Fundamentals of clinical research
3: Designing a research study
(continued)

Logan Danielson

ABSTRACT

The design of a research study is further elucidated in this article. In the process of
describing the topics needed to design a research study, we will continue to build
upon our stepwise framework for conducting and publishing research. The depth of
these topics has been curated to provide the reader with the general idea, clear up
popular misconceptions and elucidate how the reader might deepen their knowledge.

This article will conclude with what a researcher should expect from this point forward
and how to deal with it.

Key words: Clinical Research; Research study design; Evidence-based Medicine

SERIES INTRODUCTION

This series of articles is meant to provide the reader
with a framework from which to efficiently conduct
research. The content presented is intended to be of
benefit to both junior and senior researchers, as a firm
understanding of the fundamentals is always essential
to performing at the highest level. This series of five
papers on research methodology for young scientists
includes:

1. The Importance of Clinical Research
   Danielson L. Anaesth Pain & Intensive Care
   2017;21(3):289-291

2. Fundamentals of clinical research 1: Selection of a
   suitable and workable research proposal
   Danielson L. Anaesth Pain & Intensive Care
   2017;21(4):485-488

3. Fundamentals of clinical research 2: Designing a
   research study.
   Danielson L. Anaesth Pain & Intensive Care
   2018;22(1):131-138

4. Fundamentals of clinical research 3: Designing a
   research study (continued)
   Danielson L. Anaesth Pain & Intensive Care
   2018;22(2):268-278

5. Fundamentals of clinical research 4: Writing and
   publishing an original research article
   Tentative – Expected publication date is September
   2018

Picking up where we left off

In the previous article of this series, we introduced the
idea that having a method is what leads to professional
success and with this we dived into a discussion of
research methodology. We started with defining the
common types of studies before talking about data.
And we concluded with a brief overview of statistics
and the basis for hypothesis testing. In this article,
we will continue our discussion of statistics prior to
discussing the associated characteristics of a research
study, like the sample size and appropriate statistical
tests to use for analysis.

The two types of errors

There are two types of errors that one can make when making a conclusion regarding their hypothesis testing. The type I error can be defined from two standpoints: 1) Falsely rejecting a true null hypothesis or 2) Claiming one’s hypothesis as true when in fact the null hypothesis is true. Both standpoints say the same thing, but they do so slightly differently and both are included here to ease the reader’s understanding. The Greek letter α is used to portray the chance of committing a type I error. When our test statistic, i.e. the p value, is less than α we consider our result to be statistically significant. The type II error can also be defined from two standpoints: 1) Failing to reject a false null hypothesis or 2) Failing to claim a true hypothesis as true. The Greek letter β is used to portray the chance of committing a type II error. β is used in the definition of the power of a study.

The power of a study

The power of a study is equivalent to the chance one rejects the null hypothesis when it is false. Given that β is equal to the chance of failing to reject a false null hypothesis, it follows that: Power = 1 – β. This is important because for a given effect size, α and Power, the number of study participants needed to prove the statistical significance of an effect can be calculated, and we’ll touch on this subject more later in this article.

Bias and confounding

The statistics used in a study are meant to account for the random error induced by studying a sample of a population instead of the entire population. What statistics can’t account for is nonrandom bias. Bias refers to systematic errors in the collection or interpretation of data that causes deviation from the truth. There are many types of bias and knowing of them helps one avoid accidently incorporating them into their study design, data collection and analysis. These biases can be broken down into two groups: selection biases and observational biases.

Selection biases are due to inappropriate selection of or poor retention of study participants and it has the following subtypes. Ascertainment or Sampling bias is said to occur when the study population differs from the target population due to the use of nonrandom participant selection methods. Sampling bias can occur from either omission or inclusion. Omission bias occurs when certain groups are omitted from the sample, for example pregnant women, children, elderly, particular ethnic groups, the wealthy, the poor, the unemployed, or prisoners. It should be noted that this type of bias can, at times, be unavoidable. Thus, our goal as researchers is to recognize it as a potential source of bias and consider how it affects the conclusions one can draw from the results of your study. Inclusive bias occurs when samples are selected for convenience and the sample ends up not being an adequate reflection of the population. Volunteer or Referral Bias, occurs when those that want to be involved in a study are different from those that don’t. Verification bias, which is also known as Detection bias, Workup bias and Test Referral bias, is the bias that can be introduced by the selective use of a single “gold standard” diagnostic test to determine participant inclusion. This can lead to bias because, of those with the disease, those that have done the test may differ from those that have not. Compliance Bias refers to the fact that patient compliance tends to lead to better outcomes, regardless of screening. Nonresponse bias occurs when a high nonresponse rate to a survey cause systematic errors due to nonresponders differing in some way from those that did respond. Berkson bias occurs when the disease studied only uses “sicker” hospital-based patients and thus loses applicability to the target population that is outside the hospital. Length Time Bias occurs when a screening test over-represents a less aggressive form of a disease, because a disease with a slow progression will be asymptomatic for longer than one with a fast progression and as such there is there is a longer interval of time for which a screening test can diagnose a disease with a slow progression than one with fast progression, which means that any randomly chosen interval or time point of screening is more likely to diagnose a disease with a less aggressive presentation because the more aggressive forms are more likely to have developed symptoms prompting their diagnosis or death. The Lead Time Bias is also known as the Zero Time Bias, it occurs when a person with a disease is screened earlier and thus diagnosed earlier and thus appears to live with the disease for longer, when in fact, the only difference is that they knew about their diagnosis earlier than someone that was screened later in their presentation. The Over-Diagnosis Bias refers to the overinterpretation of screening tests as positive. The Prevalence or Neyman bias refers to studies missing patients who have short courses due to either a quick recovery or accelerated death. Attrition bias refers to the bias caused by the loss of study participants when those lost differ significantly to those that remain.

Observational biases are due to inaccurate
designing a research study

measurement or classification of disease, exposure, or other variable and it includes the following subtypes. Instrument bias refers to instruments reporting incorrect results and this is why instruments must be calibrated and aberrant results confirmed with additional measurements. Measurement Bias is also known as Information Bias, it occurs when measurement is consistently dissimilar in different groups of patients due to differences in data collection technique. Recall bias refers to the fact that participants with negative outcomes are more likely to report certain exposures than control participants. Observer bias refers to observers misclassifying data due to individual differences in interpretation or preconceived expectations. Procedural bias affects surveys and occurs when participants rush through the survey because the situation has led to them being motivated to finish the survey quickly, not properly. Reporting bias occurs due to participants either over or under reporting exposure history due to perceived social stigmatization. This is similar to the Hawthorne Effect, the bias induced by study participants when they change their behavior because they know they are being studied. Surveillance or Detection bias refers to the possibility that the risk factor itself caused increased monitoring in those exposed to it which increased their probability of identifying a disease relative to those that were not exposed simply because they were looking for it. Reporting bias affects meta-analyses, because research that is deemed statistically significant is more likely to be published, thus a meta-analysis of the published research is less likely to consider the trove of data that wasn't statistically significant when considered by itself.2

Confounding bias is due to a relationship between factors in a study. For example, consider a study that concluded that heavy alcohol use increased lung cancer risk, but, upon closer review, noted that the majority of heavy alcohol users were also heavy smokers. In this study, the true cause of the increased lung cancer risk could be better attributed to the smoking, not the alcohol use, as such the only way to draw an appropriate conclusion regarding alcohol use and lung cancer risk would be to account for the confounding in the data analysis. It should be noted that confounding can work either direction, i.e. it can help reject your null hypothesis or prevent that rejection.

To be a confounder, a variable must meet three criteria, it must be: 1) associated with both the outcome & the exposure; 2) independent of the exposure; and 3) not be a consequence of the exposure. Examples of types of variables that should be assessed as confounders include: age, known risk factors, and prognostic factors.

Controlling for confounders can be done either before or after the data is collected. Controlling for confounders before data collection can be done via: 1) Randomization, which ensures known and unknown confounders are evenly distributed amongst study groups; 2) Restriction, limiting participants to one category of confounder (e.g. only men); and 3) Matching, which involves enrolling controls that are similar to the cases on potential confounders. Alternatively, confounders can be controlled for after data collection via different data analysis techniques, including: 1) Stratification, which separates data into the strata of the confounder; 2) Mantel-Haenszel adjusted relative risk, which can be thought of as weighted stratification; and 3) Multivariate Analysis or Modeling, where advanced statistical techniques, like logistic regression, are used to control for confounding.

Alas, this list of biases is incomplete. There will be times when bias in one's research is unavoidable simply due to the nature of the experiment or the environment in which the experiment is conducted. While one should do what they can to minimize potential sources of bias, the most important thing to do is to think about the biases that can affect one's work and openly discuss in one's research how the conclusions drawn could be affected by bias and what was done to minimize the impact of that bias.

Association and causation

In our research, we want to prove causation. While association can be observed, causation can only be inferred. An exposure is associated with a disease when the frequency of disease differs according to presence (or intensity) of exposure. An exposure causes a disease when it has led to one or more individuals developing the disease.

There are several criteria for proving causation. 1) The variable and outcome must be associated, preferably with a high relative risk. 2) The exposure must precede the development of disease. 3) A dose response relationship must be established. 4) Reduction or elimination of exposure results in less disease. 5) The findings must be consistent between studies. 6) The causation must be plausible. 7) Other possible explanations, e.g. confounding, have been considered and ruled out.
Arguments to infer causation can be impaired by: 1) Latency, i.e. the delayed manifestation of effects post-exposure. 2) The same effect occurring due to other causes. 3) Causal factor only manifesting the outcome when additional factors are also present.

Interpreting results

The interpretation of results is an important skill that depends upon the use of other calculations and their interpretation. Some of the most common statistical measures used for binary classifications are described in Figure 1 along with their method of calculation.

The Positive Predictive Value (PPV) is the proportion of Positive results that are True. The Negative Predictive Value (NPV) is the proportion of Negative results that are True. It is important to note that both the PPV and NPV depend also on the prevalence of the disease. Also, while the PPV and post-test probability are calculated in the same manner, the term “PPV” should be used to refer to what has been established with control groups, while the term “post-test probability” should be used to refer to the probability for an individual. Sensitivity is also known as called the true positive rate, the recall, or the probability of detection. The sensitivity measures the proportion of positives that are correctly identified, it can also help to think of it as proportional to the lack of false negatives—Sensitivity: lack of FN. Specificity is also called the true negative rate. Specificity measures the proportion of negatives that are correctly identified, it can be thought of as proportional to the lack of false positives—Specificity: lack of FP. Likelihood ratios are used to assess the value of a diagnostic test. The positive Likelihood ratio (+LR) refers to value of a positive test result and is calculated by Sensitivity / (1 - Specificity), where values greater than 1 increase the probability of disease. The negative Likelihood ratio (-LR) refers to value of a negative test result and is calculated (1 - Sensitivity) / Specificity, where values between 0 and 1 decrease the probability of disease.

Risks and Ratios

There are many methods used to describe how a factor influences an outcome. The Absolute Risk is the incidence of risk with or without the risk factor. The Attributable Risk (AR) is also known as the Risk Difference and it represents the incidence of an outcome that can be attributed to the exposure. It is calculated from the absolute risk with the exposure minus the absolute risk of those without the exposure. The Relative Risk (RR) is also known as the risk ratio, it represents how much more likely an exposed person is to develop an outcome. The relative risk is calculated from the absolute risk with the exposure divided by the absolute risk of those without the exposure. The Odds Ratio (OR) is subtly different from the relative risk, it, in a way, represents how much more likely a diseased person is to have had the exposure. This difference is best seen from the calculation, of which is described in Figure 2. The relative risk is a better measure to report than an odds ratio, but it is also the more difficult number to calculate, as it technically requires using the entire population as a sample. Conveniently, when a disease is rare, the odds ratio approximates the relative risk. When a disease is not rare, the odds ratio overestimates the relative risk. Understanding the difference between these two values is important as authors have been known to confuse them in the past.1

Hazard ratios represent the instantaneous risk of some endpoint over the study period or a part of

---

### Figure 1: How to calculate Positive Predictive Value, Negative Predictive Value, Sensitivity & Specificity

<table>
<thead>
<tr>
<th>Test outcome</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Positive (TP)</strong></td>
<td>Positive predictive value</td>
<td>False Positive (FP) (Type I error)</td>
</tr>
<tr>
<td><strong>False Negative (FN) (Type II error)</strong></td>
<td>True Negative (TN)</td>
<td>Negative predictive value</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td></td>
<td>Specificity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition (as determined by the Gold standard test)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td><strong>Negative</strong></td>
</tr>
<tr>
<td><strong>True Positive (TP)</strong></td>
<td>Positive predictive value</td>
</tr>
<tr>
<td><strong>False Positive (FP)</strong> (Type I error)</td>
<td></td>
</tr>
<tr>
<td><strong>True Negative (TN)</strong></td>
<td>Negative predictive value</td>
</tr>
</tbody>
</table>

### Figure 2: Comparing the calculation of the relative risk with the odds ratio

<table>
<thead>
<tr>
<th></th>
<th>Diseased</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exposed</strong></td>
<td>$D_e$</td>
<td>$H_e$</td>
<td>$T_e = D_e + H_e$</td>
</tr>
<tr>
<td><strong>Not exposed</strong></td>
<td>$D_n$</td>
<td>$H_n$</td>
<td>$T_n = D_n + H_n$</td>
</tr>
</tbody>
</table>

$$RR = \frac{D_e / T_e}{D_n / T_n}$$

$$OR = \frac{D_e / H_e}{D_n / H_n}$$

---

designing a research study

the study period. This is different than the relative risk and odds ratio, of which represent a cumulative risk over the whole study period. Hazard ratios are calculated with regression models and are depicted with Kaplan-Meier survival curves.

The Absolute Risk Reduction (ARR) is equivalent to the experimental event rate (EER) minus the control event rate (CER). The Number Needed to Treat (NNT) is equivalent to the inverse of the ARR. NNT are useful as they allow one to express results in natural numbers, that are easily understood by non-experts. Stated plainly, the NNT is the number of patients that must be treated to produce 1 event. This number is important for pharmacoeconomics as it makes it simple to see the cost of producing the 1 event of interest. The Relative Risk Reduction is equivalent to the ARR divided by the control event rate. The RRR expresses your result relative to that of the control group. This can be useful, but, it can also obscure the reader’s understanding of the true treatment effect, when the control outcome rate is either very high or very low.

Regarding Significance

The results of a study are always interpreted with regards to their statistical significance. The statistical significance of a result answers the question of whether the result found is likely to be adequate representation of the population studied and not simply a byproduct of sampling error or other forms of random error. The clinical significance of a result is inferred by the author and it elucidates whether the result is of clinical importance. Both forms are influenced by sample size. A small sample can cause clinically significant results to not reach statistical significance, while large samples can cause clinically insignificant results to be statistically significant. Performing an adequate power analysis will help prevent these incongruencies. There is another significance though that we do not often talk about in discussions of research, that is, the significance of patient preference. The point being here that the best scientific-support doesn’t necessarily trump the significance of patient preference and that while we should know what is best supported, that knowledge should be used as a framework to help guide the care of the patient, seen as an individual, not mandate their care to a fixed path.

The Triple Aim

The triple aim was developed by Berwick, Nolan, and Whittington. They created this framework as a means of improving the U.S. health care system and, in it, they specified three goals worth simultaneous pursuit: improving the patient experience, improving population health, and reducing health care’s per capita costs. The Per capita cost can assess what the actual cost is as well as a discussion on whether that cost is worth the investment. Population health can be assessed with the NNT as well as the expected average benefit. The experience of care can be assessed by how well the patient tolerates the new treatment compared to the standard treatment. When one includes assessments of these factors in their study, the study’s marketability for publication rises. That’s not to say that every study needs to include these factors. Instead it is mentioned here to motivate the reader to think about these types of factors and those related to them at the beginning of their study so that they can include them in a meaningful way, when it makes sense to do so.

Step 3: Determine the associated characteristics of your study

Sample Size

When designing a study, trade-offs will be made based on designing the study you want and one that is feasible. One of the first characteristics to decide upon is the sample size. Using too large of a sample size wastes resources, while using too small of a sample size risks increases the chances of making a type I or type II error. The industry standard power is 80% and α of 0.05, with these numbers set and a minimum detectable/significant difference established, the required sample size can be calculated. The method of calculation can depend on the type of study and, no matter what method is used, it must be described when you present your research. If you need a number to perform this calculation and you don’t have it, then you should look to the literature to see how other authors have approached the issue. If you choose to use their number for your sample size calculation, mention it when you describe your research, if you choose to use a different number, mention how you came to that number and substantiate the decision. Your chosen sample size should consider that some participants may fall out of the study.

Inclusion and Exclusion Criteria

Inclusion and exclusion criteria must also be decided upon. These criteria can include factors such as: age, sex, race, ethnicity, type and stage of disease, previous treatment history, as well as the presence or absence of other medical or psychosocial conditions. The criteria provide three functions: 1) ensure patient safety; 2) justify participant appropriateness for the
study; and 3) minimize participant withdrawal. In general, no participant should be excluded unless there is a reason to exclude them. Factors that strongly justify exclusion include: 1) patients unable to provide informed consent; 2) placebo or study intervention would cause harm; and 3) the effect of the intervention would be challenging to interpret.5

Variable types

One also needs to determine how they plan to record their variables. The previous article in this series delved into variable types but didn't stress why the understanding was important. As an example, please consider a retrospective study where the researcher is interested in each patient's SaO2. As they collected their data, they realized that the patient with a 95% SaO2 on room air isn't truly equivalent to the one with a 95% SaO2 on 2L O2/min, so they recorded each patient's oxygen requirement as well. And with this, they created their problem: by trying to interpret two continuous variables simultaneously (SaO2 & O2 requirement), they, in effect created a two-dimensional continuous variable type. While this could be handled with a regression analysis, it would completely change their study. One solution to this would be to create a series of categories based on “oxygen status” with ranges of SaO2 & O2 requirements, but that solution imposes its own set of problems. Another solution would be to check the literature to see how other groups have handled the issue. If you use their solution, mention it. On a side note, referring to and doing as those before you have done is a useful tool and it makes any meta-analysis done with your data and theirs more accurate as there is greater homogeneity between each group's definitions.

Statistical tests

The statistical tests to be used to analyze the data should be determined prior to beginning the study. Failure to do this is likely to result in cherry-picking of one's statistics, which essentially increases one's likelihood to commit a type I error. The statistical test chosen depends on a multitude of factors, for example: variable type of the dependent and independent variables, number of independent variables, presence of repeated measurements, data distribution, and what is of interest: relationships, differences or non-inferiority? The type of test chosen matters because each test makes assumptions about the data that it is analyzing and, as such, violating the test's assumptions violates the validity of the test. While it is best to choose one's statistical tests at the beginning of the study it is possible that one could unknowingly chose inappropriate statistical tests, and this brings us to our next point: before testing one's hypothesis against the data, the appropriate action is to investigate the data with summary statistics and charts. This step helps one identify patterns in the data as well as outliers. For continuous data that is normally distributed, one should use means and standard deviations to summarize their data. For data that is skewed, the median and interquartile range should be used. This step enables one to determine whether the statistical tests they have chosen are appropriate for the data they have gathered, prior to seeing the results those tests produce. If the tests previously chosen are inappropriate, then it would be appropriate to switch to a more suitable test or, if applicable, they can transform the data, as we will discuss later (Table 1).

Table 1: Summary of descriptive and graphical statistics

<table>
<thead>
<tr>
<th>Variable type(s)</th>
<th>Chart</th>
<th>Summary Statistic(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Categorical</td>
<td>Pie or Bar Chart</td>
<td>Category percentages</td>
</tr>
<tr>
<td>Two Categorical</td>
<td>Stacked Bar Chart</td>
<td>Percentages within Categories</td>
</tr>
<tr>
<td>One Continuous</td>
<td>Histogram</td>
<td>Mean &amp; Standard Deviation</td>
</tr>
<tr>
<td>One Continuous +</td>
<td>Boxplot</td>
<td>Median and IQR</td>
</tr>
<tr>
<td>One Categorical</td>
<td>Line Chart</td>
<td>Means by time point</td>
</tr>
</tbody>
</table>

Generally, the type of variable dictates the tests that can be used, but some statisticians do make exceptions. For example, ordinal data with seven or more categories can be analyzed with parametric tests, as if it were continuous data, so long as the data is approximately normally distributed. Do note though that most surveys use a scale of 5 and as such these surveys shouldn't be analyzed with parametric tests. Another exception involves whether one should use regression or the ANOVA test. Regression can only be used for continuous or binary independent variables, while ANOVA works best with categorical variables. If a variable has only a few categories, one could recode the categories as multiple binary independent variables and hence use regression on what was categorical data.

ASSUMPTION OF NORMALITY

Parametric tests require data that is approximately normally distributed, non-parametric tests do not. When a parametric test can be used, it should be used
because parametric tests are more likely to achieve statistical significance when the effect or relationship sought truly does exist. Normality can be assessed for large samples with a histogram, and, for small samples, a QQ plot. As non-parametric tests make no assumptions about the distribution of the data, non-parametric tests can also be useful when data has unequal variances or when the normality is hard to determine.

Other common assumptions

The Chi-squared test has a couple assumptions. The assumptions mentioned here are not meant to serve as a complete list, they are instead meant to serve as examples of types of data set characteristics that one should look for as they investigate the data set’s validity. The first assumption of the Chi-square test is that 80% of the expected cell counts are equivalent to at least five. Some software suites will warn you if a single expected cell count is below five, this warning can be safely ignored so long as the assumption of meeting 80% of cells is still valid. It should also be noted that we are referring to the expected cell count, not the observed cell count— the expected count for each cell is the row total multiplied by the column total divided by the overall total. If this assumption is not met, then one could either use Fisher’s exact test or they could merge categories to enable use of the Chi-square test, if it made sense to do so and the action did not result in the loss of normality. On a side note, Fisher’s exact test is computationally extensive, especially for large tables, so some statistical suites will only perform Fisher’s exact test on 2x2 tables. The second assumption of the Chi-square test is that there are no cells with an expected frequency below one.

The Kruskal-Wallis test shall serve as our next example of common assumptions. From Table 2, you will remember this test as the non-parametric equivalent of the one-way ANOVA. These assumptions are more generalizable than those of the Chi-square test, but not every test will require all of these assumptions be met. The first assumption is that all observations are independent, if one patient’s results affects another patient’s results then this assumption has been violated. The second assumption is of similar sample sizes between groups. This assumption would be

<table>
<thead>
<tr>
<th>Table 2: Common Single Comparison Tests*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing:</td>
</tr>
<tr>
<td>Averages of 2 independent groups</td>
</tr>
<tr>
<td>Averages of 3+ independent groups</td>
</tr>
<tr>
<td>Average difference between paired (matched) samples</td>
</tr>
<tr>
<td>3+ repeated measurements</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Test of Association*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing:</td>
</tr>
<tr>
<td>Relationship between two continuous variables</td>
</tr>
<tr>
<td>Predicting the dependent variable / Assessing for relationships</td>
</tr>
<tr>
<td>Assessing relationship between two categorical variables</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4: Test to use for one scale dependent and multiple independent variables*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st independent variable</td>
</tr>
<tr>
<td>Continuous</td>
</tr>
<tr>
<td>Nominal (Independent groups)</td>
</tr>
<tr>
<td>Nominal (repeated measures)</td>
</tr>
<tr>
<td>Nominal (Independent groups)</td>
</tr>
<tr>
<td>Nominal</td>
</tr>
</tbody>
</table>
violated by a case-control study that matched three controls to each case. The third assumption is that there are at least five data points per sample.

Tests for other common assumptions

The presence of equal variances is referred to as homoscedasticity. It is an assumption made by tests that compare the means of independent groups, e.g. Independent t-test and ANOVA. The Levene's test can be used to assess the homogeneity of variances. A significant test statistic from the Levene's test implies that the assumption of equal variances is not valid.

The presence of equal variances among the differences of repeated measures is referred to as sphericity and it is an assumption of the repeated measures ANOVA test. Sphericity is assessed with Mauchly’s test and a significant test statistic implies that the assumption is not valid. The Greenhouse-Geisser correction can be used to account for an invalid assumption of sphericity.

Correcting for multiple tests

When one performs multiple tests, the chance of a type I error rises significantly. This can be best visualized by imagining the probability of flipping a coin a certain number of times and getting heads five times in a row. If one were to flip the coin 5 times, then the probability of getting heads five times in a row would be 1 in 32 or about a 3%. If we compare this to the probability of flipping a coin 20 times and getting heads five times in a row at least once in those 20 flips, then we see that the probability of our “significant event” having occurred is now about 25%—note, it is a “significant” event because its 3% probability corresponds to a p-value of < 0.05.

Point being, the more you check for something, the more likely you are to find it, so one should take this effect into account when running a multitude of tests. Moreover, the p-value refers to the chance of a type I error occurring for that one test, it does not reflect the chance of type I errors in your study as a whole. This aggregate rise in type I error can be accounted for in multiple ways and the common post hoc tests for doing so include: Tukey’s test, the Sidak correction, Scheffe’s test and the Bonferroni adjustment.

Transforming data

Table 3 mentions “Transform data” as the test to use as the non-parametric equivalent to simple linear regression, this was meant as short hand for transform the data and then run simple linear regression. Data transformation is the application of a function to all data points such that each data point takes on the value implied by the function. Transformations are done for two reasons: 1) to better approximate the assumptions of a statistical test or 2) to improve readability of a graph. There are many methods of transformation, these methods include: differencing, curve fitting (gradient descent, Gauss-Newton and the Levenberg–Marquardt algorithm), Log transformation, Square root transformations, Square transformations, Cube Root Transformations, Reciprocal Transformations, Box Cox transformation, Tukey ladder of powers, and the Fisher Z-Transformation. Superiority, Equivalence and Non-inferiority (Table 5)

All the statistical tests mentioned so far are meant for hypotheses that look to prove that a difference exists. The burden of proof for these tests lies in having found strong support that the difference truly exists. If one is trying to establish equivalence or non-inferiority, then all the statistical tests mentioned thus far are inappropriate as they place the burden of proof on the wrong hypothesis. To better understand the importance of the difference, one can think about how a court room’s defendant’s defense would change based on whether the defendant is assumed, by the court, to be innocent or to be guilty. Traditionally (as with the tests mentioned thus far), the defendant is assumed innocent until their guilt is proven. Taking this out of the metaphor, the researcher is the court that assumes there is no “significant” difference (i.e. guilt) between the experimental group (i.e. the defendant) and that of the control group (i.e. the guiltless jury) until the difference is proven to exist. Thus, here, the burden of proof is on proving guilt, i.e. proving the difference exists. If we compare this to a hypothetical court room where one is guilty until proven innocent (as is the case with equivalence and non-inferiority testing). Then we start with the assumption that a difference exists and the burden of proof lies in finding evidence that it doesn’t exist. So, in our hypothetical court where guilt is assumed, if the defendant failed to prove their innocence then they would be deemed guilty. Taking this out of the metaphor, if one tried to use a non-significant p-value from a student’s t-test, a test that is meant to prove a difference exists, as evidence of equivalence, then the appropriate conclusion of the court would be that equivalence (i.e. innocence) could not be proven. The distinction to understand here is that the quality of evidence required to come to the appropriate conclusion depends on what the court assumes to be true. To “prove” something scientifically, one must do so with a strong argument, the type of argument...
The new treatment is not designing a research study benefit & risk-benefit assessment. Consistent with the effect of the current therapy; the overall cost-seriousness of the clinical outcome; the magnitude equivalency margin should take into account: the choice being appropriate. The determination of the credibility of the study entirely depends on that must justify that choice with sound clinical reasoning judgement. No matter what value of the value of can be a somewhat subjective clinical efficacy benefits of the new treatment. In this way, clinically acceptable difference that is worth the non-standard therapy.

The equivalency margin, $\delta$, is defined from a context of the equivalency margin, $\delta$, equivalence means that the value of the treatment is within $\delta$ units of the standard therapy. Non-inferiority refers to the new therapy not being more than $\delta$ worse than the standard therapy.

The equivalency margin, $\delta$, is defined from a standpoint of practicality. It represents the maximum clinically acceptable difference that is worth the non-efficacy benefits of the new treatment. In this way, the value of $\delta$ can be a somewhat subjective clinical judgement. No matter what value of $\delta$ one chooses, one must justify that choice with sound clinical reasoning as the credibility of the study entirely depends on that choice being appropriate. The determination of the equivalency margin should take into account: 1) the seriousness of the clinical outcome; 2) the magnitude of effect of the current therapy; 3) the overall cost-benefit & risk-benefit assessment. Consistent with its importance, the sample size required by the study will depend on the size of the equivalency margin.

The most widely used statistical test for equivalence and non-inferiority is the two one-sided test (TOST). Using TOST, equivalence is established at the $\alpha$ significance level, if the $(1-2\alpha) \times 100\%$ confidence interval for the difference in efficacies is within the equivalency margin. So, for example, a 90% confidence interval yields a 0.05 significance level for testing equivalence. This should be noted in comparison to the traditional comparative study which, at the same significance level, would require a larger 95% confidence interval.

On a related note, the first table in most articles typically describes the characteristics of the patients in the control and experimental groups of the study and it has the purpose of proving that the 2 groups were established randomly by showing how the groups are demographically equivalent. The statistical tests that most authors use to “prove” this equivalence are tests designed to prove that the groups are different, yet the authors argue that the lack of statistical significance for these tests implies that the groups are equivalent. While this is technically incorrect, you'll find many published articles that make this assertion. Personally, I don't know why this is the norm, but I believe it occurs for two main reasons: 1) a lack of understanding of statistics and 2) the statistical rigor that we hold this section to is less given that we assume that our random assortment of patients to groups led to their random assortment, but if you are going to assume that it worked, then there's no reason to try to prove it with a test statistic. My suggestion for authors is to do what you think is appropriate. If you want to champion proper statistics, use an equivalence test to prove demographic equivalence and, when asked why you didn’t use the same test everyone else uses, be prepared to explain it. If you want to just not worry about it, then just do what everyone else is doing.

Step 4: Obtain IRB/ethics committee approval

It would be irresponsible to not discuss the importance of gaining the approval of your organization’s ethics committee prior to beginning one’s research. Atrocities have been committed in the name of science before and those actions and their perpetrators are universally condemned. As such, it is safe to say that, in the name of science, the ends do not justify the means. If your organization does not have an ethical review board, then you should talk to the ethical review board that is closest to you, with regards to distance or association, as appropriate. The U.S. Department of Health & Human Services has created a document listing the laws, regulations and guidelines regarding human subject protections of 130 countries and it can be found at this citation. Frequently, ethical review boards have a series of forms that must be filled out to describe the nature of your
study and how you plan to protect the participants. While the board that reviews these forms often seem to be looking for keywords and phrases to validate your study is ethical, it is advised that one sees these forms, not as rote, but as a reminder of what needs to be done to ensure that one’s research is carried out properly. Having an ethical board’s approval is one thing, knowing that you’ve thought about each step of your research and what you can do to protect and do right by the patients you seek to help is another. The forms of the ethics committee will cover topics like: risks and benefits for participants, informed consent, conflicts of interest, methodology for keeping personal information private, and policy for storing and transmitting data, etc.

This section will come to a close with the reminder that it is only ethical to compare two treatments that are believed to be of near equal utility. Comparing a new drug to a placebo when there is an established standard treatment is not ethical. Also, if during your study, a treatment is recognized as being vastly superior, then the only ethical choice is to end the study and provide all patients the superior treatment. It’s understandable that a researcher may want to continue the study because they know they performed an adequate sample size calculation and thus they know how many participants are required to recognize a statistically significant effect, but this thinking is misguided, ethical conduct in research is more important than the results it can deliver. On a side note, the researcher can probably be relieved because, if they can truly tell that one treatment is significantly better than the other prior to the data analysis when the experiment has been properly randomized and blinded, then it would follow that the statistical analysis would show relevant results as well. It should also be noted that research is often done as part of a team and while one researcher might believe there is an obvious superior treatment, the rest of the team may not have yet come to that conclusion, in this case, the team should come to a consensus on whether there is a superior treatment and then proceed in accordance to that decision.

**Step 5: Do research**

If you gotten this far along in the process, you should take a moment to congratulate yourself, you’ve done a bounty of work. Alas that work is truly just beginning. On the path to publication, you will encounter setbacks, so do your best to acknowledge them and be undeterred. This part of conducting research is a marathon, not a 100-meter dash, remind yourself that you are in this for the long haul and persevere for publication.

**Step 6**

Step six will be covered in the final article of this series. In this final article we will discuss how to prepare and publish a research manuscript. The covered topics will include: manuscript design, scientific writing, how to display research results, the selection of a journal for publication, and tips for success with the peer review process.

**Summary of steps:**

- **Step 0:** Develop and learn from your research question
- **Step 1:** Define your null hypothesis
- **Step 2:** Determine what type of study is needed
- **Step 3:** Determine the associated characteristics of your study
- **Step 4:** Obtain IRB/ethics committee approval
- **Step 5:** Do research
- **Step 6:** ?
- **Step 7:** Publish

**Conflict of interest:** Nil declared by the author.
designing a research study

REFERENCES

1. Andale S. Verification bias / detection bias: definition. Statistics How To. [Free full text]