SDSU

HealthLINK Center

Grant Writing Workshop Session 3

December 11, 2024 | 10:00am

SDSU HealthLINK Center for Transdisciplinary Health Disparities Research
Funder: National Institute on Minority Health and Health Disparities (U54MD012397 & S21MD010690);
MPIs: Guadalupe X. Ayala, PhD, MPH, MA and Kristen J. Wells, PhD, MPH

Topic #2:

Statistical Analysis; Power and Sample Size Calculations

45 minutes

SDSU
HealthLINK Center

Dr. Chii-Dean (Joey) Lin
Co-Leader, Health Data Analytic Group
SDSU HealthLINK Center
Professor
SDSU Department of Mathematics and Statistics

Dr. Shih-Fan (Sam) Lin
Co-Leader, Health Data Analytic Group & Measurement Methods Group
SDSU HealthLINK Center
Adjunct Associate Professor
SDSU School of Public Health

Basic Concepts in Data Analyses

SDSU
HealthLINK
Center

Outline

- Study Design Plan
  - Study type
  - Study design and experimental unit (EU)
  - Confounding factors
- Analysis Plan
  - Identify statistical methods
  - Data visualization and summary
  - Missing values
  - Validating assumptions

Study Design Plan

SDSU
HealthLINK
Center

Study Types: Observational vs. Experimental Study

- Observational study:
  - A research method used to gather information about a group of subjects without manipulating any variables or conditions.
  - No assignment has been made.
  - Often used to identify patterns, correlations, or associations between variables rather than establishing cause and effect.
  - Examples: Survey data, interview data, EHR

Study Types: Observational vs. Experimental Study

- Experimental study:
  - A research method where researchers manipulate a factor/treatment to observe its effect on an outcome variable(s)
  - Assignments were made
  - Can establish a cause-and-effect relationships
  - Replicable
  - Examples: biomedical experiment, randomized controlled trials, intervention studies
  - CONs: ethical issues, costly, and time-consuming

Understand the Data

- Secondary data? Primary data collection?
- For secondary data, observational or experimental (from previous trial)?
- For primary data collection, observational or experimental?
- Know the target population of the data
- **Important**: Need to assess if the collected data is representative of the target population
- Potential biases:
  - Selection bias
  - Measurement bias
  - Response bias
  - Publication bias

Diverse Perspective Bias and Others

- NIH has been emphasizing on this and provided workshops to address this
- EHR data: need to ensure equitable representation in research spaces and increasing diverse participation
- Avoid certain groups or perspectives are systematically overrepresented or underrepresented in a dataset

Identify Experimental Unit(s) (EU)

- EU: The smallest **unit** that a factor/treatment is applied
- EU and measurement unit (MU) are different concepts
- Sometimes, EU and MU are the same
- Example: ANOVA for evaluating teaching methods
  - EU= Classes
  - MU= Students in the class
- Example 2: Efficacy for a new hypertension drug
  - EU= individual patient
  - MU= individual patient

Identify an Appropriate Study Design

- Various study designs are used to “control” heterogeneous and/or dependent EUs so that any outcome difference is due to the treatment effect.
- Ideal situation: homogeneous and independent EUs
- Different EU assignments yield different study designs
- For heterogeneous EUs:
  - Strategies to employ: stratification, blocking
  - Examples: patients with various health condition, disease stage, income level
- For dependent EUs
  - Strategies to employ: clustering, nesting, longitudinal, repeated measures
  - Examples: patients from the same clinical center, participants from the same church, participants living in the same geographical area

Randomization Scheme Involved?

- Common types of randomization
  - Simple randomization
  - Block randomization
  - Stratified randomization
  - Cluster randomization
- How was the randomization applied?
- Any constraints?
- Different randomization mechanisms yield different study designs
- Self-selected sample, convenient sample – not ideal

SDSU
HealthLINK
Center

Confounding Factors?

- Factors/covariates that are not the focus of our study but could potentially impact the outcome.
- Baseline data is a natural pick
- Demographics variables, health condition index, etc. are usual suspects
- In addition to potential confounders, one should collect more covariates for future analysis (within the budget and time limit)
  - Can be used for subgroup analysis or influential points assessment

Objectives & Study Aims

- Formulate the hypothesis to be tested
- Conduct statistical tests for the hypotheses
- Have sufficient power for the test(s)?
- Need to conduct a power/sample size calculation?
- Which test should be used to perform the power/sample size calculation?
- Power/sample size calculation is usually based on the primary aim
- Important to adjust the significant level for multiple tests

Analysis Plan

SDSU
HealthLINK
Center

Intent to Treat Analysis vs. Per Protocol Analysis

- **ITT Analysis:** All randomized participants will be analyzed according to their original treatment assignment, regardless of compliance.
  - Most commonly used for the primary analysis of randomized clinical trials
  - **Preserves Randomization:** Maintains the balance of original randomization scheme
  - Maintains sample size
  - May result in an effective intervention appearing to be ineffective
  - Affected by the pattern of adherence to treatment strategies

Intent to Treat Analysis vs. Per Protocol Analysis

- **Per-Protocol (PP) Analysis:** PP analysis focuses on the subset of participants who adhered strictly to the protocol..
  - Can investigate the “true” effect of actually having received the assigned treatment strategies
  - **Potential for Bias:** Excluding non-adherent participants can introduce bias, especially if non-adherence is related to the outcome.
  - Less power
  - Need to adjust for incomplete adherence in per-protocol approach

Type of Outcomes

- Continuous
- Dichotomous
- Ordinal
- Categorical

Plot/Summarize the Data

- Examine data versus treatment factors, covariates, time, location, etc.
- Generate histograms, matrix scatterplot, boxplots (large data), spaghetti plot, bar plot, etc.
- Plots grouped by treatment condition.
- Do these plots seem to support the hypotheses of interest?
- Are there any patterns that suggest a violation or concerns about assumptions for the planned statistical analysis?
- Outliers?

Plot/Summarize the Data

- Conduct descriptive statistics- overall, by treatment group, by categorical covariates, etc.
- For continuous outcomes, use mean, standard deviation, min, max, etc.
- For categorical outcomes, use percentage in the frequency table

Statistical Methods for Analyses

- Parametric and nonparametric methods
- For parametric methods, valid assumptions are needed
- If assumptions for a parametric method are violated and cannot be “fixed”, should consider nonparametric approaches or another statistical procedure
- If assumptions are satisfied, parametric approaches are more powerful
- For dependent observations (i.e., clusters, repeated measures), use linear/nonlinear mixed model to address the dependency

Diagnostic Analyses and Assumptions Checking

- Examine outliers and influential observations
- For parametric methods,
  - Normal distribution?
  - Constant variance?

Types of Missing Values

- Missing Completely At Random (MCAR):
  - Missing values occur randomly and are unrelated to the observed or missing data.
- Missing At Random (MAR):
  - Missing values are related to the observed data but not the missing data itself.
- Missing Not At Random (MNAR):
  - Missing values are related to both the observed and missing data.

Missing Values

- It's expected to be addressed for grant proposals nowadays
- Compare covariates for missing vs. non-missing
- Conduct missing imputation
- Multiple imputation methods are standard now
  - SAS Proc MI, Proc Mianalyze
  - SPSS (Missing Imputation)
  - Stata (MI)
  - R: Amelia, MICE, RF
- Conduct sensitivity analysis

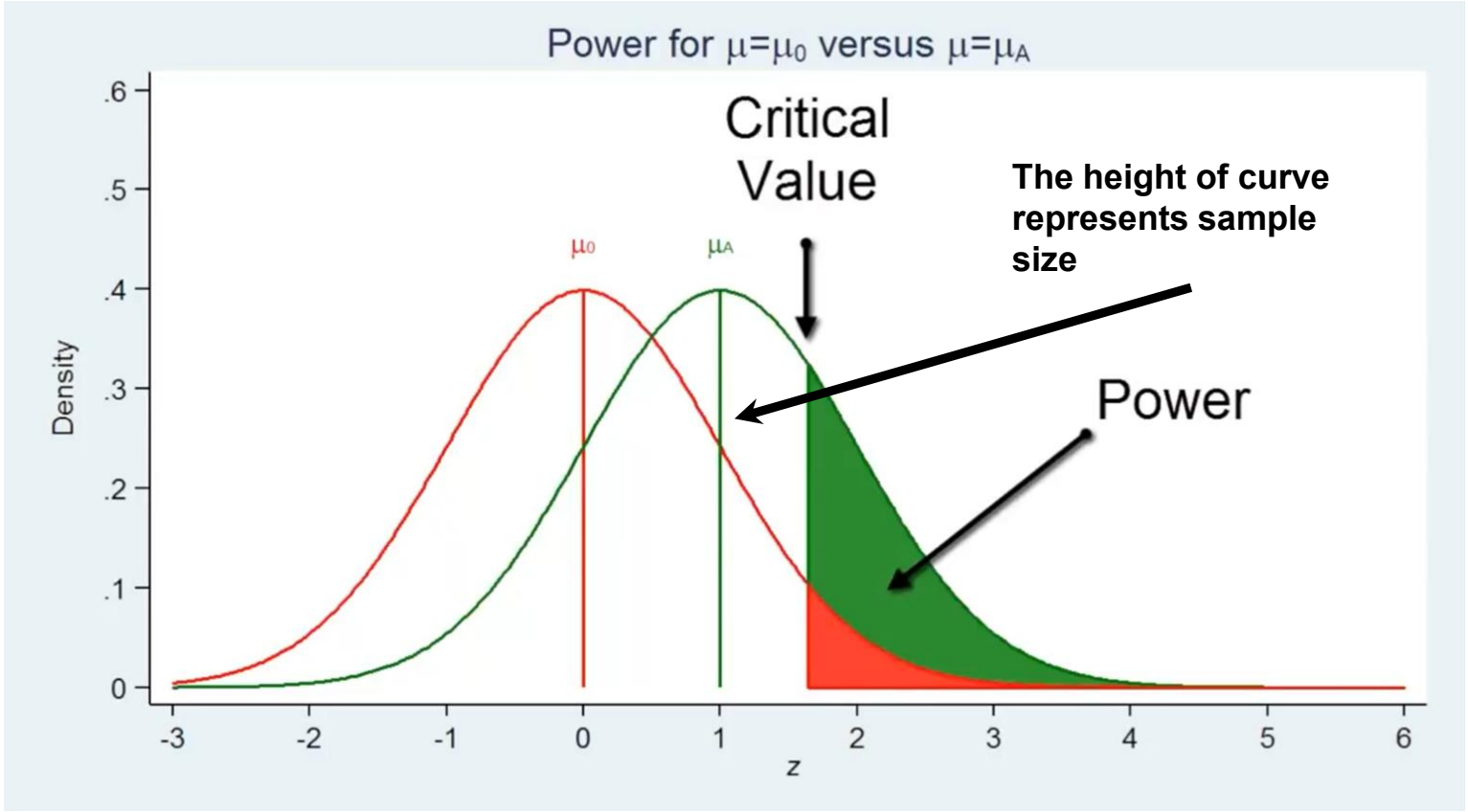
Power & Sample Size Calculations

SDSU
HealthLINK
Center

Power and sample size calculation recap

- Power is the probability to detect a true association or effect in a study
- Sample size is number of subjects needed for the study
- Power and sample size are complementary
  - When calculating the sample size, power is a required parameter
    - Primary data collection
  - When calculating power, sample size is a required parameter
    - Secondary data analysis
- Effect size
  - Magnitude of the effect of outcome to be detected by a test with a specified power
  - Power and sample size can be influenced by the effect size
- Study design can impact power/sample size calculation
  - Balanced group? Clusters? Multiple time points?

Power vs. Sample Size

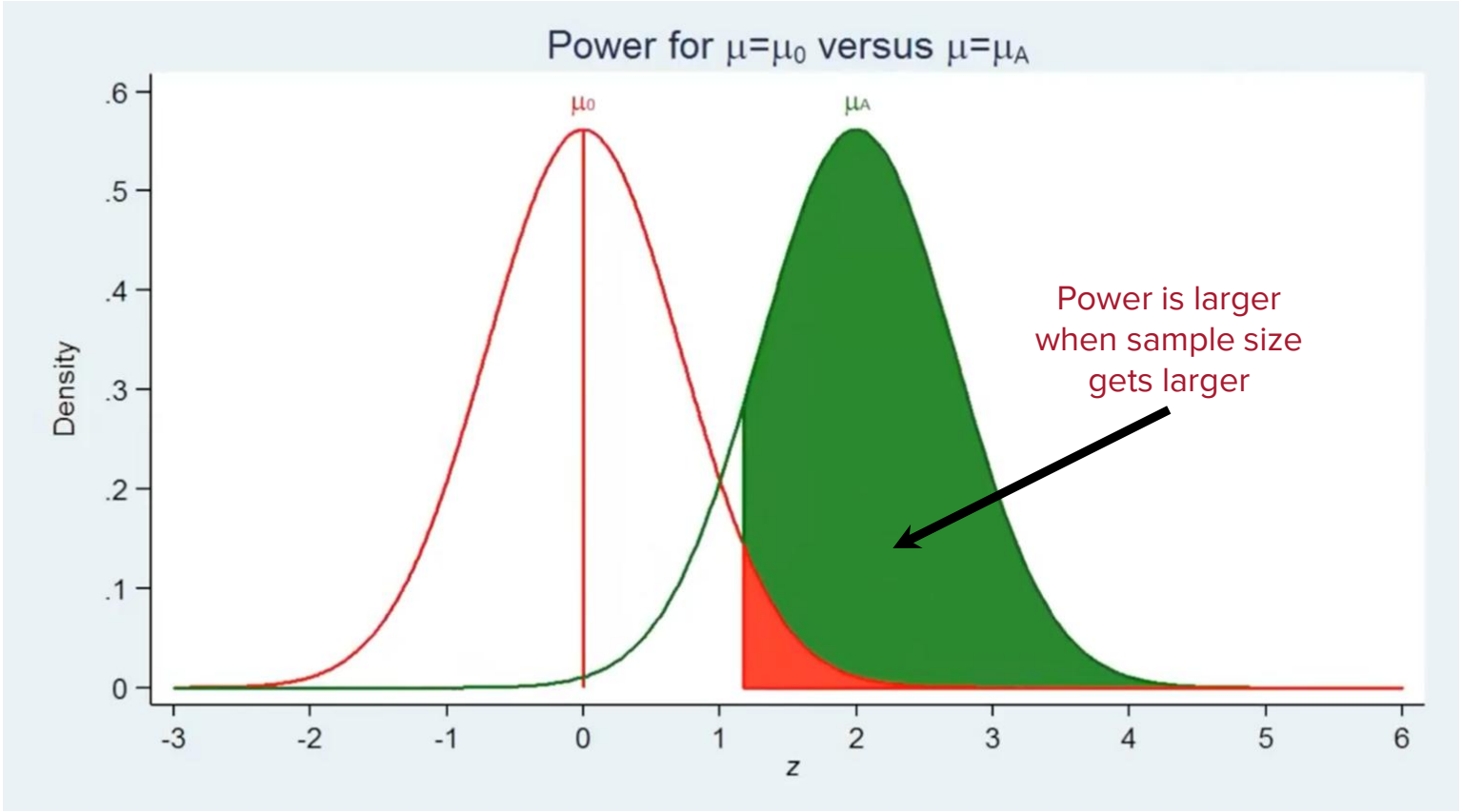


What does increasing the sample size typically do to the statistical power of a study?

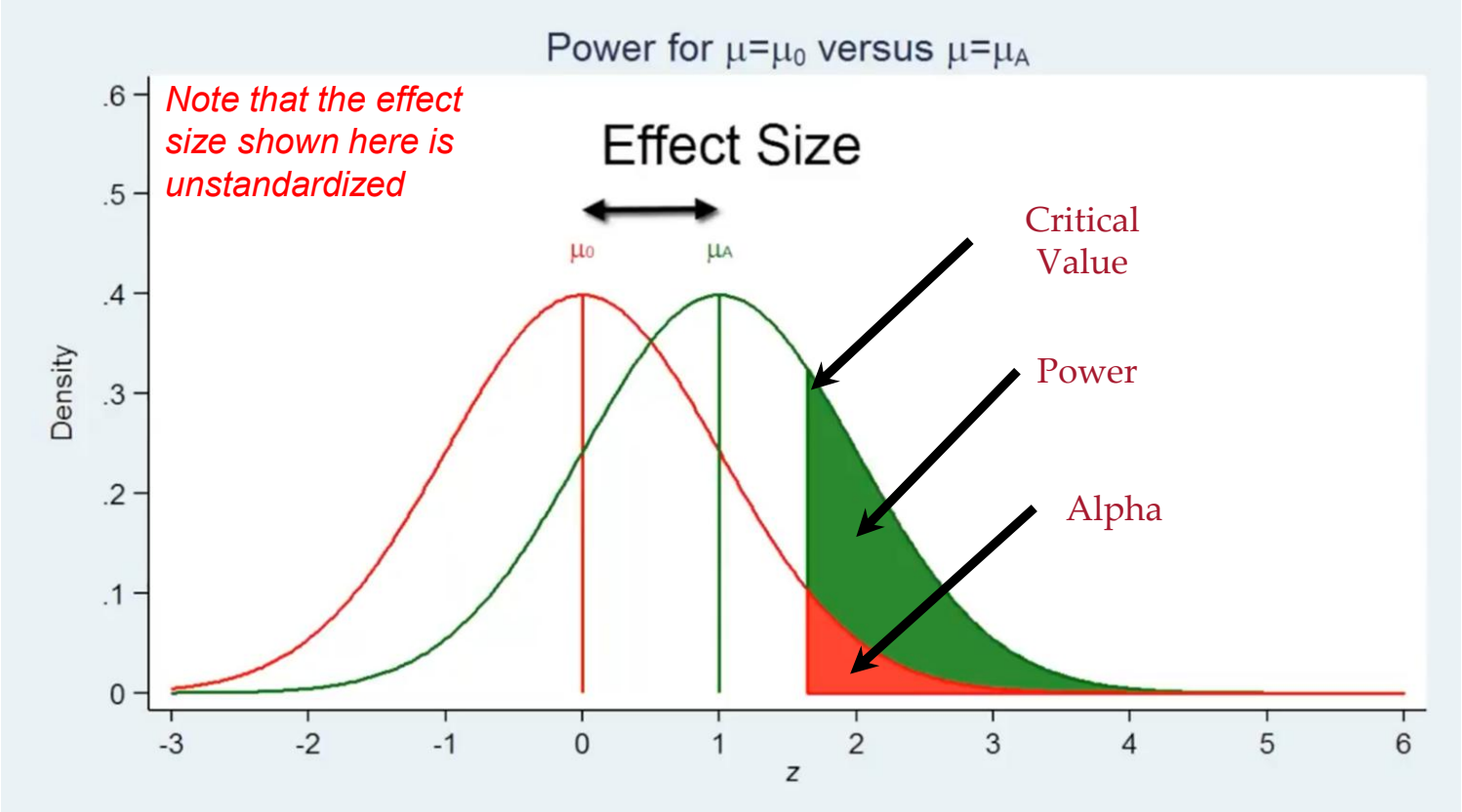
- Decreases power
- No effect
- Increases power
- Makes the effect size smaller

SDSU
HealthLINK
Center

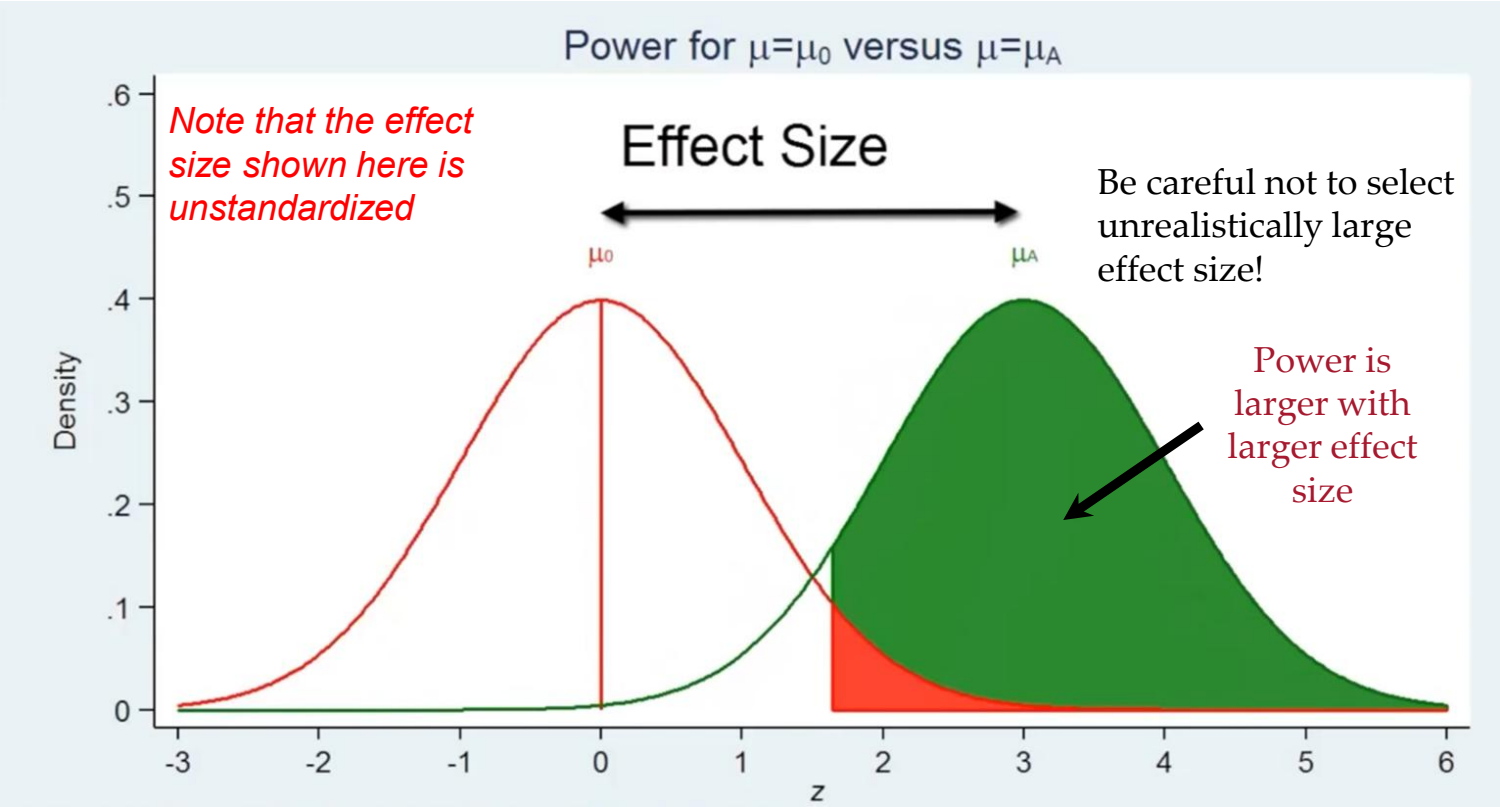
Power vs. Sample Size



Power vs. Effect Size



Power vs. Effect Size



How to determine parameters needed for power/sample size calculation?

Determine the study design*

- Cluster effects
- Number of data collection time points
- Number of confounders
- Unbalanced group
- Multiple group comparisons

*If you have more than one aims, determine the primary aim before moving on to the next step.

Determine the appropriate statistical analysis

- T-test
- Chi-square
- ANOVA/ANCOVA
- Linear/Logistic regression
- Mixed model

Determine the parameters needed for power/sample size calculation

- Alpha level
- Power/Sample size
- Effect size
- Means/Variance

SDSU
HealthLINK Center

Power and sample size parameters

- Alpha (i.e., significant level)
  - One-sided or two-sided. Typically, 0.05 or 0.01
- Sample size
  - Constrained by a determined sample size

  Or
- Power
  - Determined by tradition or funding agency guidelines
  - Traditionally 0.8
- Effect size:
  - Difference between the hypothesized values of the “primary outcome”
  - Based on previous studies or pilot data
  - If no previous studies available, an assumed effect size (e.g., small, medium, or large effect size) can be used
  - The selected effect size should be justifiable
- Mean(s) and variance(s) (if effect size is not provided)
  - Based on previous studies or pilot data

Additional considerations

- For projects with clusters, need Intraclass Correlation Coefficient (ICC) to calculate the design inflation factor
- For longitudinal projects, need within-subject correlation to compute power/sample size
- Multiple comparison tests (conducting multiple tests/post-hoc tests in a study)
  - Need to adjust the individual significance level so that the overall “familywise” error rate is remained at 0.05
  - Can use the Bonferroni adjustment
  - Adjusted $\alpha = \alpha/m$, where m=# of tests conducted in a study
- Drop-outs/Attrition
  - Inflate the sample size by applying this formula:
    - N/(1-attrition rate), where N=calculated sample size

Example 1: Cluster Randomized Controlled Trial – *Aims and Data Analysis*

Aim:

This aim will determine the efficacy of a culturally adapted tele-rehabilitation program compared to usual care in Hispanics/Latinos with chronic spine pain, based on improvement in pain intensity and interference with the Brief Pain Inventory (primary outcome). → Clearly state the purpose of the aim and specify that it is the primary aim of the study.

Analysis:

Intervention effects on changes in our primary (BPI) and secondary outcomes between baseline, 1-wk Post, and 3-mos Post assessments will be analyzed using mixed effect models to adjust for clustering by clinic site. Time by condition (CBPT vs. Usual Care) interactions will be added to the models to assess for intervention effects. If the interaction is not significant, the main effect for time and condition will be assessed. → Reiterate the data collection timepoints and describe the statistical procedure used to assess the primary outcome.

For all outcomes, the models will be adjusted for sociodemographic and clinical characteristics (see Table 1, last row) that may affect treatment response [e.g., sex (male vs. female) and age as biologic variables, community (urban vs. rural). and spine pain (neck vs. back)]. → Describe whether covariates will be adjusted in the model.

SDSU
HealthLINK Center

Example 1: Cluster Randomized Controlled Trial – *Aims and Data Analysis*

Missing Data:

All analyses for Aims 2 and 3 will use an intent-to-treat (ITT) approach. Although withdrawal bias is always a risk with longitudinal cohorts, retention rates for members of the investigative team are high (e.g. 96-98% for prospective study on risk factors for chronic neck pain,[38] and 80-98% for longitudinal observational and intervention studies in Hispanics/Latinos[56]). The characteristics of the individuals with missing values and the individuals with no missing values will be studied. Multiple imputation (SAS Proc MI and Proc MIANALYZE) technique will be performed. Longitudinal profiles will be described using descriptive statistics (such as means and standard deviations) and graphical statistics (such as spaghetti plots and trellis graphs) of the repeated measures over time to visualize trends and patterns of change.

→ Describe the plan to handle missing data, including the procedure that will be performed.

Sensitivity Analysis:

Sensitivity analysis will be conducted based on both Missing At Random (MAR) and Missing Not At Random (MNAR) assumptions and results using imputed data and non-imputed data will be compared.

→ Describe any sensitivity test (e.g., outliers, alternative model specification, subgroup analysis…etc.) that will be done

Example 1: Cluster Randomized Controlled Trial – *Sample size calculation*

The sample size estimate is based on the Brief Pain Inventory (BPI) pain interference score as primary outcome for Aim 2. → Determine the primary outcome.

In our previous CBPT intervention for post-surgical spine pain[36], a significant 1.3-point reduction in pain interference score was found in the intervention condition compared to the control condition at the 3-mos Post. Thus, an effect size of 0.5 (mean diff=1.3; sd=2.5; effect size=1.3/2.5=0.5) was used for our sample size calculation. → Determine the effect size based on pilot study (effect size=0.5).

An initial calculation determined that a sample size of 132 participants (66 per condition) will achieve 80% power to detect a difference of 1.3 in pain interference given an alpha of 0.05 → Determine the alpha level and power (80%) and compute the sample size using all of the parameters provided.

However, we adjusted the sample size calculation based on the design effect of individual participants nested within two clinic sites (or clusters). The design effect is 1+(m-1)*ICC where m is the average size of participants in a cluster (m=66). Based on previous studies conducted in primary care settings including musculoskeletal conditions and chronic pain, the median ICC reported was 0.01.[54] Considering these estimates yielded an inflation factor of 1+(66-1)*0.01 =1.65, which would require 132*1.65≈220 participants (110 per cluster). → Make additional adjustment for the clustering effect in the design.

Assuming a 20% attrition rate, and an equal distribution of participants in each condition at each site, we will recruit 220/(1-0.2)≈276 participants (138 participants per group). → Make another adjustment for expected attrition rate.

SDSU
HealthLINK Center

Example 2: Biomedical Research Study– *Aims and Data Analysis*

Aim:

Characterize the antibacterial activity exerted by direct contact of T. vaginalis with L. crispatus and L. iners.

→ Clearly state the purpose of the aim. If more than one independent experiment are proposed, each experiment should have its own analysis plan and power/sample size calculation.

Analysis:

For all assays, three biologically independent experiments will be performed utilizing triplicate technical samples in each independent experiment. → If you are repeating the same experiment, clearly indicate so in your analysis plan as the sample size would affect power.

For the viability experiments, a 2 x 2 factorial design will be tested: time (30 mins vs. 3 hours) x incubation condition (L. crispatus or L. iners alone vs. cocolonization with T. vaginalis). Appropriate two-way analysis of variance (ANOVA) statistical tests will be performed with Tukey's multiple comparison test will be employed to contrast the combination of time and incubation categories. → Describe the statistical procedures that will be used to test the outcome. If you need to do multiple comparison across groups, it should be described in the data analysis plan.

SDSU
HealthLINK Center

Example 2: Biomedical Research Study– *Power calculation*

Using a significance level of 0.008 (adjusted for multiple comparison tests),

→ Determine the significance level (alpha). The alpha level was reduced due to the adjustment of multiple testing.

a within-cell standard deviation of 4.76 X 10^6 (based on preliminary data), means and effect sizes found in our preliminary experiments (1.3 for time and 1.8 for incubation condition), we will obtain sufficient power to test the main effects (83% for time, 98% for incubation condition) and the interaction effect (94%) given 12 samples (4 cells x triplicates=12).

→ Standard deviations and means were provided from the pilot study and effect sizes were computed

→ Since the budget dictates number of samples tested in each aim, statistical power was computed based on number of sample that can be afforded.

SDSU
HealthLINK
Center

5 Minute Break

SDSU
HealthLINK
Center

Topic #3:

Data Management and Sharing (DMS) Plan

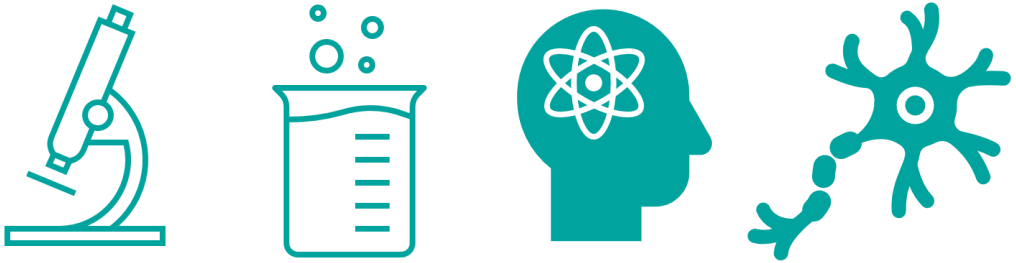
SDSU HealthLINK Center

20 minutes

Dr. Shih-Fan (Sam) Lin
Co-Leader, Health Data Analytic Group & Measurement Methods Group
SDSU HealthLINK Center
Adjunct Associate Professor
SDSU School of Public Health

Ms. Yaritza Benitez
Data Analytic and Measurement Coordinator
SDSU HealthLINK Center

Data management and sharing (DMS) plan recap



- NIH's definition of scientific data
  - Data commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications

SDSU
HealthLINK
Center

What data needs to be shared?

Share

- Adequate data to validate and replicate study findings
  - Both ***quantitative and qualitative*** data
- Data resulted from the study but not supporting a publication
- Null finding that did not result in publication

No need to share

- Laboratory notebooks
- Preliminary analyses
- Completed case report form
- Drafts of scientific papers
- Plans for future research
- Peer reviews
- Communications with colleagues
- Physical objects such as laboratory specimens

SDSU
HealthLINK
Center

When should data be shared?

No later than:

time of associated publication or

end of award period, whichever comes first

SDSU
HealthLINK Center

Acceptable reasons for **NOT** sharing data

- Existing consent prohibits sharing
- Privacy or safety of research participants would be compromised or place them at greater risk of de-identification or suffering harm
- Explicit federal, state, local, or Tribal law, regulation, or policy prohibits disclosure
- Datasets cannot practically be digitized with reasonable efforts
- Proprietary data

SDSU
HealthLINK Center

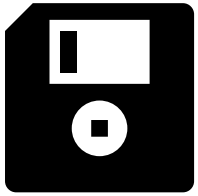
DMS plan resources

- NIH Data Sharing: https://sharing.nih.gov/
- DMS Plan
  - Template
    - https://grants.nih.gov/grants-process/write-application/forms-directory/data-management-and-sharing-plan-format-page
  - Sample plan with different types of studies
    - https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan#after
- Common Data Elements recommendations
  - Center's adult demographic survey (English and Spanish)
    - Highly encouraged to include this in your DMS plan
  - CDE with NIH's endorsement
    - NIH Toolbox/PROMIS/Neuro-QoL/ASCQ-Me
    - PhenX Toolkit
    - NIH Common Data Element Repository

Center requirements to share project-related documents

- Pilot Project Leaders will be expected to share the following:
  - Project summary
  - Dataset link (with instructions if needed) to access the data from your selected data repository
  - Metadata
  - Project associated data documentations (not data per se)
    - **Data collection instruments** (e.g., soft copy of survey, interview questions…etc.)
    - **Dataset codebook** (including variable name and label, frequency counts/percentage, response value label, standard missing codes)
    - **Protocols** (e.g., recruitment protocol, interview/focus group guides, assessment protocol such as anthropometric measurements, air quality measurement, sample processing protocol….etc.)
  - Project associated dissemination products (e.g., manuscripts, posters, presentations)
- Such information will be stored in the Center's ***Health Science Research Portal (HSRP)*** to share publicly.

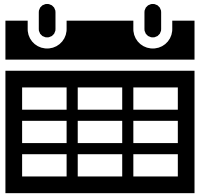
Six elements of DMS plan



1. Data type

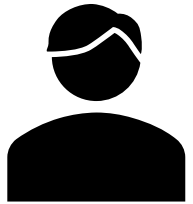
2. Related tools, software, and/or codes

3. Standards



4. Data Preservation, Access and Associated Timelines

5. Access, Distribution, or Reuse Considerations



6. Oversight of Data Management & Sharing

SDSU
HealthLINK Center

Element 1A Example: Types and amount of data to be generated

Environmental Study

As part of Aims 1 and 2, we will collect several types of data associated with environmental sites in this project: (a) microbial microscopy images with embedded metadata (in the Nikon .nd2 format, which can be opened using ImageJ), (b) metagenomic FASTQ sequencing files and.CSV files with comparative, annotated data, (c) chemical identification data collected by the non-targeted method, and (d) caregiver interview data. This project will generate approximately 2-4 TB of data. Quantitative interview data will be collected during caregiver interviews conducted as part of the home visits planned for Aim 2. Caregivers will provide open- and close-ended responses to questions on barriers to environmental control strategies, among other variables. → Clearly state the types of data generated and the size of the data.

Behavioral Intervention

Quantitative data will be collected from parents and teachers who will provide ratings of children's mental health symptoms, academic, social, and behavioral functioning prior to participating in the intervention, at post-intervention, and at a follow-up time point during the subsequent school year (for participants in the cluster randomized trial only). They will also provide their perceptions of the intervention at the post-intervention time point and parents will report their skill use during each intervention session. → Clearly state the types of data generated and when data will be collected

Qualitative data will be collected from four primary sources, including school administrators, school mental health providers (SMHPs), teachers, and parents of children with ADHD during focus groups and key informant interviews. Focus groups and interviews will be professionally transcribed and translated, and transcriptions will be coded by research team members.
→ Clearly state the types of data generated for qualitative portion of the research

Element 1B Example: Data to be preserved and shared

Environmental Study

All metagenomic sequencing files (raw and annotated) and microscopy images, will be preserved and shared to encourage data sharing and reproducibility (detailed in Element 4). Preservation of all formats will be done physically on external solid-state hard drives in the laboratories (i.e. SDSU College of Health and Human Services servers), and sequencing data will also be backed up onto SDSU HealthLINK Center's secured open data server, with consultation and support of the Center's IT Group. Sequencing data will additionally be archived in repositories (detailed in Element 4). Caregiver interview data will be captured in a de-identified fashion and stored using Qualtrics. Focus group data will be stripped of all identifying information and stored in password-protected files on the servers.

→ Clearly state where all of the data will be preserved and what data will be shared.

Behavioral Intervention

All study data will be de-identified and preserved on HIPAA-compliant servers through SDSU's College of Sciences. Cluster randomized controlled trial assessment data will be collected and stored via REDCap. Data exported from REDCap will be stored in the server discussed above for data cleaning and analysis. De-identified quantitative data related to the open and cluster randomized trials will be shared upon publication to promote open science communication (detailed in Element 4).

→ Clearly state where all of the data will be preserved and what data will be shared. Name of the repository will be discussed in Element 4.

SDSU
HealthLINK
Center

Element 1C Example: Metadata and associated documentations

Environmental Study

Detailed study protocols (robust descriptions of sample collection and extraction methods, processing strategies, and data analysis pipelines), dataset codebook, metadata, and project summary will be submitted to the Health Science Research Portal (HSRP) as mandated by Center policy. File Title Nomenclature for Organization – All files, once collected, will be labeled with the date in the file title, according to ISO 8601 for easy searchability. Common identifiers will be provided in the sample name, including the de-identified sample location and sample type (fecal, aerosols, etc.). Metadata Spreadsheets with Sample Information – The complete information, containing all de-identified pieces of information about each sample, will be provided in spreadsheet form with the Project description in the data repository. → Clearly state what documentations will be shared and where they will be shared.

Behavioral Intervention

Study protocols (e.g., recruitment protocol, intervention protocol, training and supervision protocols, …etc.), dataset codebooks, metadata, and a project summary and will be submitted to the HSRP as mandated by the Center policy. In addition, study-specific data collection instruments, such as survey, fidelity rating scales, and the mHealth application user's manual will also be shared publicly in the HSRP. → Clearly state what documentations will be shared and where they will be shared.

SDSU
HealthLINK Center

Element 2 Example: Tools, software, and/or codes

Environmental Study

No specialized software is needed to access the data. Image files can be opened using ImageJ, and large data files such as metagenomic data can be opened as CSV files in any spreadsheet software, and raw files can be opened through free software such as R. All data analyses will be performed using R.

→ Clearly state what software is needed to open the files you shared.

Behavioral Intervention

Data will be cleaned and processed primarily using SPSS. Data files will be stored in .dat and .csv formats, both of which can be exported directly from SPSS. All data cleaning and processing syntax will be saved and shared. Data shared in .dat/.csv format does not preserve variable labels and response value labels. However, a detailed codebook will be provided that includes variable labels and response value labels for each variable in the dataset. The .sav file, which contains variable/response labels will also be shared in case users also prefer to analyze data in SPSS.

→ In addition to stating the software needed to open the file, it is encouraged to share the programming codes to read in flat files (e.g., .csv, .tsv, .dat, .txt) or apply variable labels and response value labels.

Element 3 Example: Standards

Environmental Study

Sequencing Data – All sequencing data will be submitted to European Nucleotide Archive (ENA), which adheres to Minimum Information About a Next-generation Sequencing Experiment (MINSEQE) standards. We will go through each point on the checklist to ensure compliance. FASTQ files, CSV differential expression outputs, and CSV metadata files will be provided. Caregiver interview demographic data: we will follow Center requirement to adapt the Center's adult demographic items in their caregiver interview to maximize dataset interoperability. → Clearly state what standards will be applied for each data type. Also, note where Center's adult demographic items were mentioned.

Behavioral Intervention

We will follow Center's requirement to include the Center's adult demographic items for all adult participants. In addition, since the data will be submitted to the NIMH Data Archive (NDA), we will adhere to NDA's data structures in terms of variable name, variable label, value labels, and data types.

→ In addition to stating the use of Center's demographic items, using the data structure standard required by the data repository was also stated.

Element 4A Example: Repository where data & metadata will be archived

Environmental Study

All sequencing data, including metadata, will be submitted and archived with the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home). The ENA is a global nucleotide repository, with sequencing data from projects on metagenomics, microbiome, transcriptomics, and more. Distance from the "point" of exposure, used to assess correlations in this project, will be provided in a non-identifiable way (PII and geographic information removed) as ENA metadata and additional supplemental information will be provided with the resulting publication. → Clearly state where data and matadata will be archived. Note that the website for the repository was provided but it was **NOT** hyperlinked.

Behavioral Intervention

The RP Leader will make data publicly available through the NIMH Data Archive (NDA). We will work with NDA staff to develop a data submission schedule and outline data elements to be submitted. Throughout the project, the data collection will be carefully organized and documented following best practices for data file management to allow for data sharing.

→ Clearly state where data and matadata will be archived

Element 4B Example: How scientific data is findable and identifiable

Environmental Study

Upon submission to ENA, all projects are assigned an accession number in their database. The project accession number will be referenced in all resulting publications and reports for ease of access. A thorough project description will be provided so that the information is easily searchable within ENA, even without the accession number. All samples will also have adequate descriptions to complement the metadata so that project organization is obvious.

→ Clearly state whether a unique persistent identifier will be assigned to the dataset(s) submitted to the repository.

Behavioral Intervention

Datasets submitted to NDA are assigned a data submission ID. This identifier will be referenced in all publications and presentations using data from the study to link disseminated findings to the database.

→ Clearly state whether a unique persistent identifier will be assigned to the dataset(s) submitted to the repository.

Element 4C Example: When & how long the scientific data will be available?

Environmental Study

Metagenomic Data – Data submission will occur prior to, or concurrent with, the time of first publication submission or the conclusion of the funding period (whichever is first). Data will be made publicly available (embargo lifted) at the time of first publication release or the end of the funding period, whichever comes first. Data will be available in perpetuity. **Analytic Data** – The dataset used for regression analysis, including distance to the impacted environmental sites (described as TJRE in the RP) will be provided as metadata for the ENA metagenomic submission. The data will be available to the research community in perpetuity. → Clearly state when the data will be submitted vs. available if the data is not immediately available after submission. Also, provide information on how long data will be shared

Behavioral Intervention

All relevant data will be shared at the time of publication or by the end of the funding period (whichever is earlier). Data deposited in NDA will be available to the research community in perpetuity.

→ If data is immediately available after sharing, just state when the data will be shared and how long the data will be shared. Most likely the data will be shared in perpetuity.

SDSU
HealthLINK Center

Element 5A Example: Factors affecting subsequent access, distribution, or reuse of data

Behavioral Intervention **Environmental Study**

Participants' consent for data sharing will be obtained. For public sharing of the data via the data repository described above, there are no additional limitations other than control (**Element 5B**) and privacy protections (**Element 5C**) described in the following sections.

→ Clearly state whether consent for data sharing will be obtained. Make it clear that there will be no additional limitation other than control and privacy protections describe in the next sections. If you decide to do a **broad consent**, make sure you mentioned it here as well. **For those who are accessing Electronic Health Records (EHR), make sure to indicate whether a Data Use Agreement (DUA) or a Business associate Agreement (BAA) will be obtained. If the covered entity states that data sharing is prohibited in the DUA or BAA, make it clear in this section.**

Element 5B Example: Whether access to scientific data will be controlled

Environmental Study

All deidentified data will be made available via the public repository described above. Users of the data will likely need to register for an account with the repository and agree to the Terms of Use. All scientific data will be freely available via public data repositories after release.

→ Clearly state how user would need to access the data and whether there are restrictions for accessing data.

Behavioral Intervention

All deidentified data will be made available via the public repository described above. The NDA requires users to create an account, agrees to the Terms of Use and submit a Data Access Request. Upon approval, users will be able to access data in the repository. The Terms of Use will protect the study participants by limiting the use of data to scientific research and only allowing dissemination of aggregated statistical reporting. In addition, attempts to identify research participants are prohibited and users may be required to notify any disclosure of study participant identity.

→ Clearly state how user would need to access the data and whether there are restrictions for accessing data. Note that the public data repository requires users to submit a Data Access Request Form and the Terms of Use will protect human subjects' privacy.

Element 5C Example: Protections for privacy, rights, and confidentiality of human research participants

Environmental Study

Sequencing information is not considered human genomic data. However, since metagenomic data obtained from hand swabs may also contain human sequence data, we will remove host (human) sequences using Bowtie2. All data, including metadata will be de-identified prior to sharing, and all personally identifiable information will be removed. Deidentified data characteristics will be retained for reproducibility.

→ Clearly state the measures that will be taken to protect participants' identities.

Behavioral Intervention

The privacy, rights, and confidentiality of human research participants will be protected by replacing the direct identifier with study IDs. All datasets derived from data collection will only include study ID as the unique identifier. Only certain research staff will have the access to participant contact information for recruitment and data collection scheduling and incentive distribution purposes. Once data collection is completed, access to participant contact information will be removed except for Principal Investigator and Project Manager/Coordinator for future contact purposes. Access to contact information in REDCap after the end of data collection will be limited to those with Principal Investigator and Project Manager/Coordinator roles. Study participants will be asked to provide consent to data collection and sharing of data to the wider research community. They will also be informed that the information will be disseminated in the form of an aggregated statistical report.

→ Clearly state the use of study ID to conceal participants' identity and describe the purpose to obtain contact information and when this information will be removed.

Element 6 Example: Oversight of Data Management and Sharing:

Environmental Study

Compliance with this plan will be monitored and assured by both RP Co-Leaders, Drs. XXX and YYY. While both Co-Leaders are directly involved in data generation, we will be responsible for data sharing and management. As a project funded by the Center, we will also communicate these timelines and steps with the Center's IT team as large sequencing files are backed up onto Center's open data server. Progress of the data management and sharing will also be reported in the Research Performance Progress Report (RPPR).

→ Clearly state who will monitor the plan to ensure compliance, timeliness of data sharing, and data quality.

Behavioral Intervention

Compliance with this plan will be monitored by the RP Leader, Dr. XXX. He will oversee the acquisition, cleaning, processing, and documentation procedures and will work with research project staff to establish a timeline for preparing and sharing key data elements. Progress of the data management and sharing will also be reported in the Research Performance Progress Report (RPPR).

→ Clearly state who will monitor the plan to ensure compliance, timeliness of data sharing, and data quality.

SDSU
HealthLINK Center