Murray on GRTs

Chapter 1

GRTs are comparative studies where the units of assignment are identifiable groups, and the units of observation are individual members from those groups.

- Groups are not constituted at random, but they have some common physical / geographic / social trait in common.
- Different units of assignment (groups) are allocated to each study condition (trt/ctl)
- Members are nested within the groups which are nested within the condition (hierarchical model)
- The limited number of groups in most studies (10 to 15 per arm) makes us use the t-distribution rather than the z-distribution for statistical inference

Analytical issues in GRTS

Let Y i:k:l the outcome observation for the ith participant nested within the kth group nested in the lth condition. Since the groups are non-random, we have a portion of the variation in the outcome due to the group. We call this the ICC (intra class correlation: between group variance / [between group variance + within group variance])

And the group variance is $\sigma_{yg} = (\sigma_e/m) + \sigma_{g:c} = \sigma_y/m\ (1 + (m-1)*ICC)$

The condition variance is $\sigma_{yc} = \sigma_{yg} / g = \sigma_y/mg\ (1 + (m-1)*ICC)$

σ_y/mg would be the variance in the outcome if all observation within the groups were independent

(1 + (m-1)*ICC) if the variance inflation factor or design effect, which is the adjustment due to the positive correlation in the outcomes.

If the ICC > 0 then $\sigma_{yc} > \sigma_y$ so by assuming independence, our standard errors will be too small, leading to larger test statistics and false positives (statistically significant results when we shouldn't have one)

Design issues in GRTs

Since we only have a limited number of groups in GRT, it is harder to protect against biases (differences between our trt and control arm that makes our control a bad counterfactual substitute (ideally, we would like our treatment and controls to be equal in all aspects save for the treatment) so we have to be particularly careful about threats to internal validity (ability to make causal inferences with good design) in our study.

Chapter 2: Planning the trial

The research question should be crafted around these two principles: Is the study important to do? (extent of the problem and/or potential benefits). Is it the right time for the study? (Question not previously answered/good data and measurement techniques available)

Good research design includes these three aspects: control observations (appropriate counterfactual substitutes), minimizing bias in our estimate (internal validity) and maximizing the precision of the outcome (ensuring the effect isn't attributable to random variation in outcome)

Potential sources of bias: Biases create differences in the treatment and control groups as confounders (competing explanation for the trt effect) because our control are no longer counterfactual substitutes to treated.

- Selection bias, differential history, differential maturation and contamination are all biases that can be protected against with randomization (with large numbers, protects against measured and unmeasured confounders). Since groups are often limited in GRTs, a priori matching and stratification are also used in conjunction with stratification to make the groups more similar.
- Differential testing, differential instrumentation, regression to the mean, loss to follow up, rivalry or demoralization are all biases occurring after randomization, but they can be protected against using objective measures, data collectors blind to trt assignment, and measuring potential confounders.
- In the analysis, you should also look out for high variation in the outcome, high ICC, low replication as signs of inappropriate implementation of procedures that lowered the precision in our outcome. You can also use repeated observations model time for longitudinal data and use techniques such as regression adjustment and pot hoc stratification to improve precision.

Chapter 3: Research Design

Whenever you plan on doing a priori matching or stratification, you need to also collect data on your matching variables. Main effect (only one intervention of interest) vs factorial effect (multiple interventions, including their joint effects). You need a lot groups to have a good post test only analysis: otherwise you miss all baseline information (were the groups comparable at baseline?) and you cannot see if your effect changed over time, or if there was potential contamination (no pre post comparisons).

Chapter 4: Planning the analysis

Fundamentals:

- Use the research question to determine the primary and secondary endpoints
- Use the design to determine the # of conditions, # of levels per conditions, a priori group matching variables, and main effect vs factorial analyses.
- The endpoint can be continuous, categorical, count, rate or survival data
- You can use regression adjustment of covariates and model time variables to reduce bias and improve precision

Threats to the validity of the analysis

Misspecification of the model: can be due to missing measurable sources of variation (ignoring the clustering effect) or to using the wrong correlation structure for group correlation (not modeling the cluster as a RV)

GLM assumptions

- Normally distributed errors (ok if we have similar sample size of groups per condition)
- Homogeneity of errors (ok if we have similar sample size of groups per condition)
- Independence of errors (Very serious violation, because our SE is too small when we ignore the positive ICCs in our data due to clustering)

To protect the internal validity in our study, we identify all possible sources of random variation in our data, check the variance assumptions using model fits, and use robust methods in the analysis.

Statistical models

Fixed effect: we want to draw inference on specific levels of the variables on the outcome

Random effect: want to generalize the inference to a larger population (group level variable)

- General Linear Model: intercept + slope + residual
- Generalized linear model: GLM for non-normal outcomes (use the Gaussian, binomial and poission distribution in conjunction with identity, logit and log link functions respectively)
- General Linear mixed model: GLM for tow or more random variables (clusters or repeated measures
- Generalized linear mixed model: GLMM for non-normal outcomes (use the Gaussian, binomial and poission distribution in conjunction with identity, logit and log link functions respectively)

Estimation of outcomes

- Ordinary least squares: minimize the sum of squares deviation between data and predicted values in the data: best for independent residual errors
- Maximum likelihood: maximize the probability that the predicted values match the data. Can be used for both fixed and random effects. The restricted ML does not estimate the fixed and random effects at the same time, rather it estimates fixed effects first then uses the to inform random effects estimation.
- Parametric methods make assumptions on the distribution of the standard errors in our model, but non parametric methods do not.
- Sampling distribution and DF: find the post hoc condition, find the ICC, calculate the VIFF then make the sampling distribution and use the t-test with c(g-1) df to estimate your statistical significance.
- Units of analysis in the model: there are units of assignment, intervention and observations in GRTS. The units of analysis for an effect are those and only those for which this effect is assessed against the variation among these units.

PUBH 6363 Lecture Notes

Lecture 1

How to write an article review

First Paragraph: Summarize the paper, according to the authors (Think article abstract, but less formal)

Second Paragraph: Talk about the contributions of the paper and the things that the paper did well

Third Paragraph: Talk about the flaws that in your opinion this paper brings; In the first half, go into the deep critique of the paper, then in the second half of the paper, go into the miscellaneous critique of the paper.

Remember to make sure that your comments are constructive criticism; The mindset you should approach any paper you review with is: Is there sufficient data to address the question? Depending on the answer to this question, you have to decide if you accept the paper with minor revision/accept after major revisions/reject the paper.

Third paragraph: Include the comments to the editors

General research approach

Good research question -> Appropriate data to answer the research question

Study aims (what you aim to do) -> Study Design (How you collected the data to answer your question/what you did) -> Study Conclusions (Does the data work to answer the research question)

Significance of the research question -> data + Methods -> Study Findings -> Study Conclusions

When questioning various statistical models, focus on whether the different statistical models you would like to see would change the conclusions

Each table should be able to stand on its own, along with proper footnotes

Make sure that in your conclusions, you are clear about your assumptions in the limitations section

Make a skeleton table before starting a new analysis for a project/paper: This way it can focus your research and analysis in order to fill in your tables with all the data needed to answer your research questions.

When writing the paper, talk about the contribution that your findings are bringing to the body of literature in the introduction: identify a gap in knowledge, then explain how your paper is addressing said gap.

Group Randomized Trials

Feasibility/ Pilot Studies are very important in research

Make sure that baseline measurements are done before randomization and Tx group assignment: This is done to avoid bias during Tx assignment/Measurement of outcome based on knowledge of Tx assignment.

Blinding (Single or Double) is also done to ensure/remediate the bias created by knowledge of treatment assignment in some studies. To ensure compliance in blinded studies, it is important to inform participants of the protocol and role of blinding, and guarantee that the ctl group will receive delayed intervention, if the findings are beneficial

ITT Vs ATT analyses: Analysis of each Tx arm as it was randomized, regardless of actual exposure (Intent To Treat): This approach focuses on the effect of the treatment. Analysis of each Tx arm as it was exposed (As Treated Analysis): This approach focuses on the practicality of the treatment.

Causes of within group similarities: Clustering and ICC are statistically meaningful concepts, that are created by self-selection of individuals into a group, and shared experience of group members .

Lecture 2

Red Flags in Research: Lack of exchangeability between treatment arms, based on randomization process or comparison in baseline measurements. Randomization only balances out measured and unmeasured confounders when we randomized a large number of individuals.

Effect Estimation: 0.3 of SD is a small effect, 0.5 of SD is a moderate effect, and 0.8 of an SD is a large effect estimate.

Biases in effect measurement: There can be exogenous treatment effect, or contamination effect. There can also be social desirability bias which affects the response of participants. Finally, panel conditioning (the simple act of asking a question can change the response)

Dealing with withering effect in longitudinal / studies with repeated measures: You can run pre-post analysis, repeated measures, or differences in differences. However, since the effect of interventions is usually only significant shortly after the intervention and decreases thereafter, doing more baseline measurements (in social studies) or thorough wash-in periods (in biological studies) can help separate natural variations in effects (without the intervention) from the effect attributable to the intervention.

Lecture 3

The same rules that apply to randomized trials apply to group randomized trials

The main difference between RCTs and GRTs is due to the effect of Clustering and Nesting:

Clustering (Shared Experience + Self Selection) -> Leads to ICC: this is the within group correlation, also called the intra-class correlation (between group variance / between group variance + within group variance). Since the total variance is a sum of the within group variance and the between group variance) the ICC is a ratio between 0 and 1. In GRT's the majority of the variability is due to the group itself (large within group variance), so the ICC will typically be a very small number, but have a significant impact on the effect estimation.

Nesting (Based on the unit of randomization): Each group is a cohesive unit, but it is also made up from smaller units each of which have an individual error term. However, we also have to account for the

group level error term. In GRT's the group level error term helps determine the df (statistical power) of our analysis.

Nesting (also known as groups within conditions/Tx arms) implies that on top of individual errors, there are group level errors that drive the underlying association. The more measurements we take, the more precise our estimate becomes, but what matters is that you have many groups in each treatment arm, the number of people per group itself is less important.

Ecological fallacy: you cannot use a correlation or association observed at the group level to make an inference at the individual level or vice versa. Even though this may be mathematically correct (the average is a good estimate of individual values, it is not practical for making inferences about the individuals, because you are only technically right 1% of the time (when the true value is the average) but wrong 99% of the time (the rest of the distribution)

The crucial issues in GRTs are related to recognizing when a GRT is needed at the study design stage (aka knowing when you need to design and run a GRT. It is most commonly used in social interventions, where intact social groups (or their "groupiness") is of interest to the investigators. Otherwise, if the observations are treated as IID, this will lead to a falsely small SE, increase the test statistic and the type one error rate (false positives). The corollary to this is that the number of groups available is often limited, so GRTs are often underpowered study.

Lecture 4: Overview of GRTs

Overview of GRTs

The unit of randomization is a group, not individuals within a group

The unit of treatment application can either be the whole groups (limits our ability for sub-analyses) or all individuals within the group (difference from RCT's)

If you randomize within a cluster, you have a multi-site RCT, whereas randomization of the whole cluster to either treatment arms: this means that for multi-site RCTs we have the option of using either the individual effect or group level effect in the analysis, whereas for GRTs we can only estimate the group level effect

Group specific effects are important to estimate because we need to account for nesting (group level error + individual level error)

When we talk about nesting, the strata used within each treatment arm should be comparable themselves (I.e, you cannot compare three school within one city to three schools in three cities, because the true cluster in this case is the city, which is not comparable across treatment arms.)

The measurements are done at the individual level

The analysis is also done at the individual level, while taking the clustering effect as a specification into the model

In ecological GRTs, we average the individual measurements we take at the group level : $Y= ax + b + U$ (group error) + e (individual error) becomes $Y = aX + b + U$ (Group error)

There are two sample size considerations in GRTs: The total # of clusters (K), and the total number of individuals per cluster (increasing the number of groups is more statistically efficient than increasing the number of individuals per group)

You can increase the power of your analysis by adjusting for baseline values of your outcome, however, you add another correlation (pre-post) into consideration

The clusters need to have identifiable common characteristics, but watch out for a large cluster effect (the clusters in one treatment arm have a shared characteristic, whereas those in the other arm do not, resulting in a lack of exchangeability (ie comparing three schools from the same city to three schools from three cities)

During the data collection, you can use complete enumeration of all individuals in the cluster, or use simple random samples

Lecture 5: Overview of GRTs continued

Overview of GRT's (ctn'd)

The ICC (Intra-class correlation) is an estimate of the proportion of the total variance due to the group. It is a measure of clustering, and is calculated by ICC= (Between group variance)/(Between group variance + Within Group Variance). The within group variance in this case is the individual error, and the total variance is the sum of the two.

In GRT's, because the treatment is at the group level, we need to determine is the treatment effect is attributable to individual variations, or the group itself: this means that even small ICC can have a large impact on the inference for our effect.

Naturally, there is a covariance (or correlation) between our group level error and our individual error, when we look at any single group (That is, knowing something about the group can tell us something about its individual members and vice versa). Using randomization in GRT's we break that link and with enough groups randomized, the covariance between the group error and individual error becomes 0.

In order to estimate our effect, we use B/SeB for our effect estimate: B is the signal, and Se is the noise around the signal. If SeB is too small, we will pick up larger signals than we should (false positive). This can happen if we only consider the individual level error in GRT's, and must be corrected by adding the group level variability using our ICC-> This is the design effect, and is possible because there is no covariance between U and e thanks to the randomization.

Ignoring the cluster specification means ignoring the ICC (the proportion of the variance attributable to clustering i.e. between group variance / total variance): Even though the ICC is small, it has large impacts on the statistical inference.

GEE is better to assess between group comparisons, whereas GLMM is better suited for within group comparisons.

ITT is better for evaluating the biological/behavioral effect of an intervention(mechanism/Policy), whereas ATT is better suited for effect in practice (implementation in current practical settings).

Interference, Contamination and Spillover effects in GRT's: Interference is when we have spillover within a group ie when we assign treatment to some members within a group, its effects will spread to other members within the group (this is good!)

Spillover is when we have interference across groups within the same treatment arm. This is not desirable

Contamination is when we have interference across treatment arms. This is the worst case scenario of interference as it limits our ability to identify a treatment effect.

ANOVA comparisons in GRT's

ANOVA	σ^2 (Variance)	Meaning
TX	Effect of interest	The true GRT Comparison is between TX and groups (ie group means across txt conditions) This drives the df in the analysis for the treatment effect = (#groups - 1)*(#conditions)
Group	Error attributable to the group	
Individual	Error attributable to the individuals	We usually ignore the individual effect in the analysis, but more individuals increases the precision of the group means

Lecture 6: Posttest analysis in GRT's

Using simple regression (treating our observations as independent) will yield a correct effect estimate despite being misspecified. What will be wrong is the small standard error (variance over square root of n) because it will be deflated. This leads to a bigger test statistic (larger signal or false positive) and a smaller confidence interval than when the clustering is taken into account.

In more practical terms, when you assume independent assumptions, you use too many degrees of freedom (Z distribution with infinite degrees of freedom) so it is very easy to have a significant value, on top of having a deflated standard error. So even if we specify the correlation, we also have to specify the t-distribution in our model (because we rarely have more than 30 groups per condition to make the Z distribution applicable) so that we properly interpret our results.

Lecture 7: Consort Guidelines in GRTS

The first thing to remember about the consort guidelines is that they are just that: not rules, but guidelines created to encourage transparency when reporting the results of our trials. If the paper is transparent and clear as we see fit, we should not force them to stick to every point cited in the guidelines.

Safety and Stopping rules: We have a limitation to how many times we can look at our data for interim analyses in order to avoid falsely inflating our type one error rate. Therefore, we need a separate data safety monitoring board to make sure that we stop the trial when we lose equipoise (clearly beneficial or harmful results) without compromising the integrity of our data.

Blinding of participants is not usually feasible in GRTs, but blinding of data collectors is an effective way to avoid bias in our GRT (it does not touch issues such as selection bias or contamination however).

The effect size is the coefficient change in the outcome variable due to a standard deviation change in the intervention. We use the standard deviation instead of a unit change in order to compare the effect of different interventions on the same outcome

A successful trial is one that is methodologically sound: It has nothing to do with the outcome. If all you care about is the outcome, then you are just doing research. Even though you use the same methods you would, you are invested in your research so you can't be impartial and produce the rigor that you would using a neutral, scientific approach. For science, a null finding, provided it was methodologically sound is as valuable as a significant result. However, for research, this is not acceptable, and we run the danger of biasing our results to get the findings that we want.

Lecture 8: BLUE vs BLUP

How to evaluate group level (level 2) effects in our model:

The first approach we might try is to include the group level effect as indicator variables for each clusters in our model. In this case, the coefficient estimates for each indicator variables (cluster) is called the Best Linear Unbiased Estimate (BLUE).

- This accounts for clustering & uses the proper df for the clusters, but it is incorrect for GRTs because the indicator variables are based on the assumption that individual observations within each cluster are independent-> this approach would work for a multi-site RCT, because we randomize at the individual level within each site/cluster, removing any correlation between individual participants within the cluster.

Another approach would involve specifying the cluster effects as a random variable: The coefficient estimates given by the random variable are called the Best Linear Unbiased Predictors (BLUP).

$Y = \alpha + \beta*X + Z*\mu + \varepsilon$ (ε is the individual error, and μ is the group level error)

- How are these different from BLUE? They are calculated differently: In brief, the model fits a grand mean for the effect due to the cluster as a fixed effect, then it superimposes on that grand mean the individual values for each clusters to create a normal distribution for our random variable, centered around the grand mean it just calculated.
- if the value for a particular cluster is very far from the center of the distribution (towards either tail of the distribution), it gets adjusted or "smushed" towards the center of the distribution (this phenomenon "shrinkage"). Values which are already close to the center of the distribution don't get adjusted as much.
- Using BLUP in our model costs only one degree of freedom (since it is a random variable), but it does cost us one additional assumption, that we are drawing our random variable based on a random distribution.

Lecture 9: IRBs and GRT's

IRB application for studies have to rely in federal guidelines as stated in the 45 CFR section 46 for human research subjects (Animal research subjects have more stringent rules because humans are capable of consent)

A very important aspect of human research subjects is this notion of informed consent: this is the pillar of ethical research, and was developed after the conclusion of the Nuremberg trials post WWII. Legally, individuals with sufficient cognitive abilities over the age of 18 can give informed consent. Children, though able to distinguish right from wrong around the age of 7, can only give their assent, in addition to supplemental consent from one or both parents/guardians

The rules for needing or waiving parental consent, and how stringent these are, depends on the perceived risk vs potential benefits of each individual study. Section 404 is for studies with minimal risk & no known benefits, section 405 is for minimal risks and direct benefits, and section 406 is reserved for high risk, direct benefits (i.e. new cancer treatments and such).

Because of the nature of public health research and clinical equipoise in trials, we cannot guarantee the benefits of our studies, or very often they may have a contribution to generalizable knowledge, without directly benefiting an individual. Therefore, we want to overestimate the risks and underestimates the benefits, when inviting participants to a research study.

Do not make the mistake of asking for passive consent in your study. Either put the resources towards ensuring informed consent (clear consent form, questions to make sure the participants understood the consent and goals of the study since 40 – 60% of the US population has limited reading comprehension, and incentive for them to return the consent forms if it is for children research or have buy-in from their organization) or provide good justification to wave consent for your study.

Lecture 10: Pre-Post Analyses

Fixed effects: effect of the variable is replicable

Random effects: effect comes from a super population so it will not be replicable

Choosing your testing options: posttest only would be useful to avoid social desirability bias

Pre-post analysis involves making pre measurements prior to randomization, and then calculate the difference (txt – ctl) in differences (post – pre): by using the control arm as a comparison group, we remove the natural change in trend over time (because it should be the same across condition for the outcome)

- Y2 = cond + Y1 (Baseline adjusted)
- Y2-Y1 = cond (Differences in differences)
- Yi = Condi (Repeated measured analysis)

The pre-post analysis compares the results from before the intervention (preferably taken after randomization) to those after the intervention.

Baseline adjustment methods (Y2 = cond + Y1) give us the difference in the predicted value of the outcome when everyone's baseline value is the same ("adjusted" for baseline).

Change in score method (Delta, or Y2-Y1 = condition) gives the difference in the person change on average between treatment and control arms.

Member cohort GRT follow the same person at baseline and follow up, whereas member cross section analyses follow the same group/clusters, but measures different individuals at each time point

Degrees of freedom in a pre-post analysis: by adding time, we go from c(g-1) to (t-1)*c*(g-1). As you can notice, for a simple prepost analysis, the degrees of freedom are the same as a post test only analysis.