Three phases

Oncology research developed a standardized sequence for clinical trials of new drugs and devices.

- Phase I: Find the best dosage of drug to use in humans. Get preliminary evidence on safety and efficacy.
- Phase II: Determine whether a treatment is worth taking to large-scale, definitive, comparative study.
- Phase III: Definitive study, usually randomized and usually comparative against best current options.

Additional possibilities:

- Preliminary comparative study in animals.
- Post-approval “Phase IV" follow-up in general market for side effects, adverse outcomes, differences in efficacy.
- Some AIDS trials compressed to Phase 1.5 and Phase 2.5.

Some questions for thought:

- When can you skip a Phase I trial? Give some examples.
- When can you skip a Phase II trial? Is the answer the same as for Phase I?
- How do you decide what is the experimental unit in a trial? Patient, physician, hospital, community?
- Is the experimental unit always the same as the unit of measurement?
- We usually think of a clinical trial as testing a drug. What changes if you are looking at a device? A surgical technique? A vaccine? A public health message like “eat more vegetables" or “stop smoking"?

Sources of error in clinical trials: random and bias

Random error: purely chance sources. On average, neither for nor against a specific treatment. Examples:

- Who was randomly assigned to each treatment arm.
- Random time of enrollment, affecting length of follow-up.
- Random assignment to batch for assay of outcome.

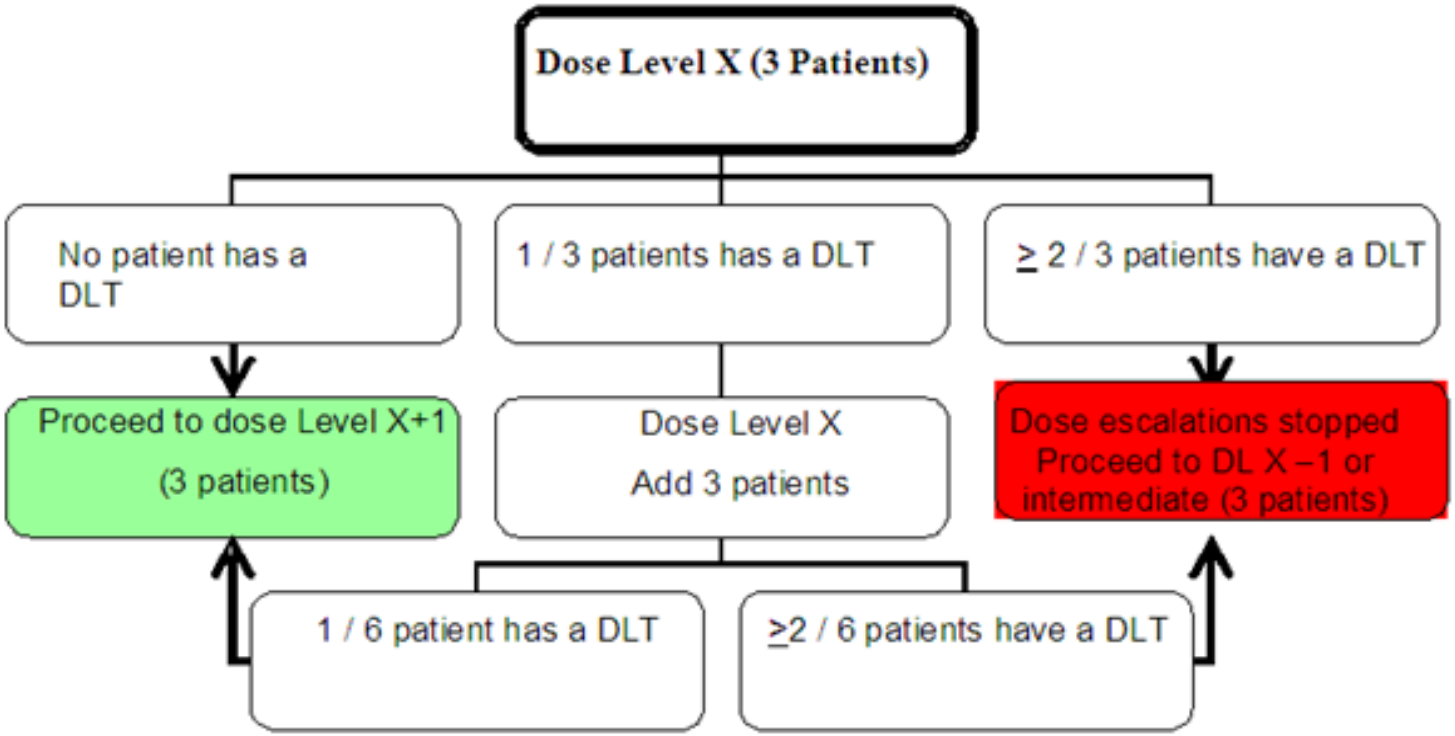
Bias: Error component that systematically over/under-estimates effect of treatment in population of interest. Examples:

- Inclusion/ exclusion criteria, dropout.
- Use of historical controls.
- Non-random determination of treatment (patient or physician preference)
- Non-blinded assessment of outcome.
- Bad analysis.

Phase I trials: dose finding and safety

- Cancer trials typically want maximum tolerated dose (MTD), without dose-limiting toxicities (DLT).
- Vaccine may want minimum dose that evokes protective response.
- Escalate to find best dose fast,
- But avoid giving a toxic dose.
- Might start with one-tenth the mouse dose.
- Escalation gets complicated for combination therapy, where you need to find dose of each component.

Phase I trials: 3+3 design



Summarize: Report the MTD, the number and type of adverse events.

Phase I trial example: INC280 + erlotinib, lung cancer

- Background: Overexpression of the MET oncogene is associated with resistance to standard treatment.
- Rationale: INC280 is a MET inhibitor; combining with erlotinib might give better outcomes.
- Phase I trial goal: Find MTD of INC280, combined with erlotinib, standard 3+3 design.
- Study population: lung cancer patients with MET overexpression in tumor.
- Outcomes: adverse events, pharmacokinetics, response.
- Results presented at ASCO 2015.

INC280 trial results: descriptive statistics

- 18 patients total were treated.
- INC280 escalated from 100 to 600 mg po/bid; DLT at 600, MTD one level down.
- AE: diarrhea and rash (47% each), fatigue (40%), etc.
- For pharmacokinetics, a linear model worked. Systemic exposure summarized by area under curve.
- 12 patients evaluable for response; 6 had stable disease, 2 had minor response.
- Statistical analysis: simple descriptive summaries, with 95% CI; mixed-effect regression analysis for PK.

Phase II trials: Does evidence support further study?

- A first look at efficacy, while continuing to look at safety.
- Decide on primary efficacy outcome:
  - Dichotomous (e.g. complete or partial response vs. none, worse)
  - Quantitative (e.g. change in cognitive function)
  - Censored (e.g. time to death or disease progression)
- Usually one sided alternative: only pursue if improves over standard, don't bother if the same or worse.
- Sometimes a randomized Phase II: which of two options to pursue, leading to 2-sided alternative.
- For power, need to know expected outcome under standard care, desired improvement.

Phase II trials: more considerations

- Should there be an interim analysis?
  - Usually only for futility, to avoid treating patients unsuccessfully.
  - Usually around halfway through; various optimization criteria.
  - Usually not enough evidence to stop for success.
  - Separate from standard monitoring from safety and accrual.
- Should you stratify? (separate arms depending on patient characteristics.)
- Should you randomize? (Two versions of new drug, or new vs. old)

Phase II trials: prostate cancer example

- Study population: castration-resistant prostate cancer, pretreated with docetaxel
- Treatment: Low-dose ketaconazole plus hydrocortisone
- Outcome: PSA response ($\geq$ 50% decrease in PSA)
- $H_0 \leq$ 5% respond, H_A: $\geq$25% respond.
- To achieve 80% power at $\alpha = 0.05$, one-sided, set $n = 25$.
- No interim analysis.
- No randomized control group.
- Therefore not blinded, but objective outcome.

Prostate cancer example results

- 30 patients accrued, 29 evaluable for response and toxicity.
- PSA response (drop of 50% or more) in 48%; 59% had drop of 30% or greater.
- Median progression-free survival 138 days.
- 12 patients experienced grade 3 or 4 AE's, but only 3 of the 17 events were treatment related.
- Conclusion: appears to be a safe, inexpensive, and clinically active treatment option.
- Further study warranted.
- Ref: Lara et al, *Prostate Cancer and Prostatic Diseases* 2015.

Some statistical notes for Phase I and Phase II studies

- Analyses generally straightforward, except pharmacokinetics. Basic SAS code:
  - Proportions: `proc freq; tables orr/ binomial(exact) alpha=.05;`
  - Means: `proc univariate cibasic(alpha=0.05); var psachange;`
  - Survival time: See documentation for PROC LIFETEST.
- Follow best practices for data checking, code checking, documentation.
- Journals increasingly require “reproducible research" practices.
- FDA has even higher standards.

Randomized Phase III studies: typical features

- Large-scale, comparative trials, intended to give definitive answers.
- Concurrent comparison with best available treatment (sometimes placebo).
- Randomized assignment to treatment.
- Objective evaluation of response:
  - By objective endpoints if possible,
  - Otherwise by masking participants and evaluators to treatment.
- Pre-planned analysis.
- Usually two-sided hypotheses, not one-sided. (Why?)

Phase III studies: complicating features

- More than two treatment arms, e.g. 2x2 design.
- Unbalanced assignment to treatment.
- Need to block by site or by patient characteristics.
- Non-compliance, drop-out.
- Interim analyses.
- Testing for noninferiority instead of improvement over standard.
- Multiple outcome measures, oddly distributed outcomes.
- Cluster-randomization.

Avoiding bias: randomization, masking are important

- Review found failure to randomize led to biased effect sizes (150% greater - 90% smaller) (Kunz and Oxman, 1998).
- Randomization should be generated by computer, and program saved including the “seed".
- Fancier features: blocking (e.g. making sure each site has balanced numbers).
- How are treatment assignments communicated? Possibilities: online program, list to each site, coordinated with pharmacy that creates masked packets?
- How is masking preserved? Is masking checked?
- Who has access to the treatment assignments?
- How is masking broken if needed for clinical reasons?

Historical example: 1954 Salk polio vaccine trial(s).

RCT, double blind:

- About 200K children got Salk vaccine, 200K got placebo.

Initial study plan:

- 220K 2nd graders vaccinated, 124K parents refused; 725K unvaccinated 1st and 3rd graders as controls.

Results of the study:

- Polio rate in RCT vaccinated: 28/100K; in placebo controls 71/100K. Among those whose parents declined: 54/100K.
- In non-randomized study: vaccinated 2nd graders: 25/100K, observed controls 54/100K, refusers 44/100K.
- About 10% of reported polio cases could not be confirmed by lab tests.

Results support importance of randomization, masking.

A more complicated example: Women's Health Initiative (WHI)

A huge study of health impacts of multiple interventions in post-menopausal women 50-79.

- Initial stage: one year of a 2×2 design with factors:
  - Dietary modification (DM) - less fat, more vegetables, fruit, grains
  - Hormone replacement (HT) individualized for woman.
- Second stage: for subset of women, RCT of calcium plus vitamin D (CVD) vs. placebo
- We will focus on the CVD component in this example.

Outcome possibilities for CVD vs. placebo in WHI

Choice of outcome affects analysis plan, sample size. Several possible outcomes for bone health in post-menopausal women:

- Incidence of fracture, type of fracture.
- Time to fracture.
- (For subset of women) - bone mineral density (BMD) at baseline, 3, 6, 9 years.
- Adverse events: death, kidney stones, other symptoms.

Should you plan statistical correction for multiple outcomes? For example, take α= 0.05/3 for 3 outcomes.

Complicating factors for planning and analysis

- Drop-outs from study, censoring, loss to follow-up
- Non-compliance with planned treatment.
- Stratification: which arm at baseline, study site, ethnicity.
- Baseline risk may vary (with BMI, age, ethnicity, baseline BMD)
- Some women were on HRT - could there be effect modification?
- Some women used personal calcium supplements or had high dietary intake.
- Some women started taking osteoporosis medications.

Raises questions about Phase II and III studies:

- Who gets into the analysis?
  - Intent-to-treat group: Everyone enrolled and randomized (even if no treatment? Dropped out partway?)
  - Safety group: Those who received at least one dose.
  - Per-protocol: Only those who received treatment as intended (how to handle noncompliance?)
- What if groups came out imbalanced?
  - Include covariates?
  - Primary or secondary analysis?

Results for CVD and bone health, *NEJM* 2006

Study participants:

- Randomized 18,176 to CVD, 18,106 to placebo.
- About 1,200 in each group got BMD measurement.
- At close-out, about 16,900 in each group. Rest withdrew or lost to follow-up.
- About 60% compliant (took at least 80% of pills), another 20% took more than half.
- Baseline mean hip BMD -0.7 (T score vs. young healthy), but lots of variation.
- Baseline mean calcium intake 1200 mg/day but lots of variation.
- Poses challenges: how to handle drop-out, noncompliance, baseline differences.

Results of WHI study: fracture incidence by site

- Fractures not common: 170 per 10,000 person years.
- Primary analysis: intent-to-treat groups.
- Time to fracture used Cox proportional hazards models.
- Hazard ratios for hip, vertebra, lower arm were NS.
- For hip, hazard in intent-to-treat was 0.88 (95% CI 0.72–1.08).
- When restricted to adherent women, hazard was 0.71 (0.52–0.97), significant reduction.

Fracture incidence analysis details

- Study designed for 85% power for total fractures, apparently $\alpha = 0.05$.
- Primary analysis total fractures, also looked at 95% CI for hazard for each site.
- Handled drop-out via censoring, stratified by age group, prior fracture, and what group they were in for the parent 2×2 study.
- Secondary analysis looked at risk factors as effect modifiers.
- Non-adherence examined in sensitivity analyses.
- Cox proportional hazards models like this can be done in SAS PROC PHREG.

Results of WHI study: BMD outcome

- Analysis complicated by baseline differences.
- Various strategies proposed for this.
- They reported percent change in BMD from baseline.
- Many statisticians would recommend analysis of covariance; may not make much difference in this RCT.
- Mean percent difference in hip BMD favored CVD group:
  - 0.59%, 0.86%, 1.06% greater BMD at 3, 6, 9 years respectively.
- Differences for spine and whole body BMD in same direction but NS.

Additional results: subgroups, effect modification

- Among women 60+, seems more protective against fracture than among younger women.
- Maybe more protective among those without history of recent falls.
- No evidence of effect modification by HRT.
- No difference in serum vitamin D for hip fractures vs. matched controls in nested case-control substudy.
- Slight increase in risk of kidney stones.
- No other significant risks or benefits for other disease outcomes, including cardiovascular and cancer.

More complex analysis: how to deal with non-adherence

WHI found only 60-80% compliance on CVD. Some took part of pills, some quit taking.

- How do you define analysis groups?
  - ITT: based on randomization, even if never got treatment, quit taking, or were not supposed to but did take.
  - Per protocol: how do you define who to exclude?
  - Adjusting for treatment received: how do you do this?
- Most experts recommend ITT as primary for efficacy. Other approaches tend to be biased, often in favor of alternative.
- For toxicities, consider treatment actually received.
- Biological questions like surrogate marker analyses also need to look at treatment received.
- Causal modeling tries to dig deeper into this. See papers by Pear, Robins, Rubin, and others.

Some other topics: very small trials

Sometimes pilot studies have very small samples. They are still clinical trials and still important!

One nice example in translational oncology: Monzajeb, Kent, et al., *Clinical Cancer Res.* 2016. Pilot of experimental therapy to block immune rebound and boost effects of radio-immunotherapy in dogs.

- Only 5 dogs, spontaneous tumors.
- Dog tumors more like human than are mice.
- All 5 dogs showed desired response.

Other complications: cluster-randomized designs

So far all our examples have had randomization and outcome measures both at the level of individual patient.
What happens if you randomize a whole group, but measure outcome for individuals within the group?

- Randomize hospital unit to certain infection control protocol, measure infections at patient level.
- Randomize junior high to program against uptake of smoking, measure outcome at child level.
- Often used for changes at provider level (physician, nurse, practice, hospital)
- Also used for community-level interventions.
- Interesting current example: recovery of guns from armed and prohibited persons.

Specific challenges of cluster-randomized designs

Implications of these designs:

- Effective sample size is between the total number of groups and the total number of individuals.
- Depends on the within-group correlation.
- Even modest correlations can give big reduction in effective sample size.
- Sadly, they are often planned wrong and analyzed wrong.
- Will defer details to another talk.

Reporting clinical trials: CONSORT guidelines

- This statement gives very clear guidelines to what should be in a published paper from a clinical trial.
- I advise pre-planning and blocking out the tables and diagrams as shown, at beginning, and updating during the trial.
- It's also very helpful as a structure for presenting to DSMC meetings.
- Combined with Reproducible Research expectations, tells you what to think about and what to save for presentation.
- FDA requirements are more complex; see their website.

Partial topics list from CONSORT paper: Methods

- Participants: inclusion, exclusion, settings, location.
- Interventions: precise details
- Objectives: state hypotheses
- Outcomes: Define primary, secondary measures, how assessed, how quality assured.
- Sample size: how determined, whether any interim analyses were planned.

Partial topics list from CONSORT paper: More methods

- Randomization: how was sequence generated, including any restrictions?
- How was allocation concealed until intervention was assigned?
- How was random assignment implemented?
- Masking: How was treatment concealed from patients, treating physicians, and those assessing outcomes?
- Masking: If masked, was success evaluated?
- Statistical methods: How were groups compared, and methods for secondary analyses.

Partial topics list from CONSORT paper: Results

- Patient flow: they suggest a diagram, and I find this helpful for interim reports also.
- Recruitment: Dates defining periods.
- Baseline data: The classic Table 1, who's in the study.
- Numbers analyzed: Who is in each analysis group, how defined.
- Outcomes: results for each primary and secondary outcome, including effect size and its precision as a 95% confidence interval.
- Ancillary analyses
- Adverse events

Some parting thoughts on clinical trial analysis

Current issues of top-tier journals give good examples.

- *NEJM* has a Phase I study of a drug that might lower LDL cholesterol in healthy volunteers. Several arms with different regimens. No serious adverse events up to maximum dose, and some preliminary suggestion of reduction in a key biomarker and LDL.
- *JAMA* has a randomized Phase III open label, non-inferiority trial of zolendronic acid for bone metastases in metastatic cancer, every 12 weeks vs. every 4 weeks. Found every 12 weeks was not inferior.

Some references

- Piantidosi, *Clinical Trials: A Methodologic Perspective*.
- Pocock, Clinical Trials: A Practical Approach.
- Green, Benedetti, and Crowley, *Clinical Trials in Oncology*.
- Ellenberg, Fleming, and DeMets, Data Monitoring Committees in Clinical Trials: A Practical Perspective.
- Moher *et al*, The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials. (*JAMA 2001* and other places)