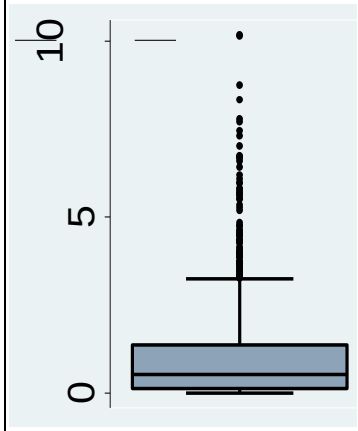
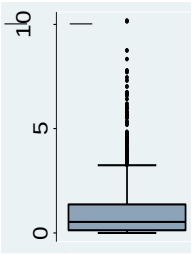
Analytical and Design Issues in Translational Research

Charles E. McCulloch
Head,
Division of Biostatistics



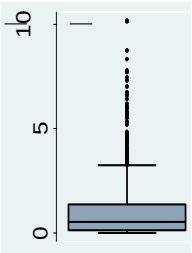
May 22, 2014

Outline



- Introduction
- Examples
- Design and analysis issues
- Cluster randomized trials (CRTs)
  - Statistical analysis of CRTs
  - Design considerations for CRTs
- Interrupted time series designs (IRTs)
  - Statistical analysis of IRTs
  - Design considerations for IRTs
- Design variations
- Summary

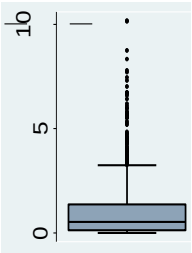
Example 1: Decision Aids in Breast Cancer Treatment



Does use of a decision aid increase patient knowledge of surgical treatment of breast cancer? Surgeons were randomly assigned to use or not use the decision aid in surgical consultations. A few days after the consultation, patients' knowledge and decisional conflict were measured with questionnaires.

(JAMA, 2004)

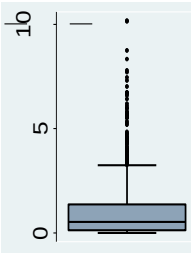
Example 2: Insurance usage in rural India



Can after sales service (S) or prospective reimbursement (P) increase usage of health insurance? 4 subdistricts were randomly assigned to each of 3 interventions and a control (C=control, S, P, S+P). Claim rates were measured before and after the interventions.

Vimo SEWA study

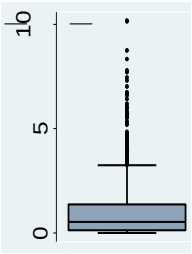
Example 3: Radiologist referral patterns



What is the impact on referral patterns of the dissemination of a booklet on *Making the Best Use of a Department of Radiology* to general practitioners?

The booklet was introduced on January 1, 1990. In the study, the number and type of referrals were measured for the year before and the year after the introduction of the guidelines. There was a reduction in radiological requests after the guidelines were introduced, and it was concluded that the guidelines had changed practice.

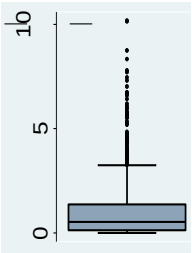
Example 4: QIDS



Quality Improvement Demonstration Study

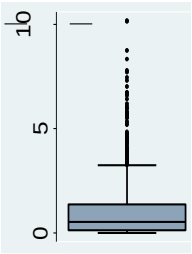
Can child health in the Philippines be improved by expanded insurance coverage (A for Access) or physician pay-for-performance (B for Bonus) incentives? 30 public hospitals were randomized to one of three arms: A, B or a control arm (C). Physician performance was measured before the intervention and every six months thereafter for 3 years.

Design and Analysis Issues



- Depending on the scale of the intervention, extra considerations in the design and analysis of the study are often required.
- Typical design features:
  - Cluster randomization
  - Before/after designs
  - Interrupted time series (repeated before/after)

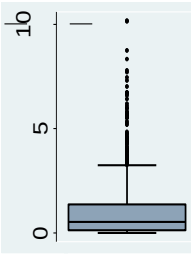
Cluster Randomized Trials



CRT = “Courting Real Troubles” – so why use?

- Intervention may dictate administration at the cluster level.
- May be logistically more feasible.
- Avoid contamination of the control subjects that might occur with a within-cluster (subject-specific) randomization.

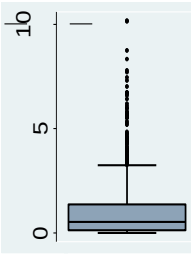
Cluster Randomized Trials



Disadvantages

- Analysis scheme may be more complicated to accommodate clustering.
- Required sample size is invariably larger.
- More complicated planning/design, e.g., tradeoff between number of clusters and number of subjects per cluster.

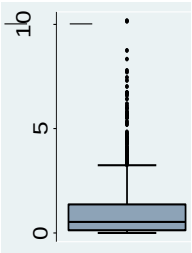
Interrupted Time Series



Advantages

- Intervention may dictate administration at a single or very few sites.
- Multiple measurements before and after allows assessment of trend and autocorrelation.
- Use site as its own “control” by looking before and after.

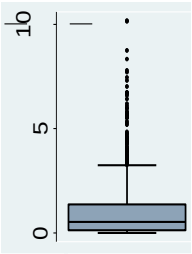
Interrupted Time Series



Disadvantages

- Analysis scheme is more complicated.
- Required sample size may be larger.
- Often not randomized, so inference of causation is weaker. Can, at best, assert that there is a statistically significant difference that occurred at the time of intervention.

CRT - analyses

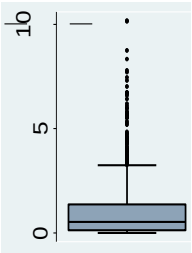


Consider Example 2: Insurance claims rates in rural India. Made up data for support and supervision (intervention) and control conditions, post intervention. M=4 subdistricts, n=1000 women assessed per subdistrict:

	Claim rates per 1,000 women
Intervention:	0.15, 0.12, 0.09, 0.11 (ave of 0.1175)
Control:	0.09, 0.10, 0.08, 0.07 (ave of 0.085)

So, in the intervention condition there were a total of 150+120+90+110 = 470 out of 4,000 who filed claims. In the control condition there were 340 out of 4,000.

CRT - analyses

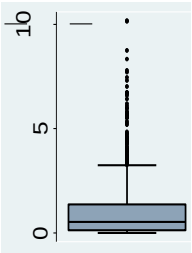



```
. csi 470 340 3530 3660

             |   Exposed   Unexposed  |       Total
-------------+------------------------+------------
       Cases |         470         340  |         810
    Noncases |        3530        3660  |        7190
-------------+------------------------+------------
       Total |        4000        4000  |        8000
             |                          |
        Risk |       .1175        .085  |      .10125

                 chi2(1) =    23.21  Pr>chi2 = 0.0000
```

CRT - analyses

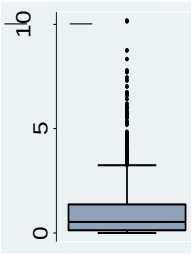



```
. ttest Intervention=Control, unpaired unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
Variable |  Obs    Mean Std.Err. Std.Dev. [95% CInt]
---------+--------------------------------------------------------------------
Interv~n |    4    .118    .013     .025  .077 .157
 Control |    4    .085    .006     .013  .064 .106
---------+--------------------------------------------------------------------
    diff |         .035    .014          -.005 .070
------------------------------------------------------------------------------
diff = mean(Interv) - mean(Control)    t =   2.3102
Ho: diff = 0         Satterthwaite's d.f. =  4.49378

Ha: diff < 0      Ha: diff != 0        Ha: diff > 0
Pr(T<t)=0.96      Pr(|T|>|t|)=0.075    Pr(T>t)=0.037
```

CRT analyses



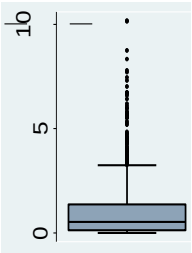
- Or could do a clustered data analysis.

```
xtgee tot_claims Intervention,
  i(siteid) family(binomial
  Nsampled) link(identity)
  corr(indep) robust nmp
```

- Which gives results
  - 0.035 estimated diff, SE=0.013, z=2.50, p=0.013

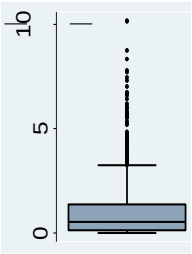
  Compared to the t-test which gave
  - 0.035 estimated diff, SE=0.014, t=2.31, p=0.075

CRT analyses



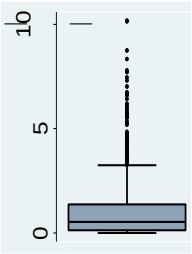
- The key to proper analysis of a CRT is to accommodate the clustered nature of the design and not analyze each individual as if they were independent.
- So chi-square test is wrong.
- With equal number of observations per cluster and no subject level covariates a very effective analysis is simply to average or total values to the cluster level and perform a simple, non-clustered analysis.
- In our example – just do a t-test.

CRT analyses



- In this simple situation `xtgee` isn't as good as a t-test.
- First, because it uses a z-test when there are, in essence, only 8 observations.
- And, further, the t-test can allow for different variances in the two groups.
- It does it by adjusting the d.f. down to about 4.5.

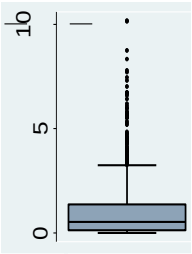
CRT analyses



- In more complicated situations, e.g.,
  - Subject specific covariates
  - Unequal sample sizes

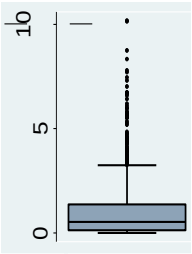
  we need to use clustered data techniques.
- Typical statistical methods
  - `xtgee, melogit, mepoisson, meglm, mixed` in Stata
  - `Proc MIXED, NLMIXED, GENMOD (with REPEATED)` in SAS

CRT Sample size planning



- Very important observation: With equal number of observations per cluster and no subject level covariates a very effective analysis is simply to average or total values to the cluster level and perform a simple, non-clustered analysis.
- For example, the t-test for the insurance claim data.
- The CRT design is much less precise than a design with a random sample of 4,000 woman for each group (which chi-square assumes).
- Tells us how to design studies and do sample size planning.

Sample size considerations



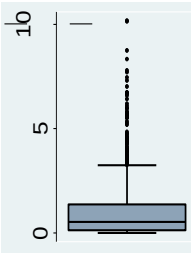
Sample size basic:

For a given effect size, sample size depends on the standard error (SE) of the estimate, or, more precisely, it is proportional to the square of the SE, i.e., the variance:

$$\text{sample size} \propto \text{variance of estimate}$$

So, if a design doubles, say, the variance then the required sample size doubles.

Sample size considerations

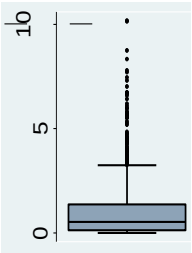


So the key is to figure out how much a CRT increases the variance of the estimate. Then we know how much larger a sample size will be required. This proportional increase in variance is known as the *design effect (Deff)* and is given by the formula:

$$Deff = 1+ (n-1)\rho,$$

Where n is the sample size per cluster and ρ is the intraclass correlation within a cluster.

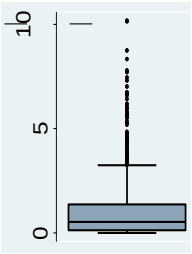
Sample size considerations



Suppose a subject-level randomized trial requires 100 subjects per arm. How many more are needed for a cluster randomized trial?

Cluster size	ρ	Deff	Required total number of subjects	Number of clusters
25	0.05	2.2	220	9
250	0.05	13.45	1,345	6
25	0.5	13	1,300	52
250	0.5	125.5	12,550	51

Design considerations



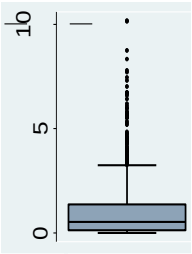
Another way to look at the same idea is that the variability of the estimate has two components: the within cluster variation, σ_W^2 , and the between cluster variation, σ_B^2 :

Var(estimate) $= \frac{\sigma_B^2}{m} + \frac{\sigma_W^2}{mn}$

where *m* = number of clusters and

n = number of observations per cluster

Design considerations

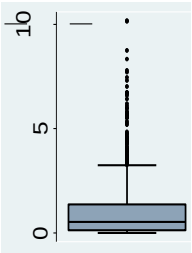


The intraclass correlation is given by

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \quad .$$

So when the intraclass correlation is 0.5, the two components of the variance are equal. Consider the case when they are both equal to, say, 5 and let's see what happens when we try different values of m and n.

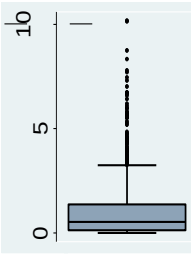
Design considerations



Values of $\frac{\sigma_B^2}{m} + \frac{\sigma_W^2}{mn}$ when both variances are 5 for various combinations of *m* and *n*.

m=no. of clusters	*n*=no. of obs per cluster	Variance	Total sample size
100	1	0.1	100
50	25	0.104	1,250
10	10	0.550	100
5	20	1.050	100

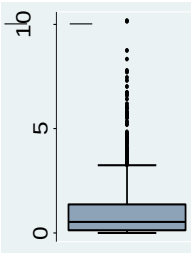
Design considerations



If you randomize observations within a cluster, then the between cluster variation is eliminated from the calculation. A consequence is that you may be able to withstand considerable contamination and still gain power over a cluster randomized design. However, you will underestimate the true effect.

I advocate doing the calculations for a range of degrees of contamination to see which design is more powerful.

Example 1: Decision Aids



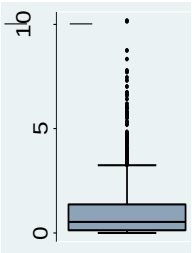
For the Decision Aid study there were 27 physicians and 201 patients for an average number of patients per physician (cluster) of 7.4. They assume a within physician correlation of 0.3.

So the *Design Effect* is 1+(7.4-1)*(0.3) = 2.92.

And they could have withstood a contamination that reduced the effect size by 40% and had more power with a within physician randomization.

Usual sample size calculations to detect an effect size of 0.3 with a SD of 0.5 says to use 45 per group, so the CRT would need 2.92*45 or 131 per group with 131/7.4 or about 18 providers per arm.

Interrupted time series analysis



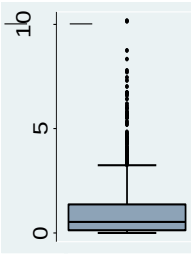
Three years of observation before an intervention and three years after. Perform a t-test to compare the three before with the three after, p = 0.0047.

What is wrong with this analysis?

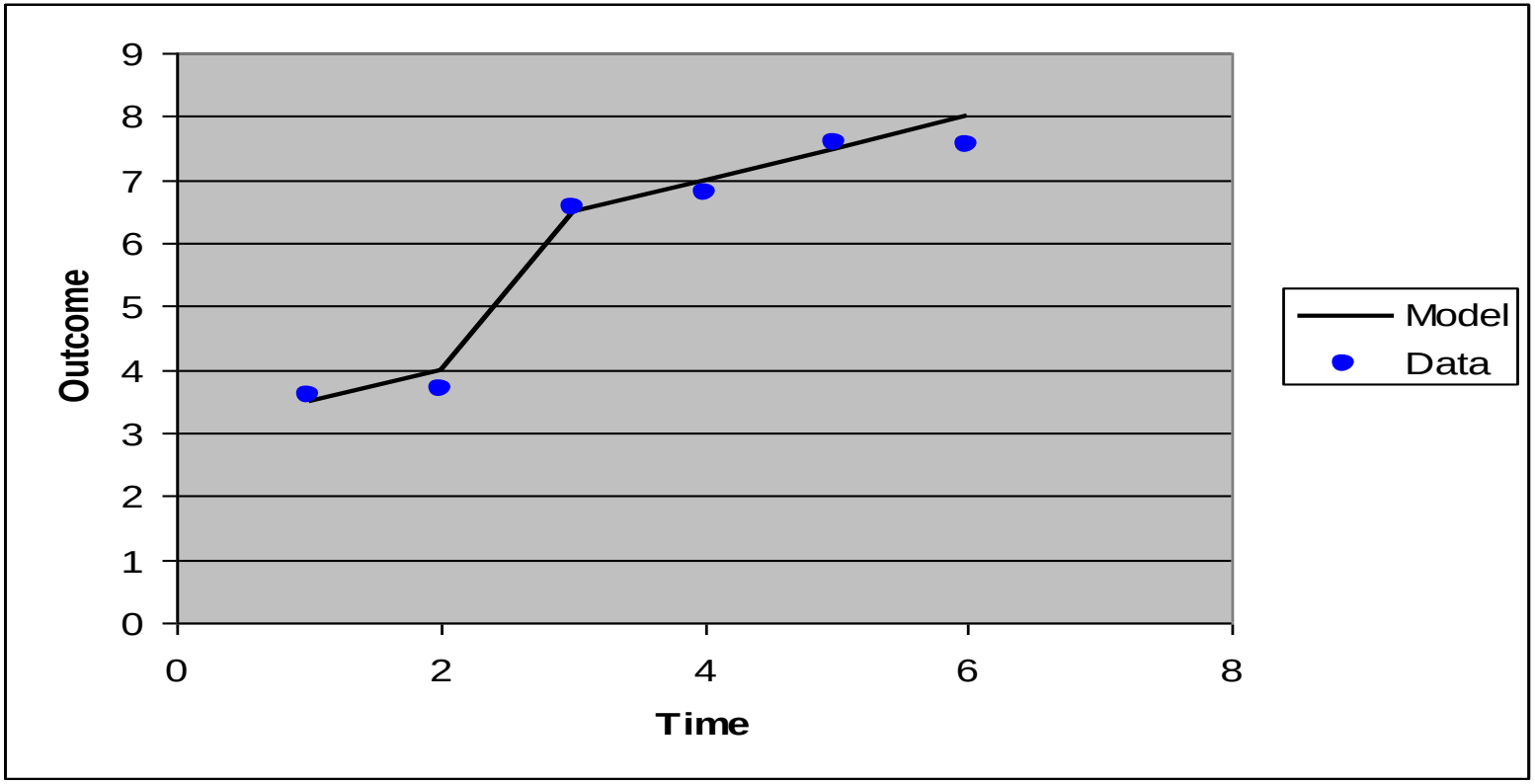
What if the values were: 0.07, 0.08, 0.09, [intervention], 0.10, 0.11, 0.12?

Autocorrelation?

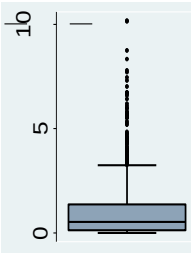
Interrupted time series analysis



Usual strategy: fit a secular trend (simplest is linear trend over time), allow a “bump” at the intervention.



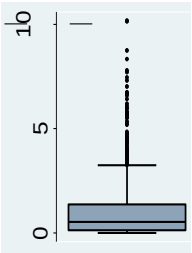
Interrupted time series analysis



Allow for autocorrelation over time. Autocorrelation just means that current values are correlated with previous values.

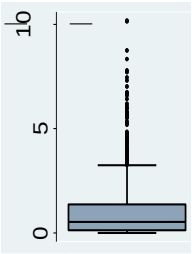
If there are multiple sites, use a longitudinal analysis with autocorrelated errors. Uncorrelated between sites, correlated within.

Interrupted time series analysis



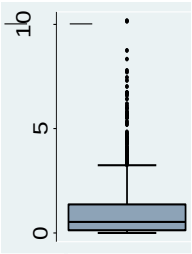
See variance calculator

ITS analyses



- Typically use time-series analysis methods
- Stata: `arima`
  - `AutoRegressive and Integrated Moving Average`
- SAS: `ARIMA, AUTOREG`
- Or longitudinal methods that allow for autocorrelation structures (e.g., `mixed`)

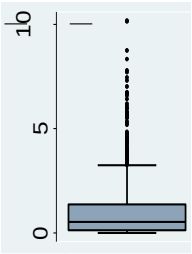
Design variations



Best designs use

- Multiple sites
- Multiple observations before and after the intervention
- Stagger the interventions across the sites so the intervention is not completely confounded with whatever else may have occurred at the same time (stepped wedge).

Summary



- Cluster randomized or interrupted time series designs may be the most appropriate designs for large scale interventions
- But they come with significant drawbacks
- Each require special analysis methods (clustered data techniques or time-series methods)
- Beware the design effect
- Use between and within variance calculations or design effects for sample size and resource allocation decisions in CRTs