Improving the Quality of EOS Clinical Research: A Step-by-Step Guide

Hiroko Matsumoto, PhD[1,2] and Brian Snyder, MD, PhD[3,4]

[1] Department of Orthopaedic Surgery, Columbia University Irving Medical Center, New York, NY; [2]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY; [3]Cerebral Palsy Center, Boston Children's Hospital, Boston, MA; [4]Department of Orthopaedic Surgery, Harvard Medical School, Boston, MA

Abstract:

Conducting high-quality research in early onset scoliosis (EOS) is challenging, requiring trained biostatisticians who develop theoretical and statistical methods to analyze data in support of evidence-based decision-making. Epidemiologists provide empirical confirmation of disease processes, identifying factors that affect prognosis to guide the process toward clinical relevancy. Within each step in the study process, there are important principles that investigators can apply to improve the quality of research in EOS.

One must ask an important research question that tests a focused, testable hypothesis. From this, create a study design with appropriate patient cohorts according to established inclusion/exclusion criteria. Specify the variables hypothesized to impact dependent measures of outcomes that reflect disease pathophysiology, treatment, and/or prevention. The data is to be analyzed with applicable statistical tests based upon power calculations with an estimate of the extent of variation in the dependent variables. Finally, we interpret results established on appropriately powered statistical tests in support/rejection of the hypothesis.

These points, as relevant to early onset scoliosis (EOS) research, can be illustrated through an example of a retrospective *de novo* study identifying risk factors for increased mortality and decreased health related quality of life (HRQoL) in EOS patients with cerebral palsy (CP) undergoing spine surgery.

Key Concepts:

- There are many unanswered questions in the management of early onset scoliosis.
- Impactful clinical research in this field and in all of pediatric orthopaedics requires a team of clinicians, epidemiologists, and biostatisticians, each contributing in their areas of expertise.
- Determining a testable hypothesis is the first step of careful study design with clearly defined independent and dependent variables.
- A variety of study designs can be considered, each with their own potential bias, confounding effects, chance, and risk for reverse causation.
- A basic understanding of p-values, accuracy, precision, and relative risk is important to consider when determining if statistical findings have clinical importance.

Introduction:
Before You Start

Early onset scoliosis (EOS) is a challenging field with many questions that remain unanswered. To create a high-quality research project to answer these, first ask a question that is worth answering: Will the answer change the way you practice? How will the knowledge gained advance the treatment of EOS? Avoid derivative, incremental, non-impactful retrospective research studies. Think about how you will answer the question. Develop a preliminary hypothesis and research plan that answers the question. "Brainstorm" your ideas with colleagues and mentors, seek critical feedback—revise and rethink your approach.

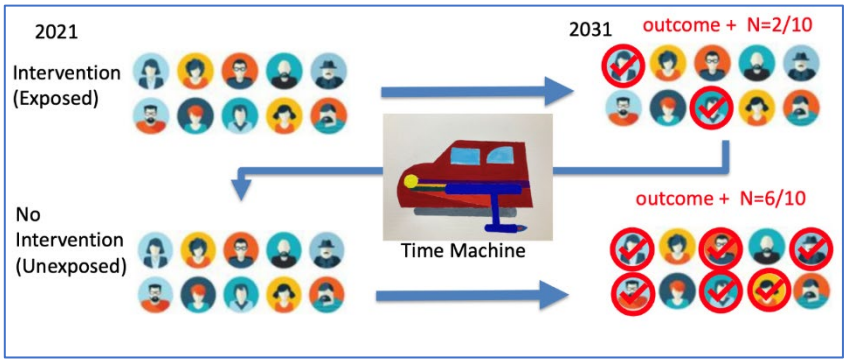


Figure 1. *Hypothetical model for establishing causation*

Causal Inference from Observed Associations

The primary mission of clinical research is to provide evidence of causality. For example, does a novel surgical technique lead to better clinical outcomes compared to traditional methods or does a specific patient characteristic increase the risk of postoperative complications? To answer these questions requires an understanding of causality; however, simply identifying an association between a patient characteristic (independent variable) and an outcome (dependent variable) *does not establish causality*. To understand the actual effect of an intervention on an outcome for a specific patient cohort, one would theoretically need to prospectively intervene on study participants today, follow them for a specific amount of time, and measure clinical outcomes, then on the same cohort, put them in a time machine, bring them back to the beginning of the study and NOT intervene (i.e., untreated or natural history), follow them for the same amount of time, and measure the same clinical outcomes (Figure 1). Since the cohort is identical, the only difference affecting the future clinical outcome (dependent variable) is whether the patient was exposed or not to the intervention (independent variable). While this hypothetical experiment establishes true causation, it is impossible to perform. In the real-world causation is inferred

True Result

	Outcome +	Outcome -	Risk	Risk Ratio
Exposed	50	50	0.5	1
Unexposed	50	50	0.5	

Biased Result

	Outcome +	Outcome -	Risk	Risk Ratio
Exposed	50	50	0.5	0.6
Unexposed	50	10	0.8	
Dropouts		40		

Figure 2*. Likelihood of being retained in the study leading to selection bias*

from observed associations (***causal inference***). Validation of causation must be proven based on pathophysiological mechanisms and additional objective clinical evidence.[1]

In casual inference, we compare the likelihood of an outcome in the exposed group to the likelihood of an outcome in the unexposed group; the groups are presumed to be interchangeable with respect to all patient-related characteristics that might affect clinical outcomes.[2,3] To ensure that causal inference is the sole plausible explanation for an observed association, alternative explanations for that association must be systematically eliminated, including selection or information bias, confounding, reverse causation, and chance. ***Selection bias*** is when

subjects meeting inclusion criteria are enrolled into a study but may not fully reflect the characteristics and outcomes of the entire targeted patient population. Consider, for example, if patients unexposed to a treatment with a good outcome are retained until the study is complete (Figure 2: N=50→N=50), while patients with less favorable outcomes drop out (Figure 2: N=50→N=10)). A less obvious example is how preoperative health status can unintentionally introduce selection bias. Patients who are healthier preoperatively are more likely to have better postoperative outcomes. Medical clearance to help providers identify children least likely to suffer serious postoperative complications and most likely to have improved surgical results inadvertently introduces selection bias by encouraging surgeons to operate on the least sick patients, even though it is the sickest patients who are most in need of an intervention. ***Information bias*** results from misclassifying treatment exposure or outcomes during the collection, recall, and processing of information (Figure 3).

Confounding occurs when an unaccounted, extraneous factor is associated with both the treatment exposure and the outcome being investigated (Figure 4). Randomized controlled trial (RCT) study design removes confounding by randomly assigning subjects to the exposed vs. unexposed treatment groups, thereby eliminating any systematic association between potential confounders and treatment exposure. While causation is implied if an exposure precedes an outcome, in ***reversed causation,*** that temporal order is reversed. Reverse causation can also explain an observed association. ***Chance*** is the likelihood that random error produced an association between an exposure and an outcome. In most clinical studies, the probability value (p-value) is used to evaluate the likelihood that an observed association occurred by random chance.

True Result

	Outcome +	Outcome -	Risk	Risk Ratio
Exposed	50	50	0.5	1
Unexposed	50	50	0.5	

Biased Result

	Outcome +	Outcome -	Risk	Risk Ratio
Exposed	10 →	90	0.1	0.2
Unexposed	50	50	0.5	

Figure 3. *Misclassification of outcome leading to information bias*

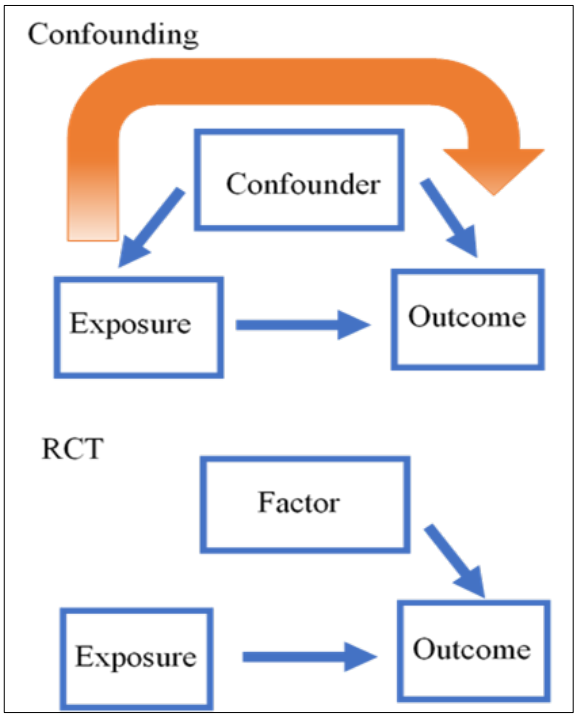


Figure 4. *Directed acyclic graph (DAG)[4] demonstrating confounding effect and removal by RCT*

1. Develop Testable Hypotheses

Avoid collecting "data" and taking a "shotgun" approach that surveys the data for associations between variables. Ask a research question that tests a hypothesis or formulate a hypothesis that answers a research question. Clinical studies should be hypothesis driven. The hypothesis forms the conceptual foundation of your investigation; it needs to be focused and testable by including independent and dependent (outcome) variables that can be evaluated objectively and

EOS Example

P = EOS patients with CP who have major scoliosis with Cobb angle >40°

I = Surgery for their spinal deformity

C: No Surgery for their spinal deformity

O: Mortality and HRQoL

T: At 10-year postoperative assessment

Hypothesis 1: For EOS CP patients with scoliosis >40° (P), those treated surgically to correct their spinal deformity (I) have decreased risk of mortality (O) at 10-year postoperative follow-up (T) compared to those patients who did not have surgical correction of their scoliosis (C).

- *Specific Aim 1:* Compare mortality (O) at 10-year follow-up (T) in EOS CP patients with scoliosis >40° (P) treated surgically (I) compared to those EOS CP patients with 40° scoliosis who did not have spine surgery (C).

Hypothesis 2: For EOS CP patients with scoliosis >40°(P), those treated surgically to correct their spinal deformity (I) have decreased risk of deterioration in HRQoL (O) at 10-year postoperative follow-up compared to those EOS CP patients who did not have surgical correction of their scoliosis (C)

- *Specific Aim 2:* Compare HRQoL (O) at 10-year follow-up (T) in EOS CP patients with scoliosis >40°(P) treated surgically (I) compared those EOS CP patients with 40° scoliosis who did not have spine surgery (C).

compared statistically. The PICOT format organizes proposed research questions into focused hypotheses and facilitates the development of specific aims and methods that test hypotheses:[5,6]

P: *Population:* specify study subjects—inclusion/exclusion criteria

I: *Intervention/Exposure:* exposure to intervention or independent risk factor to be evaluated. One treatment or risk factor should be evaluated for each hypothesis.

C: *Comparison:* alternative intervention or unexposed control group

O: *Outcome:* dependent variables—specific outcomes of interest (clinical biomarkers, imaging spine/thoracic anatomy, mortality, HRQoL)

T: *Time:* time duration for the monitored outcome to occur

For poorly understood events, exploratory research questions can be posed such as "What patient and/or surgical risk factors predict an unplanned return to the operating room (UPROR)?" Univariable analyses can be performed to investigate associations between specific risk factors and UPROR. Once risk factors are identified, the PICOT format can be used to formulate testable hypothesis and specific aims.

2. Study Design

The study design should establish causal inference as the sole plausible explanation for an association between independent (intervention/comparison) and dependent (outcome) variables, while eliminating alternative explanations for the observed association.[2,7–9] Selecting an appropriate study design should take into consideration the ethics of subject accrual and intervention assignment, the quality of measured outcomes, available resources, budget, and allocated time.

Experimental vs. Observational Studies

In experimental studies, investigators prove causal inference for an association between an intervention and a clinical outcome by prospectively assigning eligible subjects to different interventions and measuring the same outcomes (clinical biomarkers, image based anatomy, HRQoL) afterwards for the exposed and unexposed groups.[2,7–9] In an observational study, patient selection for an intervention is not actively managed—subjects meeting inclusion criteria are exposed (or not) to an intervention independently. Subsequent outcomes are measured identically for the exposed and unexposed groups, thereby allowing researchers to make inferences as to the likelihood of causation between the

intervention/exposure and ensuing outcomes.[2,7–9] Observational studies can be prospective or retrospective and can include case-control cohorts.

1) Randomized Control Study

A ***randomized clinical trial*** (RCT) is an experimental research study in which eligible subjects are prospectively allocated randomly to exposed or unexposed groups by chance.[2,7–11] This design is superior to other research designs because the random assignment of suitable patients to either the exposed or unexposed groups minimizes the risk of selection bias and mitigates risk that unknown confounding factors may bias or contaminate subsequent outcomes (Figure 4). Blinding observers evaluating outcomes as to the assignment of subjects to the exposed vs. the unexposed intervention groups reduces the possibility of measurement bias. Since the allocation of patients to the exposed vs. unexposed groups occurs before outcomes are measured, reverse causation is eliminated as an explanation for observed associations. The CONSORT (CONsolidated Standards of Reporting Trials) 2010 guidelines are helpful for understanding the methodology and reporting of RCTs.[12,13] Computer programs facilitate the random assignment of patients to exposed/unexposed cohorts (http://random.org/). In EOS, parallel group randomized trials[13] or randomized crossover trials[12] can be useful.

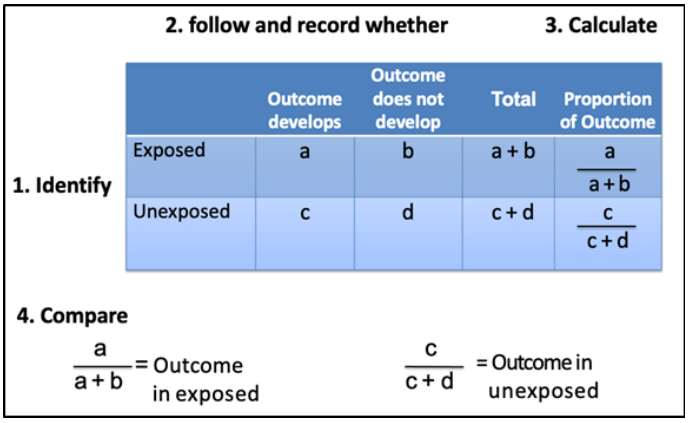


Figure 5. *Cohort study*

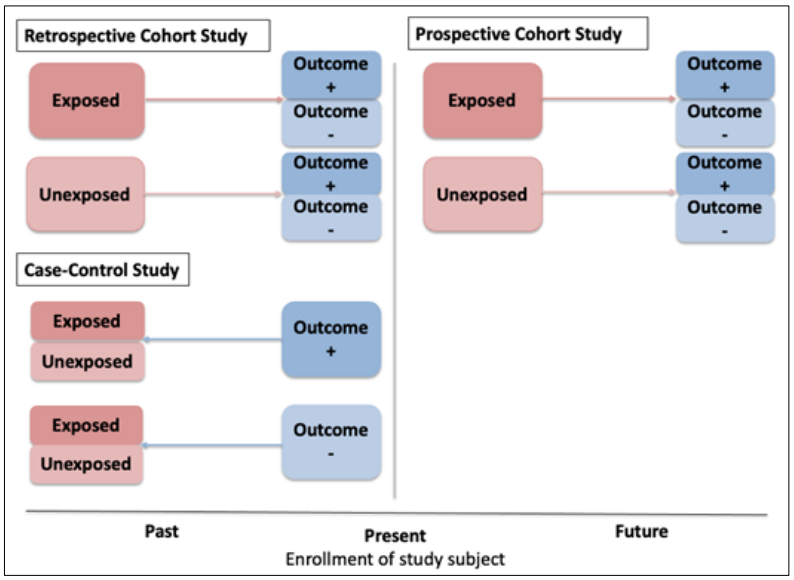


Figure 6. *Prospective and retrospective cohort study and case-control study in relation to the enrollment of study subject*

2) Cohort Studies

Cohort studies can be prospective or retrospective (Figure 5). PICOT should be defined prior to conducting the study: eligible subjects (P) are either prospectively assigned to the exposed (I) vs. unexposed (C) intervention groups or retrospectively identified as having been exposed or unexposed to an intervention. Subsequently both groups are monitored over time (T) for development of specific outcomes (O). The proportion of exposed vs. unexposed subjects who exhibit these outcomes are calculated and compared. In ***prospective cohort studies***, investigators allocate suitable patients to be exposed or unexposed to an intervention (Figure 6). Subjects are monitored for specific outcomes, which may take substantial time to develop. In ***retrospective cohort studies***, the outcome has already developed (Figure 6) at the time of patient evaluation, shortening the time duration of the study. ***Combined prospective and retrospective cohort studies*** are where investigators use existing data to retrospectively identify appropriate subjects of interest based on exposure status but subsequently followed prospectively to monitor outcomes. Compared to RCTs, cohort studies are more susceptible to selection bias and information bias. Investigators may unintentionally assign qualified patients to a particular exposure group based on factors that affect outcomes. For longitudinal studies, the main concern is differential loss of follow-up, which can contribute to selection bias.[3] Since

retrospective cohort studies are planned after the data is collected compared to prospective cohort studies, retrospective cohort studies are more vulnerable to bias induced by unmeasured confounders as well as ***information bias*** introduced by misclassification of exposure status or outcome status. Sensitivity analysis should be conducted to assess these biases.[3] Multi-center registries are especially useful for aggregating large numbers of eligible patients who were exposed or not to an intervention and consequently monitored over time for the development of outcomes that reveal the safety or efficacy of an intervention. The limitation of using large registry datasets is that although the data was collected contemporaneously, the data was not explicitly collected to prospectively test a specific hypothesis. The data set aggregates a myriad of patients from multiple sites, exhibiting a large age range, disparate diagnoses, and co-morbidities who may have been followed inconsistently over time for measured outcome variables. Selection bias due to missing values and inconsistent follow-up are especially problematic (Figure 2). Selection bias is introduced if missing outcome metrics and/or loss to follow-up was not random, influenced by the intervention and/or outcome status. For example, if patients exposed to a specific procedure develop unfavorable outcomes which dissuade them from pursuing 2-year follow-up, these poor outcome cases will be excluded from the analysis and be underrepresented in the registry. For studies utilizing registry data, conducting a ***sensitivity analysis*** that evaluates all eligible patients initially exposed to a particular intervention must be compared to those patients who completed follow-up for which outcome data are available. If the distributions are similar, selection bias is less likely to be present; however, if the distributions are different, selection bias limits confidence in study results.[14,15]

3) Case-Control Study

Case-control studies are the converse of cohort studies: eligible subjects are segregated according to a specific outcome (cases) vs. subjects without that outcome (controls). The proportion of exposed subjects comprising the cases vs. control groups establishes the association

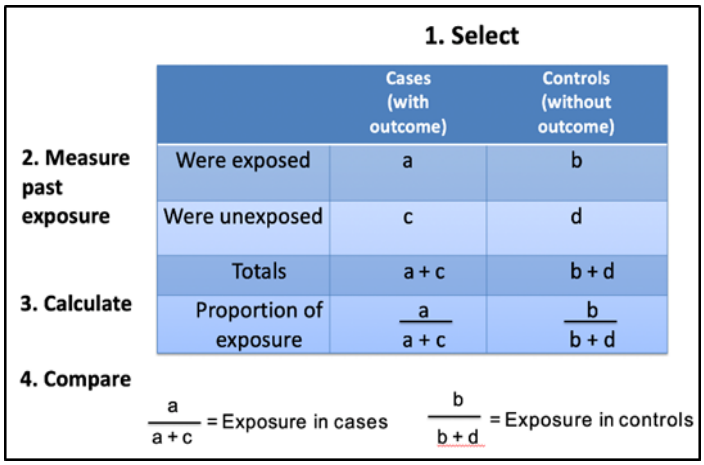


Figure 7. *Case-control study*

between the exposure and the outcome (Figure 7). Case-control studies offer logistical efficiency over cohort studies when the outcome is rare: cohort studies require a large number of exposed subjects to be evaluated so as to attain a sufficient number of cases, whereas in a case-control study design, cases are proactively identified. Case-control studies are beneficial when the outcome has a long induction and latency period. While case-control study designs offer efficiency and optimize resources, establishing causality between a rare outcome and an exposure is imprecise since only odds ratios are calculated. The risk or rate ratio of the outcome to the exposure cannot be estimated directly. Odds ratio approximate risk in cases if the proportion of exposed subjects is low (< 20%). Cases may be over-sampled if the proportion of subjects exhibiting an outcome is greater than that in the underlying population of eligible subjects. Case-control studies are susceptible to selection and information bias as well as being prone to confounding since unmeasured confounders cannot be easily adjusted for. Therefore, case-control studies should be utilized only when an outcome is rare and resources are limited. In EOS, a case-control study design can be utilized for ***exploratory analyses*** to formulate hypotheses in preparation for conducting a retrospective cohort study to obtain more accurate results.

3. Power and Sample Size Estimation

Power analysis or sample size estimation needs to be performed in order to determine how many study

patients are required to answer the research questions and test hypotheses. This is performed a priori before conducting the study. *An underpowered test result does not mean there is "no statistical difference" just because the p value is >0.05!* The number of patients needed depends on four factors (reference 55, 56):

1. The *precision and variance of the measurements* for the independent and dependent variables. The coefficient of variation (relative standard deviation) is a statistical measure of the dispersion of data points around the mean, calculated by the standard deviation divided by the mean. Confidence intervals (CI) represent uncertainty in a sample variable—i.e., the range of values for a variable, bounded above and below its mean calculated from the standard deviation multiplied by the Z value, which is determined by the selected confidence interval (typically 95%) and the number of data values for that variable. The more subjects or data points evaluated, the more precise the estimate of true population value (i.e., narrower CI).

2. The *effect size* –i.e., the magnitude of the difference that we are trying to detect. Meaningful statistical analyses should detect medically important differences (reference 55). To detect small differences (effect size) between comparison groups requires precise estimates with small variance and therefore, large numbers of subjects.

3. The *acceptance of possible error*. Conventionally, in clinical research, we choose < 0.05 for a type I error and < 0.20 for a type II error. When we try to determine whether two comparison groups are the same (accepting the null hypothesis) or if they are different (accepting the alternative hypothesis), type I error (accepting the null incorrectly) and type II (rejecting the null incorrectly) can occur. The lower the type 2 error, the higher the power of a test. To avoid invalid conclusions, by convention the minimum threshold of power is $\geq 80\%$, which means that there is an 80% chance of detecting whether the specified effect exists (and in turn 20% probability of being wrong—type 2 error).

4. The *type of statistical test* used for the analysis affects how the sample size and power are calculated. For example, a non-parametric test (e.g., Kruskal-Wallis test) will need more patients than a parametric test (e.g., independent t test). Programs such as STATA, PASS, or R are used to calculate the appropriate sample size and power estimates. These programs require specification of at least three of these four factors.

Power analysis is also used to check and validate the results and findings from retrospective studies. For example, if we specify the clinically meaningful difference, sample size, and significance level (type I error), we can calculate the power of an experiment to check whether type 2 error probability is within an acceptable range. Power curves demonstrate the inter-relationships among effect size, sample size, and the power of a statistical test at a given significance level.

4. Selecting Study Patients

Using the PICOT format to develop specific aims and testable hypotheses requires identifying the target population (P). Detailed inclusion and exclusion criteria allow specification of the patient population to be examined.[18] Inclusion criteria should consider patient demographics, gender, ethnicity, socioeconomic status, physical capabilities, diagnoses, medical co-morbidities, habits, and medication/drug exposure. To avoid potential ethical dilemmas or problems with data analysis, well-defined exclusion criteria need to be outlined. Even though patients may have missing data, data availability should not be an exclusion criterion as this can introduce selection bias. Sensitivity analysis can be performed to determine if patients with missing data are different from patients with complete datasets. It is important to recognize that subjects may be too young or cognitively impaired to provide informed consent for enrollment into a research study; thus, surrogates (parents, guardians) will need to be involved.

5. Study Measures

1) Exposure

In formulating the hypothesis, the exposure is the independent variable that is related to an outcome measure.[19] The exposure can be an intervention, shared factor of interest, unexposed control, or alternative intervention. Exposure factors can be continuous or categorical variables. For instance, if the patient's age during the surgery (continuous) was the exposure factor, the risk of an outcome decreases/increases by X% as age increases by 1 year. If there is a specific age which affects the risk of an outcome, the continuous variable, age, can be discretized at that threshold, e.g., > 6 years old or ≤ 6 years old (Figure 8).

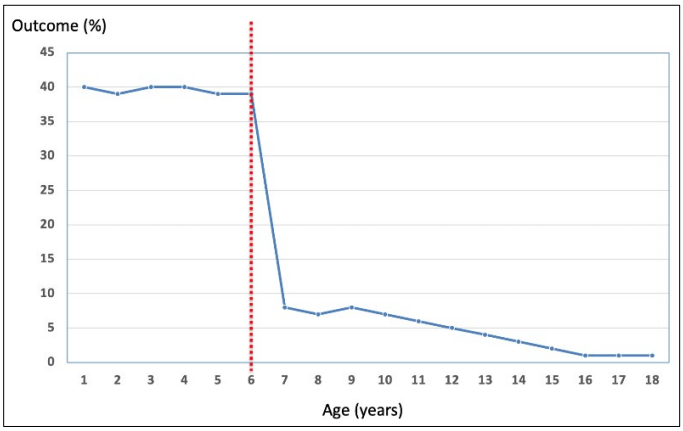


Figure 8. *Age threshold*

2) Outcome

Outcome measures are dependent variables; they can be continuous or categorical. Outcome measures determine the type of statistical analyses to be performed (Figure 9). A contingency table is used to summarize data from an experimental or observational study with two or more categorical variables. The $\chi 2$ chi-squared test of independence, Fisher exact as well as regression analyses are used to determine the association between categorical variables. The null hypothesis is that the two categorical variables are independent, while the alternative hypothesis is that the two variables are related. For binary outcomes where the categorical outcome is counted (e.g., number of complications) a Poisson regression is used. When the outcome is continuous, t-tests, Analysis of Variance (ANOVA) and regression analyses are applied.

3) Confounders

When examining associations observed between an exposure and an outcome, before assuming a causal inference, it is important to consider confounders. The apparent association between an exposure and an outcome may fail to reflect the effect of confounders and ***collider bias***, an alternative pathway relating the exposure to the outcome (Figure 10).[20,21] To minimize the effect of confounders and collider bias during the analysis, it is necessary to collect data on colliders and confounders and adjust for them using a DAG model.[22] (DAGitty is a free website to create DAGs and identify which variables to control for.)

4) Effect Modifiers

When the magnitude of the effect of an exposure on an outcome depends on a third variable, this third variable is called an effect modifier (Figure 11). The association strength, demonstrated by the risk ratio (RR), can differ for each component of the effect modifier (e.g., EOS etiology). The magnitudes and directions of the associated risk ratio can be variable.[23] It is important to differentiate effect modifiers from confounders. Adjusting for confounders in statistical models assumes that the magnitude of the effect of an exposure on an outcome is similar across groups. For instance, if EOS etiology is treated as a confounder and adjusted for, the correction will be the same for all etiologies. However, in EOS studies, the magnitude of the effect of an exposure on an outcome is different across etiologies. To differentiate confounders from effect modifiers, requires comparing effect measures, such as risk ratio or rate ratio stratified by the candidate variable. If the effect measures are substantially different, it is an effect modifier, whereas if they are similar, it is a confounder. (Breslow Day Test can be used to test the null hypothesis that the effect measures for the strata are all equal – i.e., a confounder, if $p \leq 0.05$, and the null hypothesis is rejected, the variable is a modifier).

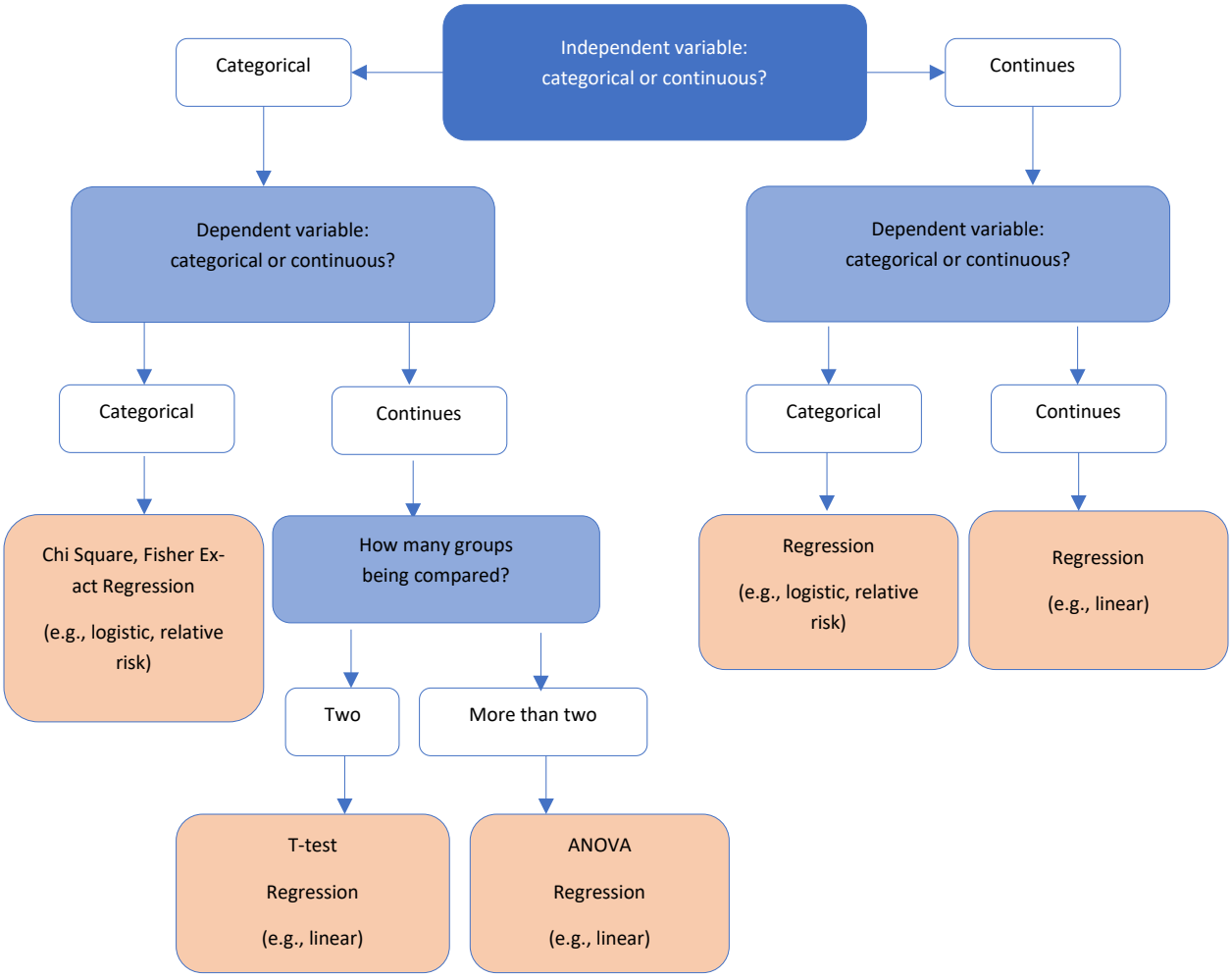


Figure 9. *Choosing statistical tests*

5) Mediators

When a mediator is assumed to be present, the total effect can be separated into direct and indirect effects (Figure 12).[24,25] A direct effect is defined as the effect of the exposure on the outcome when the mediator is absent, whereas an indirect effect is defined as the effect of the exposure on the outcome, reconciled through the mediator. There are traditional and modern approaches to estimate direct and indirect effects.[26] Mediation analysis strengthens the evidence of a causal inference between an exposure and an outcome, to better understand the mechanisms of causation to improve the efficacy of interventions.

6. StatisticL Analyses

1) Terminology

Risk: the probability of an outcome = chance of the outcome of interest developing in study patients over a specified time period

Rate: the number of study patients who develop the outcome divided by the person-time at risk in a fixed study cohort

Prevalence: the proportion of patients with an outcome during a selected time period

Incidence the proportion of new cases of a disease during a designated time period

Comparisons of these outcome measures between the exposed and unexposed groups are conducted by matrix operations such as ***ratio*** and ***difference***.

2) Description of Cohorts

General characteristics of study patients are reported in a table that exposes potential threats to the internal (e.g., confounding, bias) and external validity (e.g., generalization) of the study.[28] Rows identify patient characteristics and variables of interests (including potential confounders). Columns stratified by exposure provide insights to both internal and external validity.[28] For example, when there is a significant difference in a variable (e.g., age) between the exposed and the unexposed group, confounders may be suspected.[29]

Descriptive statistics: categorical variables are reported using n (%) and continuous variables are reported using mean (standard deviation) or median (25^{th}-75^{th} percentile or minimum-maximum).[29] Radiographic descriptions of pathoanatomy: deviation from the center to the right or left is usually reported as positive and negative values, respectively (e.g., coronal balance 5 cm to the right of center sacral line is coded as +5 while 5 cm to the left is coded as -5). Use the mean of absolute values to indicate the average magnitude of musculoskeletal deformity. Avoid using means to describe the distribution of categorical descriptions of the study cohort (e.g., GMFCS levels: I, II, III, IV, V).

When examining categorical variables, Chi-squared (when all cells have n≥5) or Fisher exact (when one or more cells have n <5) test can be used. For normally distributed continuous variables, t-tests can be used to compare means. However, when the normality assumption is violated, nonparametric tests such as the Mann Whitney U test, the Wilcoxon Signed Rank Test, or the Kruskal Wallis test must be used.[30,31] To test whether the variable is normally distributed, the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Shapiro-Wilk test can be used, where statistical significance (e.g., p<0.05) denotes that the data does not follow a normal distribution.[32,33]

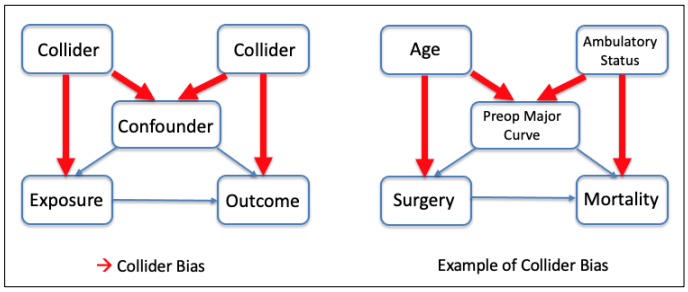


Figure 10. *Collider bias illustrated for CP EOS*

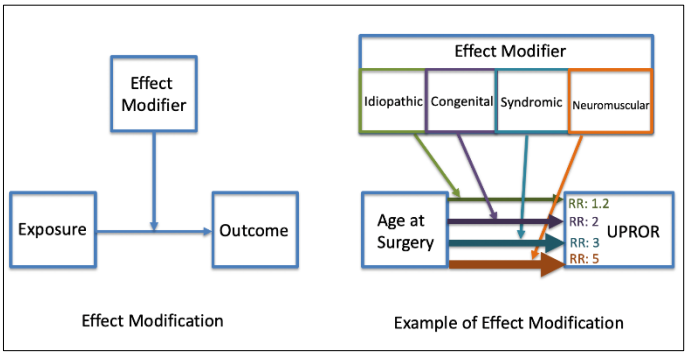


Figure 11. *Effect modification*

There has been controversy regarding the inclusion of p-values in the table to assess the potential influence of confounders as being statistically significant. As this is often misinterpreted,[34–36] instead of identifying confounders based only on p-values, investigators should consider the clinical soundness of the relationship between the exposure and the candidate confounder and whether the magnitude of the observed differences for suspected confounders is meaningful.[29,34,35,37]

3) Main Effect Analyses

The analysis investigating the association between the exposure and the outcome is called the main effect analysis. The type of statistical analyses depends on the outcome (categorical vs. continuous) and observation time (Figure 13).[38,39] Logistic regression can only be used when individual patients have equal follow-up time. Relative risk regression is preferred over logistic regression if the outcome is common (e.g., prevalence of the outcome is > 10%).[40–42] It is important to be aware that specific statistical analyses have specific assumptions that must be fulfilled for the analysis to be applicable and to

perform assumption testing and make appropriate modifications when violations are identified.[49–51] For instance, linear regression analysis requires the association between the exposure and the outcome to be linear, observations to be independent of one another and the values of the independent and dependent variables to be normally distributed, and homoscedastic (i.e., the variance of residuals is the same for any value of the exposure). Logistic and relative risk regressions require the independency assumption, while the Cox-proportional hazard requires both the independency and proportional hazard assumptions.[43–48]

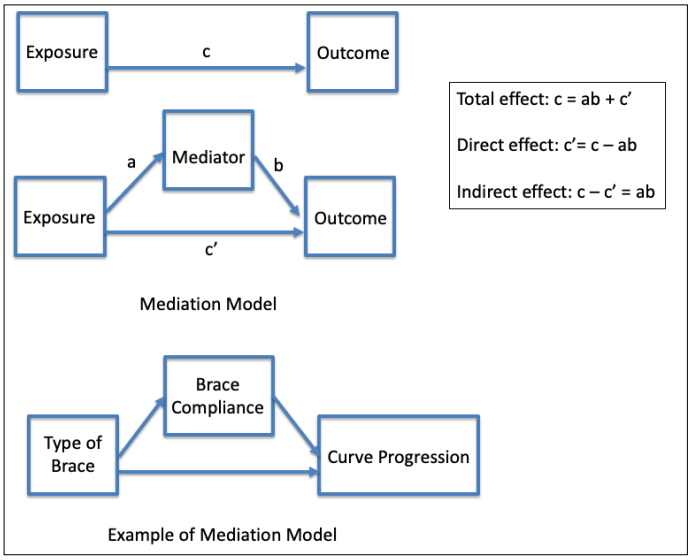
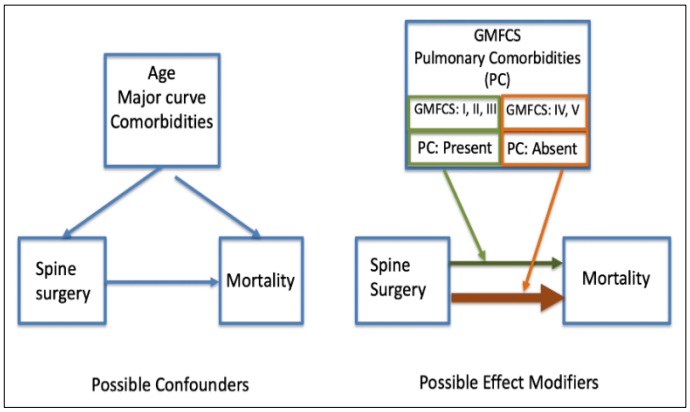


Figure 12. *Mediation*

After selecting the appropriate statistical test, the analysis is performed using an unadjusted model where the exposure is the independent variable, and the outcome is the dependent variable. Upon completion of this initial analysis, note the significance and value of the regression coefficient (beta) for the exposure variable. Subsequently, an ***adjusted model*** is developed entering confounders in addition to the exposure (independent variable) and the outcomes (dependent variables). If there is more than a 10% difference in the regression coefficients for the exposure variable in the unadjusted and adjusted models, there are confounders that must be considered as represented in the adjusted model.[38,39]



EOS CP example for possible confounders and effect modifiers

- Exposure measures: Spine surgery (coded as 1); no spine surgery (coded as 0)
- Outcome measures: Mortality within 10 years postoperatively for surgical patients and 10 years after reaching scoliosis >40° (death coded 1; alive coded 0)
- Example of possible confounders: age, major coronal curve, comorbidities (Figure 13)
- Example of possible effect modifiers: Gross Motor Function Classification System (GMFCS).[27]

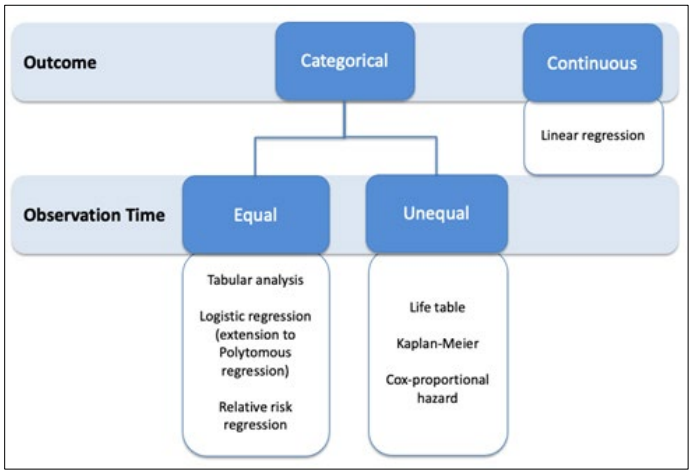


Figure 13. *Example of type of analysis*

7. Interpreting Study Results

The causal inference as the sole plausible explanation for the observed association should be described here. P-values used to assess the statistical significance of these associations are often misinterpreted. [29,35,37,52–54] P-value of ≤ 0.05 does not necessarily indicate a meaningful difference or the presence of associations, and p-values of > 0.05 does not indicate a lack of difference between independent and dependent variables.[35,37,52–54] Greenland et al. describes 25 misinterpretations of p-values and statistical significance.[53] When we interpret the main effect to

prove or disprove the study hypothesis, instead of making the judgement based on p-values, we should consider if the magnitude of differences or association observed is clinically meaningful.[29,35,37,52–54]

One of the most important responsibilities of investigators is to identify and report the limitations of a study from the clinical, methodological and epidemiological perspective. The potential of alternative explanations for the observed associations and likelihood of contamination by selection bias, information bias, confounding effects, chance, and reverse causation need to be carefully investigated and discussed.

Conclusion

Performing high-quality research requires biostatistical and epidemiologic understanding of not only strict research methodology but also the question you are searching for an answer to. In an area such as EOS, working with colleagues with research expertise can help ensure not only that appropriate research principles are followed but that the results of the study will lead to practice-changing answers for our patients.

EOS CP Example:

- Mortality Outcome: Relative risk regression
  - Mortality is a binary outcome
  - The follow-up is at 10-year assessment (equal follow-up time)
  - The outcome is not rare (mortality is > 10% at 10 years)
  - Each patient has one observation (independency assumption)
- HRQoL outcome: Linear regression
  - HRQoL score is a continuous outcome
  - Linearity, independency, normality, and homoscedasticity are assumed
  - If not, HRQoL score will be categorized, and logistic or relative risk regression will be performed

References

1. Holland PW. Statistics and Causal Inference. *J Am Stat Assoc*. 1986;81(396):945-960. doi:10.2307/2289064

2. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.

3. Rothman KJ. Causes. *Am J Epidemiol*. 1976;104(6):587-592. http://www.ncbi.nlm.nih.gov/pubmed/998606. Accessed July 19, 2017.

4. S Greenland, J Pearl, J M Robins. Causal Diagrams for Epidemiologic Research. *Epidemiology*. 1999;10(1):37-48. doi:10.1055/S-0031-1291192

5. D Sackett, WS Richardson, W Rosenburg, RB Haynes. *Evidence-Based Medicine: How to Practice and Teach Evidence Based Medicine*. 2nd ed. Churchill Livingstone; 1997. doi:10.1177/088506660101600307

6. Hulley S, Cummings S, Browner W, Grady D, Newman T. *Designing Clinical Research*. 3rd ed. Lippincott Williams and Wilkins; 2007.

7. Kenneth J Rothman. *Epidemiology: An Introduction* . 2nd ed. New York : Oxford University Press; 2012.

8. Keyes KM, Galea S. *Epidemiology Matters: A New Introduction to Methodological Foundations*. Oxford University Press; 2014. doi:10.1093/MED/9780199331246.001.0001

9. MA H, S H-D, JM R. *A Structural Approach to Selection Bias*. Vol 15. Epidemiology; 2004. doi:10.1097/01.EDE.0000135174.63482.43

10. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomized trials. *BMJ*. 2011;343(7829). doi:10.1136/BMJ.D5928

11. MA M, JP H, JA S, MA H. Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists. *Epidemiol* . 2017;28(1):54-59. doi:10.1097/EDE.0000000000000564

12. K D, T L, DG A, D E. CONSORT 2010 statement: extension to randomised crossover trials. *BMJ*. 2019;366. doi:10.1136/BMJ.L4378

13. KF S, DG A, D M. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med*. 2010;8. doi:10.1186/1741-7015-8-18

14. H M, N K, T S, et al. Short fusion with vertebrectomy during growth in congenital spinal deformity: is early surgical intervention recommended? *Spine Deform*. 2020;8(4):733-742. doi:10.1007/S43390-020-00082-9

15. H M, DL S, BA A, et al. Comparing health-related quality of life and burden of care between early-onset scoliosis patients treated with magnetically controlled growing rods and traditional growing rods: a multicenter study. *Spine Deform*. 2021;9(1):239-245. doi:10.1007/S43390-020-00173-7

16. Matsumoto H, Fano AN, Herman ET, et al. Mortality in Neuromuscular Early Onset Scoliosis Following Spinal Deformity Surgery. In: *28th IMAST International Meeting on Advanced Spine Techniques*. Virtual; 2021:36.

17. RL D, JA V, PE M, BD S, BJ S. Health-related quality of life and caregiver burden after hip reconstruction and spinal fusion in children with spastic cerebral palsy. *Dev Med Child Neurol*. 2021. doi:10.1111/DMCN.14994

18. EA P, CB T. Sampling methods: selecting your subjects. *Air Med J*. 2007;26(2):75-78. doi:10.1016/J.AMJ.2007.01.001

19. LH I, K B-R, MA N. Experimental manipulation of psychosocial exposure and questionnaire sensitivity in a simulated manufacturing setting. *Int Arch Occup Environ Health*. 2009;82(6):735-746. doi:10.1007/S00420-008-0364-7

20. Kenneth J Rothman, Sander Greenland, Timothy L Lash. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

21. MA L-F, M S, D R-S, M JSP, A V, ME S. Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application. *Int J Epidemiol*. 2019;48(2):640-653. doi:10.1093/IJE/DYY275

22. DAGitty - drawing and analyzing causal diagrams (DAGs). http://www.dagitty.net/. Accessed September 17, 2021.

23. Y E, J H, C J, et al. Can distraction-based surgeries achieve minimum 18 cm thoracic height for patients with early onset scoliosis? *Spine Deform*. 2021;9(2):603-608. doi:10.1007/S43390-020-00230-1

24. RM B, DA K. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173-1182. doi:10.1037//0022-3514.51.6.1173

25. VanderWeele TJ. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press; 2015.

26. Causal Mediation | Columbia Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/causal-mediation. Accessed September 17, 2021.

27. RJ P, L A, JW G, B G, SW M. Stability of the Gross Motor Function Classification System, Manual Ability Classification System, and Communication Function Classification System. *Dev Med Child Neurol*. 2018;60(10):1026-1032. doi:10.1111/DMCN.13903

28. Fisch RZ, Alexandrowitz A. Drug points: Delirium in a patient treated with mainserin. *Br Med J (Clin Res Ed)*. 1988;296(6615):137. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2544753/. Accessed September 17, 2021.

29. D M, S H, KF S, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Int J Surg*. 2012;10(1):28-55. doi:10.1016/J.IJSU.2011.10.001

30. Conover WJ. *Practical Nonparametric Statistics*. Vol 350. John Wiley & Sons; 1998.

31. S H, JR D, JR C. Parametric and Nonparametric Tests in Spine Research: Why Do They Matter? *Glob spine J*. 2018;8(6):652-654. doi:10.1177/2192568218782679

32. D'Agostino RB. *Goodness-of-Fit-Techniques*. Vol 68. CRC press; 1986.

33. AR H. Testing experimental data for univariate normality. *Clin Chim Acta*. 2006;366(1-2):112-129. doi:10.1016/J.CCA.2005.11.007

34. MJ K, RH G, DE G. P-values in baseline tables of randomised controlled trials are inappropriate but still common in high impact journals. *Eur J Prev Cardiol*. 2012;19(2):231-232. doi:10.1177/1741826711421688

35. M B. Statisticians issue warning over misuse of P values. *Nature*. 2016;531(7593):151. doi:10.1038/NATURE.2016.19503

36. MR de B, WE W, LD K, IH S, JW T. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *Int J Behav Nutr Phys Act*. 2015;12(1). doi:10.1186/S12966-015-0162-Z

37. YY P. Some common misperceptions about P values. *Stroke*. 2014;45(12):e244-e246. doi:10.1161/STROKEAHA.114.006138

38. Vittinghoff E. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Spriger; 2005.

39. JJM R, JWR T, I E, MW H. Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. *BMC Med Res Methodol*. 2019;19(1). doi:10.1186/S12874-018-0654-Z

40. J Z, KF Y. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*. 1998;280(19):1690-1691. doi:10.1001/JAMA.280.19.1690

41. LA M, C W, X X, JP H. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157(10):940-943. doi:10.1093/AJE/KWG074

42. GS L, LJ U, D J, C R, C W, A R. At Odds: Concerns Raised by Using Odds Ratios for Continuous or Common Dichotomous Outcomes in Research on Physical Activity and Obesity. *Open Epidemiol J*. 2012;5(1):13-17. doi:10.2174/1874297101205010013

43. David GK, Mitchel K. *Survival Analysis: A Self-Learning Text*. Spinger; 2012.

44. TG C, MJ B, SB L, DG A. Survival analysis part IV: further concepts and methods in survival analysis. *Br J Cancer*. 2003;89(5):781-786. doi:10.1038/SJ.BJC.6601117

45. PJ M. Current treatment of psoriatic arthritis. *Rheum Dis Clin North Am*. 2003;29(3):495-511. doi:10.1016/S0889-857X(03)00047-4

46. IM F, MS B, TJ S. Treatment of a resorbed maxilla with sinus grafting, implants, and spark erosion overdenture: clinical report. *Implant Dent*. 1992;1(2):150-153. doi:10.1097/00008505-199205000-00008

47. TG C, MJ B, SB L, DG A. Survival analysis part I: basic concepts and first analyses. *Br J Cancer*. 2003;89(2):232-238. doi:10.1038/SJ.BJC.6601118

48. P H. Fundamentals of survival data. *Biometrics*. 1999;55(1):13-22. doi:10.1111/J.0006-341X.1999.00013.X

49. DSS - Introduction to Regression. https://dss.princeton.edu/online_help/analysis/regression_intro.htm. Accessed September 17, 2021.

50. What are the four assumptions of linear regression? – Gaurav Bansal. https://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-are-the-four-assumptions-of-linear-regression/. Accessed September 17, 2021.

51. Testing the assumptions of linear regression. https://people.duke.edu/~rnau/testing.htm. Accessed September 17, 2021.

52. S G. A dirty dozen: twelve p-value misconceptions. *Semin Hematol*. 2008;45(3):135-140. doi:10.1053/J.SEMINHEMATOL.2008.04.003

53. S G, SJ S, KJ R, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350. doi:10.1007/S10654-016-0149-3

54. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. *https://doi.org/101080/0003130520161154108*. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108

55. Arkin CF, Wachtel MS. How Many Patients Are Necessary to Assess Test Performance? *JAMA* [electronic article]. 1990;263(2):275–278.

56. SR J, S C, M H. An introduction to power and sample size estimation. *Emerg. Med. J.* [electronic article]. 2003;20(5):453–458.