Journal of the Academy of Forensic Nursing

Research Corner

Trustworthiness of Research Results: Significant or Not?

Catherine J. Carter-Snell PhD RN SANE-A DF-AFN [1]
Shaminder Singh PhD, MSc (Psych) RN [2]

Received: October 25, 2023 Accepted: November 22, 2023

© Carter-Snell & Singh 2023. This is an Open Access article distributed under the terms of the Creative Commons-Attribution-Noncommercial-Share Alike License 4.0 International (http://creativecommons.org/licenses/by-nc-sa/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly attributed, not used for commercial purposes, and, if transformed, the resulting work is redistributed under the same or similar license to this one.

Corresponding author: Catherine Carter-Snell
Rm Y355, Mount Royal University,
4825 Mount Royal Gate SW, Calgary AB CanadaT3E 6K6
Email ccartersnell@shaw.ca

Affiliations: 1- Professor, School of Nursing and Midwifery, Mount Royal University
2- Assistant Professor, School of Nursing and Midwifery, Mount Royal University

Abstract

Evidence-informed practice relies on integration of trusted research into decisions about care in collaboration with the patient's wishes, available resources, and professional knowledge. Determining whether the research is trustworthy, the professional requires an understanding of the quality of the research and potential for errors. Nurses receive a basic research course in their baccalaureate training, but sometimes find it is difficult to apply research knowledge, relying only on word-of-mouth best practice statements. Assessing the trustworthiness of research is important to treatment decisions, to patient teaching, and to use of the evidence in court. The focus of the article is to review core concepts central to error and design. Using the decision tree method helps recognize and apply error and design concepts to determine trustworthiness. By using the method, all nurses can examine current and future study findings with confidence to support and defend evidence informed decision-making in their practices.

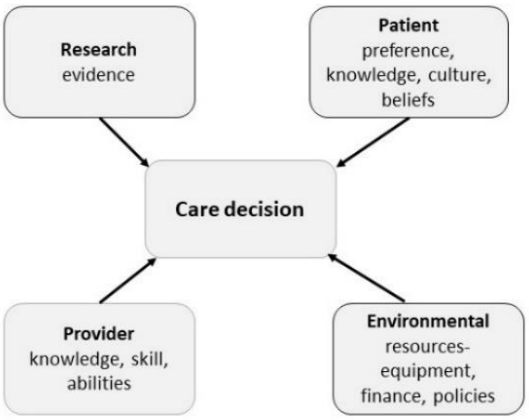
Keywords: research, evidence informed decision making, type I error, Type II error, significance, power

Introduction

The importance of supporting nursing practice with research is well recognized. The concept of *evidence-**based** medicine* emerged in the 1990s, which focused on ensuring all practice was supported by the best research (Guyatt et al., 1992. Evidence-based medicine gives preference to high levels of research quality, specifically the randomized controlled trial (RCT) or systematic review. These will be discussed later, but the designs control for error and alternate explanations for results. There are not RCTs available to support all nursing practices and are sometimes not possible or ethical to conduct. In addition, it is also recognized that research is not the only source of evidence to support practice. The authors of the evidence-based medicine expanded their definition to include the role of clinician expertise (Sackett et al., 1996). There are additional sources, however, including and patient culture, experience, or beliefs, resources, and the environment (physical, social, political) in which practice occurs. Alternatively, the concept of *evidence-**informed** practice* has been proposed (Kumah, McSherry, Bettany-Saltikov, & van Schaik, 2022. Evidence-informed practice uses all forms of evidence (Figure 1) to reach a collaborative decision with the patient (Kumah, McSherry, Bettany-Saltikov, & van Schaik, 2022). Evidence-informed practice is also seen as more inclusive of critical thinking than evidence-based practice (Kumah, McSherry, Bettany-Saltikov, van Shaik, et al., 2022), which is important for nursing practice.

Figure 1.

Evidence-informed Practice



Used with permission C. Carter-Snell

If considering research evidence for potential use in practice, nurses must assess the quality of the research before deciding whether to use it in practice. The research results may help support decisions to use new products or treatments, to answer patients' questions about treatments they read about, or to support updated nursing curricula. If required to give expert opinion when testifying as a nurse, the essential element is that the nurse understand the difference between opinion and the quality and trustworthiness of the science supporting their clinical decisions and interventions. For the nurse, the essential element is to ensure that the scientific evidence is safe and of sufficient quality to support the nurse's decisions.

Two cases are included in this article to illustrate how to make trustworthiness decisions. One case has significant results, and one has non-significant results. After reviewing some core research concepts, these cases will be used to illustrate the use of a decision tree will be used to help guide you through determining a study's trustworthiness.

> **Case 1.** A medical supply representative is encouraging a hospital department to purchase a new type of skin disinfectant. He produces a study the company sponsored in which they had a large sample of female patients in two hospitals – three hundred in one and 420 in the other – who had each undergone laparoscopic surgery for various conditions. Hospital A used the traditional disinfectant pre-procedure and Hospital B used the new one. The outcome, rates of infected patients as measured by white blood counts and differentials, was significantly lower for hospital B (with the new disinfectant). They concluded that their product was safer for clients.
>
> **Case 2.** A client asks you about the effectiveness of a new method of contraception. There are not a lot of studies available, but the one the medical supply representative shows the unit is published in a reputable journal. They compared the new method (Brand Z) with a well-known similar method (Brand A), both progestin based. Women ranged from 18 to 30 years in each group. A sample of twenty women in each group were already taking one or the other method of contraception and researchers compared unplanned pregnancy rates. The rates of unplanned pregnancy were significantly higher in the Brand A group, with researchers concluding that there is no difference between the contraceptives.

The two scenarios highlight the importance of diminishing the word-of-mouth recommendations for practice and promote understanding about how to judge the trustworthiness of research findings. Factors that falsely create significant results (type I error) and factors preventing readers from finding significant differences (type II error) are two types of errors. The focus of the article is to guide readers through the factors that help decide whether to trust the results regardless of their significance.

Significance Measures

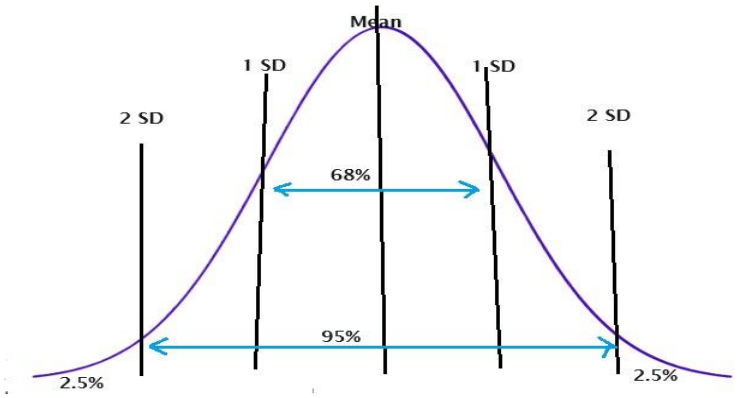
One way to assess trustworthiness is to start by looking at significance of the study results. Then the study design, controls and participants are examined in the study methods to determine whether to trust the significance or lack of it. Two of the main indicators used for significance are the use of probability or *p*-values, and the use of confidence intervals.

Probability to Determine Significance

The *obtained probability* value (p-value) is the assessment of the likelihood of results occurring by random chance. In other words, *probability* is the word to describe how likely something is to happen, like flipping a coin. Each flip is 50% likelihood of landing on heads (or tails). If data are *normally distributed*, half of the coin flips will be heads, and the other half will be heads, split by a normally distributed (or bell) curve. Consequently, Figure 2 is a typical normally distributed curve.

Figure 2

Normal Distribution



Used with permission C. Carter-Snell

At least 68% of normally distributed research result are within 1 standard deviation (SD) from the mean (centre line) and 95% of all results are within 2 standard deviations (Andrade, 2019). Where the statistical results sit on the bell curve helps researchers to determine if the results are statistically significant. The cutoff (alpha or α level) for significance is determined prior to the study and is the probability of obtaining that result by chance. That means that to be significant, the results must be an outlier, or outside more than two (2) standard deviations. The alpha level is traditionally 0.05 or 5%, This means that these extreme results are only due to error five (5) times out of one hundred, or 5% of the time.

If the obtained probability (*p*-value) from the statistical tests is higher than a pre-set alpha, then the result is not significant (Table 1). If the obtained probability is less than the alpha, then results are significant and less likely to be by *error* (although *error* is still possible). Quantitative research is about hypothesis testing. When asking if there IS a difference or association, that is referred to as the "alternative hypothesis". The often unspoken "null" hypothesis is that there is NO difference or association. When the *obtained probability* is significant, the researchers *reject the null hypothesis*, This does not "prove" there IS a difference or the size of difference, only indicates that the absence of a difference is false.

Table 1

Examples of p-values

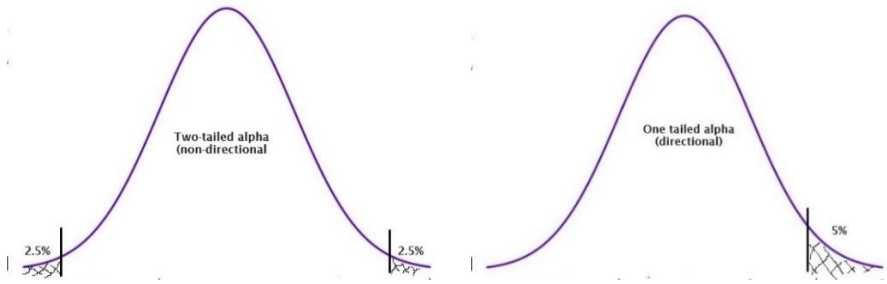
	Significant p value (< alpha)	**Not Significant p value (> alpha)**
Alpha 0.05	0.04	0.06
Alpha 0.01	0.005	0.015

If choosing an alpha of 0.05, there is a 2.5% probability on either side of the mean (the highest part of the bell curve). When results are on both sides of the normal curve, as in Figure 2 above, the name of the is *non-directional* or *two-tailed hypothesis*. Researchers use this in their statistical tests if they are not sure if results are positive or negative, meaning landing on

the left or right of the bell curve. The obtained probability is extreme when the result is beyond two standard deviations, or in other words, within either of those outer positive or negative 2.5% regions. To researchers, being in the 2.5% regions means that a larger sample is necessary to detect the significance of the results. If the data indicate the result is one direction only leaning positive or negative, the researchers often choose a one-tailed hypothesis, also called an *alpha*. All five percent is on one side of the curve as in Figure 3, the research knows there is an increased likelihood of finding a significant result. When there are one-sided results, smaller samples are used and there is a greater potential for detecting a difference. The risk for researchers is that if results go in the other direction, they often miss the significance, revealing an *error* (keep reading to find out about errors).

Figure 3

Two-tailed v one-tailed hypotheses



Used with permission C. Carter-Snell

Sometimes the authors do not state their desired alpha level of significance in their study. The reader then determines the cutoff. If not stated, traditionally an alpha (α) of 0.05 is used. If the subject is a treatment that is toxic or with harmful side effects, then a smaller level of significance should be used by the reader (e.g., 0.01 or 0.001), leaving less room for *error*.

A *p*-value only establishes the *potential* meaningfulness of the effect (e.g., significance). It informs about the strength or possibility of an association or difference. Researchers never can say that the study *proved* a difference, because there are too many sources of error. For this reason, when repeating the study multiple times, the result or obtained probability varies from study to study. Significant *p*-values that are repeated and found in varied populations, however, do demonstrate that there may be a difference or relationship that is more than *error*. Reviewers and consumers look at the design and controls used by the researchers to determine if the design and methods to trustworthy or due to error. The true value or effect of the study is not known with small single studies- it takes multiple studies or large samples until there is confidence in the finding. This is the value of a systematic review, in which similar studies are pooled to determine a closer range of the true effect.

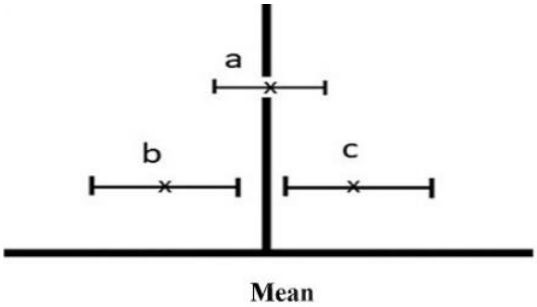
It should be noted that significant results do not indicate if the size of the difference is **clinically** significant or relevant. For instance, a drop in systolic blood pressure of 4 mmHg may be statistically significant but the drop of 4 mmHg is not clinically significant enough to make a difference. The *p*-value also does not inform readers about the strength of the association or how much variability is in the results.

Confidence Intervals and Significance

Researchers may use confidence intervals (CI) to inform readers about the variability in the results. Sometimes researchers use the CI instead of *p*-values to interpret the significance of the association. The confidence intervals (CI) give an estimate of an upper and lower range of values that researchers use to explain true population parameters, such as mean or proportion. Like the *p*-value, the accepted cutoff for the CI is either 95% or 99% depending on the risk levels (e.g., treatment that is toxic or with harmful side effects) of the intervention. The CI determines range of the sample statistic value and gives an estimate of the level of certainty or confidence in the interval, hence CI. By using CI, the researcher predicts the population parameters without having to repeat the re-draw of samples from the same population. The CI is a highly relevant tool to appraise research for practical applications, such as nursing interventions. Smaller samples have more error therefore more variability, therefore, the CI range is wider. The CI also reflects significance even in the absence of an *obtained probability* value, e.g., *p*-value. Figure 4 shows three CIs. The vertical line is the mean, which would be the existing level of a variable of interest (e.g. mean pain level, mean number of infections) or, if using odds ratios or relative risk, it would be 1.

Figure 4

Confidence intervals



Used with permission C. Carter-Snell

In example CI "a", the results for the experimental group range from below the mean to above the mean. The treatment is therefore ranging from better to worse. For instance, if the current pain medication results in a mean pain level of 6/10, then the obtained pain levels with the new treatment would be compared to this mean level of 6. The horizontal line shows the mean of the study results and variability around that result in relation to the mean. A CI range from 5 to 7 as in example "a" tells the reader that the treatment ranges from less pain to more pain than the existing treatment, therefore is not significant. If the pain CI was 4 to 5, the range of results is consistently below the mean pain level of 6 (example "b"), therefore significantly better than the current treatment. Example c shows the reader that there was an increase in mean or risk since the entire c is to the right of the mean a line. A CI of 6.5 to 7.5, the effects of the risk or treatment result in higher pain levels consistently, which is significantly worse (example "c").

Researchers sometimes calculate probabilities using odds ratio and relative risk ratio (Norton et al., 2018). The odds ratio answers the question: What are the odds of an outcome in people exposed to a risk factor? The risk ratio answers the question: How much more is the person at risk after exposure? In other words, one exposure to second hand smoke increases the

numerical odds ratio of cancer, but the risk ratio tells us what the risk of developing cancer over 20 years if exposed to second hand smoke. The statistical analysis is different, where an epidemiologist researcher measures the proportion of people with an outcome of those exposed (morbidity) v those not exposed (healthy), used to calculate their risk of developing disease; and risk ratio measures the mortality from exposure 20 years ago, e.g., 9/11 (in the USA). Both use unique mathematical formulas in a specific geographic population. Regardless, the odds ratios and risk ratios, use a mean of 1 if there is no difference. If probabilities are the same between occurring or not occurring, then when divided the odds are 1(Ranganathan et al., 2015). Similarly, if the risk is the same, it will also be 1. The CI can therefore be used to determine the significance of odds and risk ratios. If 1 is the mean in the center of the graph in figure 3, then to be significant, the CI would all be higher than 1 or all be lower than 1 (example b or c). If 1 is crossed by the CI it is not significant.

Error

When looking at significance levels, there are two main types of error. Type I (α or alpha) error is when the obtained probability or CI is significant but is falsely significant. It relates to the alpha level therefore alpha error. If the results are not significant, then there is a risk of type II (β or beta) error. This is when the results are falsely non-significant- the study was not well enough designed to detect a difference (power).

Type I (α) Error

Type I error is only a consideration if the results are significant. Are they truly significant or are they in error (falsely significant). An obtained probability of 0.03 would mean that there is still a 3% chance of error, so the study controls for variability need to be examined. How likely is it that the results are within this error? The design of the study, as well as measures to control error and threats to internal validity should be assessed when looking at significant results. Controlling for alternate explanations is considered "controlling for **internal validity**" (Singh & Thirsk, 2022). No study will have everything controlled, but we can assess generally how well they attempted to control variability or alternate explanations for the results. Sufficient internal validity must be present before we can consider generalizing the findings to other settings (**external validity**). Internal validity can be controlled by several strategies including research design, study elements, managing threats, measurement, and statistical testing.

a) Research Design

There are three main categories of quantitative research designs. These types of designs affect the potential for error in a study. These are the experimental, quasi-experimental, and non-experimental designs.

Experimental – the randomized controlled trial (RCT) is considered the only "true" experimental design. In this design, participants are randomly assigned to a control and a treatment or intervention group(s) (Jakubec & Astle, 2021). The RCT is the only design where researchers can *infer causation* (i.e. That a caused b). As such, the RCT is the *gold standard* in research designs as RCT attempts to distribute error through randomization between groups, minimizing the effect of Type I errors. Ideally the RCT will also include random *selection* of participants, to improve generalizability, but usually only random assignment is seen. That is the most important factor for controlling extraneous variables/other explanations and improving internal validity.

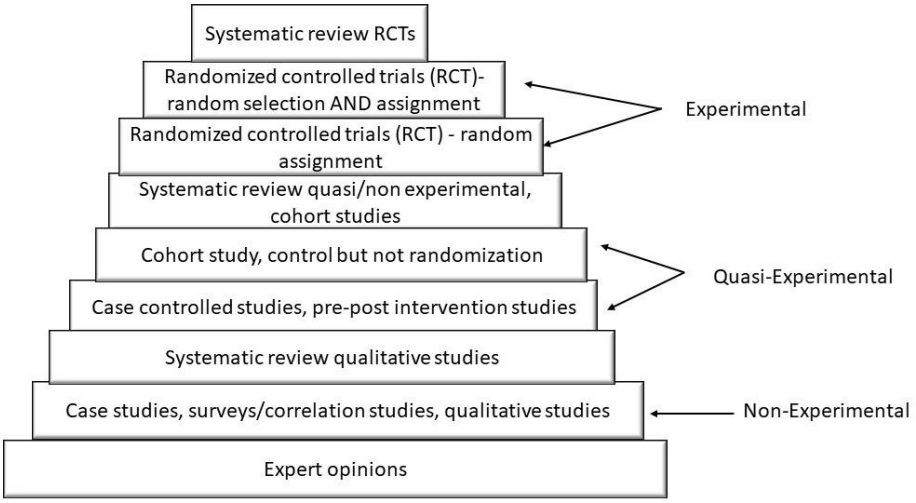
Quasi-experimental designs. Study designs include an intervention but are either missing random assignment or a control group. An example is a pre-post-test design such as where levels of education are measured after an educational intervention or treatment. Without adequate controls one could argue that learning happened outside the intervention or that the testing itself affected learning.

Non-experimental designs. These studies do not have an intervention nor control. They are only measuring what is happening. Examples include survey research, naturalistic observation, and archival research. For example, researchers may collect multiple variables from a population to predict influences on blood pressure. Although a significant association may be found, there are many other unmeasured variables that may also affect blood pressure. It was once thought rain caused malaria, as rates of malaria went up in rainy seasons. Eventually researchers found that rainy conditions allowed more breeding of mosquitoes, which were the actual carrier of malaria.

There are **levels** of research based mainly on features of the study designs described above, including randomization, control groups and the source of participants. While there is some variation, a summary of these from common nursing sources has been created in figure 5 (Melnyk & Fineout-Overholt, 2015; Woo, 2019). Single studies, even RCTs may still have error, depending on other measures of control discussed. Systematic reviews are therefore considered a higher level of research quality. In these, the researcher pools results from multiple smaller studies with similar designs and outcomes. The increased sample size reduces error and therefore a narrower confidence interval and clearer idea of "true" effect of the treatment.

Figure 5

Levels of Research Quality



Used with permission C. Carter-Snell

Systematic reviews of RCTs are the highest quality, but systematic reviews of cohort/observational studies are also a strong source of evidence. Systematic reviews of qualitative studies are beyond the scope of this article but are also considered stronger than single qualitative studies. The strength of research varies as the body of research evolves and

expands as more becomes available and as we gain more understanding of the topic. Single small studies have variable results, especially if they involve small convenience samples in unique populations. Eventually there may be enough studies to compare each of the small studies in a systematic review, getting a closer idea about the real effect of an intervention. Consider the COVID pandemic experience below considering research quality.

> Initially during the COVID pandemic, it was unknown if masks helped prevent transmission. The mode of transmission was relatively unknown. Some small studies indicated there was a difference in transmission while others did not. Eventually the researchers found the virus was airborne, and the question changed to, what types of masks were effective? Mask types initially thought to be helpful (e.g., homemade, or single layer) were determined to be ineffective, while N95 masks or masks with a triple layer were more effective. Some described this as indecision or "unreliable" science, but it demonstrated the growth of research evidence. It evolved from expert opinion to lab studies, to single observational studies with sometimes conflicting results, to RCTs. Eventually there were enough studies to conduct systematic reviews, increasing confidence in the results over time.

Randomized controlled trials are difficult in unstructured clinical settings and nursing research is therefore more likely to consist of smaller cohort quasi-experimental studies. In some settings, there often is not prior research to support the practice. Single studies are considered more trustworthy if there are similar studies with comparable results across settings or populations, which is the value of replicating studies. Smaller studies with similar outcomes also provide a basis for subsequent larger studies.

These design examples indicate why such emphasis is placed on the RCT. Given that many nursing questions cannot be randomized or controlled, correlational type designs (quasi-experimental and non-experimental) are more common. It therefore becomes very important to find ways to control for other possible explanations for the results.

b) Study Elements to Improve Internal Validity

Researchers will describe various measures in their study that they used to attempt to establish internal validity/ control error and extraneous variable. Some of these are described below.

Random assignment. The first and most effective is *random assignment* of participants into comparison groups. Ideally, there is more equal distribution of the variables brought by individual participants. Random assignment to a group also increases confidence in the results.

Case control methods. When random assignment is prohibitive, researchers use *case control methods*, matching patients from one group with patients from another group who have similar characteristics but don't receive the intervention.

Sample selection. Random selection (probability sampling) is the ideal method to best represent a population and attempt to generalize the study sample results to the larger population (external validity). In situations where it is difficult to recruit, and randomly select, *convenience samples* are used (i.e., those persons easily available and willing to participate). Convenience samples pose limitations, as the participants have more uncontrolled variables with more alternate explanations than randomly selected samples. *Cohort* samples are a whole group (e.g. an entire class, or everyone born in a certain year), but again may not get the entire group and those studied may have variables not seen with other groups that could affect results.

Selection or inclusion/exclusion criteria. The researchers will determine who is eligible/ineligible to participate in the study to reduce variability or alternate explanations. If

studying effects of dogs on stress, researchers might exclude those with allergies or prior negative dog experiences. When choosing where to select sample, it would be important to get a varied cross section. For instance, if assessing satisfaction and surveying only those outside the patient advocate/complaint department, the results will be different than if participants came from all areas of the agency.

c) Threats to Internal Validity

Key threats to internal validity are summarized in Table 2 (Flannelly et al., 2018; Singh & Thirsk, 2022). When any of these threats are present and left mostly uncontrolled, type I error may occur (the results may be falsely significant).

Table 2
Threats to internal validity

Threat	Explanation
History	An experience of an external event during or prior to study that affects the outcome (e.g., a prior dog attack affecting a study on using dogs for reducing stress)
Selection bias	Without random assignment or selection, the group characteristics may differ between groups before the study even starts. (e.g., if studying effects of dog therapy- only those who like dogs are likely to volunteer which gives different results than a random sample of students)
Maturation	Body development changes over time that may affect the outcome. This is most common in pre-post measures (e.g., fatigue levels in women measured but some become pregnant after premeasuring and not an exclusion criterion).
Instrumentation	Using different equipment to measure the outcome could produce errors if not all calibrated the same way, the same type, or if used in different ways. Using several raters could also produce different results.
Testing	Taking a test repeatedly could result in increased scores even without an intervention. A control group who also does the tests but has no intervention is preferred.
Mortality	Losing participants during a study can remove the benefits of selection and random assignment. The remaining participants may be quite different than intended between groups. An example is long term studies in which fewer people complete the 2nd measure than the 1st.
Statistical regression	Choosing participants who at extreme ends (high or low) will result in their subsequent scores moving closer to the regardless of intervention.

d) Measurement Reliability and Validity

Measurement reliability and validity are important factors. While they also affect external validity (the ability to generalize the results to other populations), using unreliable or inconsistent measures may produce error. The reliability of the measurement tool means that it consistently has similar results on the same participants. If given a test or variations of the test repeatedly the results would be the same. Measurement validity is that the tool measures what is intended. For instance, the tool might be measuring stress instead of resilience, or sedation effect rather than pain relief. Establishing measurement reliability and validity are separate studies. Strong research usually includes instruments that have already had reliability and validity established and the researchers report this in their description of the design.

Measurement reliability and validity, especially sensitivity to detect small changes, can also affect type II error as discussed later.

e) Statistical Testing

The type and numbers of statistical tests used also has a role in controlling error. If too many statistical tests are conducted on related data in a study, then the risk of false significance increases (Andrade, 2019). Multivariate analysis is one alternative, as it uses statistical techniques to reduce overlapping sources of random error or variability. For instance, instead of multiple t-tests on each of the variables, the researchers would ideally use multiple regression for associations, or analysis of variance for differences. If multiple tests are not present, however, this isn't an issue.

Type II Error (β)

The ability to detect a difference, if it exists, is the "power" of the study. Ideally studies should have a power of 0.80 or higher, meaning the results have at least an 80% chance to detect a difference if it exists. Type II error, also called beta error (β) is failure to detect a difference when it exists. Power is calculated as 1-β. If the power is 0.8, then the risk of type II error is 0.2 (20%). Note that the tolerance for type II error (e.g. 20%) is much higher than for type I error (e.g., 5%). This is because a significant result is likely to result in changes in practice if the result can be trusted. If results are not significant, then change is unlikely to happen. Power is dependent upon three key factors: the alpha level (α), the size of the sample, and the size of the effect between groups.

a) Alpha α and Power

Smaller alpha probability levels of 0.90 (*p-value* of 0.10) or 0.95 (*p-value* of 0.05) are used when there is high risk to using results with the participant populations. Consider if the alpha level is very small when planning a study, the researchers require a larger sample or a larger effect size between the groups. Table 3 shows the difference in sample size required in alpha probability group if a significance cutoff or alpha level of 0.01 rather than 0.05 is used. power 0.80 assuming a medium effect size and two- sided alpha level (Cohen, 1988).

Table 3

Alpha α level and participants needed per sample .

Alpha 0.01	Alpha 0.05
95	64

b) Effect Size

The *effect size* is the percentage of a standard deviation difference between groups or magnitude of the difference. If it is small (e.g., 0.25 of a standard deviation for many tests), or medium (0.5 SD) then larger sample sizes are required or more precise measurements. If the effect size is large (e.g., 0.75 SD) it would likely be observed without the study. If you have a small effect size, more precise measurements are recommended. The way in which data are measured can also affect the ability to detect differences in effect size. Data measurement is either continuous or categorical. **Continuous** measures are those with equal intervals from each other. Heart rate, temperature, quantitative laboratory results and size measurements are all

continuous. **Categorical** measurements are binary (e.g. yes/no) or levels such as small/medium/large, or Likert scales of 7. These are not as sensitive as continuous as they do not have equal intervals or midpoints- a score of 4 is more than 2 but is unlikely to be exactly twice as much. Larger samples are required for categorical measures as they are less precise and may miss subtle differences or small effect size. Consider the following example. A pain scale of 0 to 10 is treated as continuous. A categorical measure of pain sometimes used with triage is small (a level of 1-3), medium (4-6), or large (7 to 10). If a medication resulted in a decrease of pain from 6 to 4 on a continuous scale it might be significant. If a categorial level was used, however, the level would remain medium in both measures and no significant difference would be identified. Using reliable and valid measurements is also important to detect the measure of interest consistently.

c) Sample size

The size of the sample matters. Choosing a sample size depends on the statistical test, desired power level, and in the case of multivariate analysis, the number of variables measured. A crude rule is 20 per group, or 20 per variable in multivariate analysis but this is still low. The level of measurement is also an issue. One of the classic sources that researchers might cite for calculating sufficient sample size is "Statistical Power Analysis for the Behavioral Sciences"(Cohen, 1988), in which recommended sample size can be estimated by each type of statistical test and alpha (one or two-sided). There are also online calculators available. A well-designed study should state in the methods section what their desired sample size was based on power calculations and include a final power calculation once the study is complete. The final sample may be lower than planned due to loss of participants, or issues that occur during the study. For example, in a study of nosocomial pneumonia post-endotracheal tube suctioning, a power of 0.80 was desired and would require a sample of 169 randomly assigned to two groups of suction catheters (Snell, 1988). Changes in ICU routines during the study resulted in endotracheal tubes being removed in 24 hours for most patients, rather than the 48 hours needed to determine the pneumonia to be nosocomial. A total of 171 patients were entered into the study but the final sample was reduced to 69 total due to the change. The comparison was not significant, but the sample loss resulted in a final power of 0.51 with a medium effect (or less power if only a small effect). This meant there was a 49% chance of type II error (that a difference was missed), therefore the results could not be trusted.

Trustworthiness Decision Tree

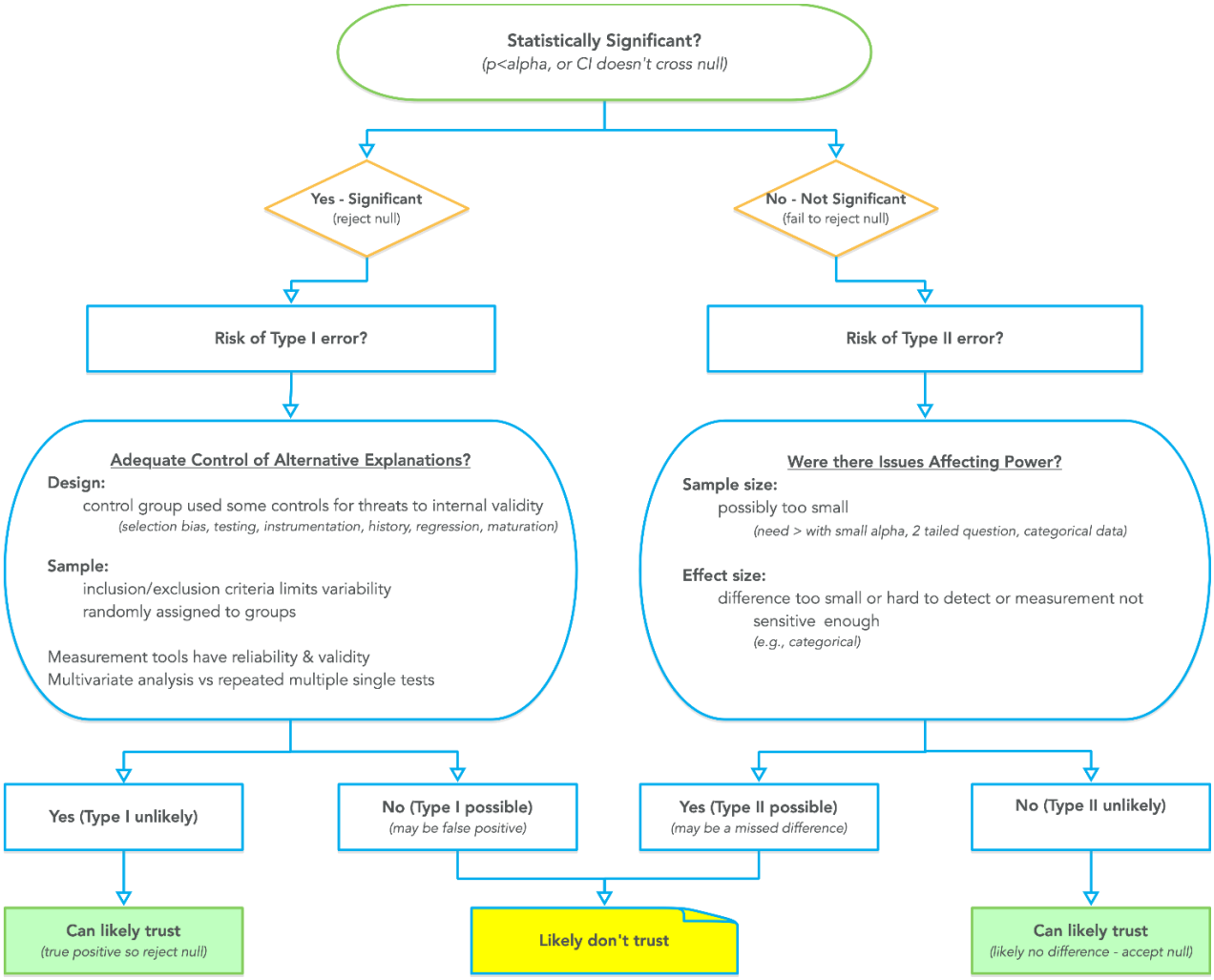
Decisions to trust the significance/non-significance of the research combine all the above information. Before getting there, however, it should be noted that you should also assess the credibility/trustworthiness of the **authors** and the **journal**. Key questions about the author include whether the authors are qualified to do this kind of research (research preparation such as master's or PhD, ORCID registration, prior research in the area), do they have knowledge of content area, and/or have relevant agency or work affiliation. When looking at the journal, is it from a credible/well known publisher, do they publish similar types of articles, are they highly rated (impact factor), and are the references relatively recent (i.e. within 5 years) and relevant. A rough estimate of a five-year range for references is typically used, except for "gold-standard" primary or key articles. If citing primary sources (e.g. Selye's stress) then they may use the original study from Selye as a gold standard primary source and should be used. Also consider whether the references include key authors or topics in this area. Also consider

the risk of publication bias especially in terms of funding. If funded by someone who can profit from results it is of concern as to bias- do they only circulate research they funded that was significant or do they also have non-significant studies? Most research grants come from government sources or community agencies so generally this is not a problem. The next part is looking at the results. Are they significant or not? If they are significant then it may be true, or it may be the result of random error through lack of controls for internal validity or measurement error. If they are not significant, then again it may be true or the result of lack of power. This is shown in Figure 4.

If authors and journals are deemed credible, then look at the trustworthiness of the results. If the results of a study are significant, then they can either be trusted or there is a chance of type I error (false significance). If the results are not significant, then the issue becomes the risk of type II error (false non-significance) or trustworthiness. These pathways can be followed in figure 6.

Figure 6

Trustworthiness Decision Tree



Used with permission: C. Carter-Snell

If results are significant, then go into the methods section of the paper to examine the design and ways in which researchers attempted to control internal validity. This is seen in the left of the pathway. If reasonable efforts were made to control variability with some or all these measures, then the results can likely be trusted. If there was a key source of variability possible, then it would be advisable to not trust the results. It would be recommended to wait for further research to be done before making any changes as the probability of type I error is likely.

If the results are not significant then we follow the right side of the decision tree and again look at methods and result. Was there a reasonable sample size in the final sample? Did they measure the changes with a validated and reliable measure that would be sensitive to changes (e.g. continuous vs categorical)? Was the alpha level reasonable or very strict? If any of these are a concern, then the results may be a type II error (missing a difference). If not, they may be trusted.

Application

Let's return to the cases in the beginning of the article. As you look at the two cases introduced in the beginning of the article, use the decision tree (Figure 4) to examine each case.

Case 1 - Significant Results

In this case the results were significantly different, in favour of the new disinfectant. They are either truly significant, or there is a chance of type I error (finding a difference when it does not exist or was only by error). In this situation, on the surface it looks like the study should be trustworthy due to the large sample size and precision of measurement (white blood counts and differentials). The first "red flag" however, is that it is sponsored by the company that makes the disinfectant. There is a concern in healthcare research that, if sponsored, the company may not release non-significant results if they occur. This is considered a form of research bias. If we move beyond that concern to the decision tree, we can examine other sources of potential error. The key to type I error is to consider if the researchers have put in sufficient controls for random error between groups. They are unlikely to have controlled everything but is what they did control sufficient? One of these is inclusion/exclusion criteria. In this study they used any patient receiving laparoscopic surgery, regardless of type, gender, or other pre-existing illnesses. They did have a control group (the other hospital), but it was not randomly assigned who got which disinfectant. One control could have been case matching (e.g., a person from each hospital of similar age, health status and similar reason for surgery) but this was not done either. A strength was the use of a precise measure such as blood results that are continuous in nature and precise, so instrumentation is not likely, and the measures should be reliable and valid. The use of a convenience sample (anyone who met criteria) could lead to selection bias or unequal distribution of other characteristics (extraneous variables) that could impact results if not controlled. For instance, there may have been a hospital acquired infection going through hospital A that could account for the higher infection rate- the type of infection was not specified. Sometimes random error can be limited through use of multivariate analysis. It does not remove the error but limits some of its' effects. In this case, it was a straight comparison of rates of infection so multivariate analysis was not used. The final decision- there were not enough controls on internal validity to trust the results. If we can't trust the internal validity, then we cannot or should not try to apply or generalize the results to other populations. Further research with more controls is required.

Case 2- Nonsignificant Results

This case involved non-significant findings for methods of contraception. You would start on the right side of the decision tree for non-significant findings. The concern then becomes whether the findings are truly non-significant (no difference) or whether there is a risk of type II error (missing a difference when it exists). While there may also be design/internal validity issues, this is not the primary concern. We need to understand if there is potential for the researchers to have missed a difference.

Conclusion

Although research findings are only one aspect of evidence informed practice, the significance of findings and trustworthiness of results requires understanding of the practitioner wanting to apply the information promoted as evidence. Small significant studies that have internal validity but have numerous errors may be correct. However, untested in the provider's communities means that there is no internal (applicable to the agency) or external validity (generalizability to agencies outside the participant sample) and *changes should not be made to practice* unless there are other similar studies in similar communities to support the results. Systematic reviews inform our practices and are considered a gold standard for research findings because they combine and analyze similar studies for common samples, sizes, and outcomes. The statistical analysis of the variability estimates a *true effect*.

In healthcare a systematic review is not always available, therefore it is important that professionals understand that word-of-mouth is insufficient to support practice. The paper presented information about how to evaluate the quality and trustworthiness of single studies and use of a decision tree to evaluate the potential impact of a single study on their practice.

References

Andrade, C. (2019). The p-value and statistical significance: Misunderstandings, explanations, challenges and alternatives. *Indian Journal Of Psychological Medicine*, *41*(3), 210-215. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19

Andrade, C. (2019b). Multiple testing and protection against a type I (false positive) error using the Bonferroni and Hochberg corrections. *Indian Journal Of Psychological Medicine*, *41*(1), 99-100. https://doi.org/10.4103/IJPSYM.IJPSYM_499_18

Ciliska, D. (2012). *Introduction to evidence-informed decision making*. Canadian Institutes of Health Research. Retrieved Nov 22 from https://cihr-irsc.gc.ca/e/documents/Introduction_to_EIDM.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Djurisic, S., Rath, A., Gaber, S., Garattini, S., Bertele, V., Ngwabyt, S. N., Hivert, V., Neugebauer, E. A. M., Laville, M., Hiesmayr, M., Demotes-Mainard, J., Kubiak, C., Jakobsen, J. C., & Gluud, C. (2017). Barriers to the conduct of randomised clinical trials within all disease areas. *Trials*, *18*(1), 360. https://doi.org/10.1186/s13063-017-2099-9

Flannelly, K. J., Flannelly, L. T., & Jankowski, K. R. B. (2018). Threats to the internal validity of experimental and quasi-experimental research in healthcare. *JOurnal of Health Care Chaplaincy*, *24*, 107-130. https://doi.org/10.1080/08854726.2017.1421019

Guyatt, G., Cairns, J., & Churchill, D. (1992). A new approach to teaching the practice of medicine. *JAMA*, *268*(17), 2420-2425.

Jakubec, S. L., & Astle, B. J. (2021). *Research literacy for health and community practice* (2 ed.). Canadian Scholar's Press.

Kumah, E. A., McSherry, R., Bettany-Saltikov, J., & van Schaik, P. (2022). Evidence-informed practice: simplifying and applying the concept for nursing students and academics. *British journal of nursing (Mark Allen Publishing)*, *31*(6), 322-330. https://doi.org/10.12968/bjon.2022.31.6.322

Kumah, E. A., McSherry, R., Bettany-Saltikov, J., van Shaik, P., Hamilton, S., Hogg, J., & Whittaker, V. (2022). Evidence-informed practice versus evidence-based practice educational interventions for improving knowledge, attitudes, understanding, and behavior toward the application of evidence into practice:A comprehensive systematic review of UG students. *Campbell Systematic Reviews,*, *15*(e1233), 1-39. https://doi.org/10.1002/cl2.1233

Melnyk, B. M., & Fineout-Overholt, E. (2015). *Evidence-based practice in nursing and healthcare: A guide to best practice*. Wolters Kluwer.

Norton, E. C., Dowd, B. E., & Maciejewski, M. L. (2018). Odds ratios- Current best practice and use. *JAMA*, *320*(1), 84-85. https://www.feinberg.northwestern.edu/sites/firstdailylife/docs/resources-docs/jama.2018.norton.guidetostatisticsandmedicine.odds-ratioscurrent-best-practice-and-use.pdf

Ranganathan, P., Aggarwal, R., & Pramesh, C. S. (2015). Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in Clinical Research*, *6*(4), 222-224. https://doi.org/10.4103/2229-3485.167092

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *Bmj*, *312*, 71-72.

Singh, M. D., & Thirsk, L. (2022). *LoBiondo-Wood and Haber's nursing research in Canada: Methods, critical appraisal and utilization* (5 ed.). Elsevier.

Snell, C. J. C. (1988). *An evaluation of two suction techniques in relation to ICU nosocomial pneumonias* University of Alberta]. Edmonton, AB.

Woo, K. (2019). *Polit & Beck Canadian Essentials of Nursing Research*. Wolters Kluwer.