

# How to Critically Review the Medical Literature

Molly A. Brewer, DVM, MD, MS  
Professor and Chair

Christopher M. Morosky, MD, MS  
Associate Professor

Department of Obstetrics and Gynecology  
University of Connecticut

April 20, 2022

# DISCLOSURES

- We have no disclosures



*Data, data everywhere, but not a thought to think.*

- Jesse Shera

*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.*

- H.G. Wells, 1866-1946



# Learning Objectives

- How to read the literature and decide if you will adopt a practice
- Review important aspects of trial design
- Review study designs and strengths and weaknesses of both
- Review statistical methods



# As an Example – The ARRIVE Trial

## *The* NEW ENGLAND JOURNAL *of* MEDICINE

ESTABLISHED IN 1812

AUGUST 9, 2018


VOL. 379 NO. 6

### Labor Induction versus Expectant Management in Low-Risk Nulliparous Women

William A. Grobman, M.D., Madeline M. Rice, Ph.D., Uma M. Reddy, M.D., M.P.H., Alan T.N. Tita, M.D., Ph.D., Robert M. Silver, M.D., Gail Mallett, R.N., M.S., C.C.R.C., Kim Hill, R.N., B.S.N., Elizabeth A. Thom, Ph.D., Yasser Y. El-Sayed, M.D., Annette Perez-Delboy, M.D., Dwight J. Rouse, M.D., George R. Saade, M.D., Kim A. Boggess, M.D., Suneet P. Chauhan, M.D., Jay D. Iams, M.D., Edward K. Chien, M.D., Brian M. Casey, M.D., Ronald S. Gibbs, M.D., Sindhu K. Srinivas, M.D., M.S.C.E., Geeta K. Swamy, M.D., Hyagriv N. Simhan, M.D., and George A. Macones, M.D., M.S.C.E., for the Eunice Kennedy Shriver National Institute of Child Health and Human Development Maternal–Fetal Medicine Units Network\*

# How to read the literature and decide if you will adopt a practice

- Read the abstract and decide if you are interested
- Does the introduction state a hypothesis?
- Read the Materials and Methods: is the study design appropriate for the question asked?
  - Is there a control group that is comparable to the study group?
  - Is the statistical approach reasonable?
  - What biases and confounders are inherent in the study design, and do they invalidate the findings?
- Does the data support the conclusions reached?
- Do the authors state conclusions that were not tested?

A large, stylized oak leaf graphic in a dark blue color, positioned on the left side of the slide. It has a prominent vein structure and a lobed edge.

**ABSTRACT**  
(aka, should I bother to  
read this?)


# Abstract

- The purpose of the abstract is to provide a concise overview of the study
- A good abstract will highlight the primary results and make a brief statement about the significance of the findings
- For original research, most abstracts will contain Objective, Materials and Methods, Results, and Conclusion sections



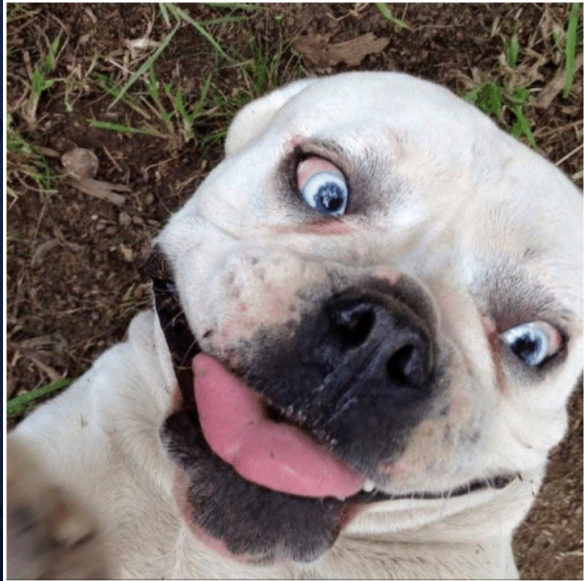
# Questions to ask about the abstract

- Does the abstract adequately summarize the article's content?
- Are there major discrepancies between the abstract and the body of the article?
- Pitkin et al. found that discrepancies occurred in 18–68% of the articles that they reviewed
- Does the abstract's conclusion address the specific aim of the investigation?

A large, stylized oak leaf graphic in a darker shade of blue, positioned on the left side of the slide, partially overlapping the text.

**INTRODUCTION**  
(aka, am I really  
interested in this paper?)

# Introduction



Total Excitement

- Is this new or confirmatory?
- If the authors' question is not clear, it raises concerns about the validity of the research
- Is there a rationale? Do we care? (For real, like in my bones)
- Do the authors build a logical case and context for their hypothesis?
- Clearly state the

HYPOTHESIS


**UConn**  
HEALTH

# ARRIVE Trial Introduction

However, these conclusions were derived largely from observational studies in which labor induction was compared with spontaneous labor.<sup>4-6</sup> Such a comparison provides little insight into clinical management, because spontaneous labor is not a certain alternative to labor induction. Most observational studies that have used the clinically relevant comparator of expectant management have not shown a higher risk of adverse outcomes with labor induction; instead, some of these studies have shown that induction of labor resulted in a lower frequency of cesarean delivery and more favorable perinatal outcomes than expectant management.<sup>7-11</sup>

[Previous] conclusions were derived largely from observational studies in which labor induction was compared with spontaneous labor. Such a comparison provides little insight into clinical management, because *spontaneous labor is not a certain alternative to labor induction.*



A large, stylized oak leaf graphic in a dark blue color, positioned on the left side of the slide, partially overlapping the text.

**METHODS**  
(aka, the most important  
part of the paper!!!)





# METHODS

## Study Design



# ARRIVE Trial – Study Design

## METHODS

### TRIAL OVERSIGHT

We conducted this multicenter, randomized, controlled, parallel-group, unmasked trial at 41 hospitals participating in the Maternal–Fetal Medicine Units Network of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The protocol (available with the full text of this article at [NEJM.org](http://NEJM.org)) was approved by the institutional review board at each hospital before participant enrollment. Written informed consent was obtained from all participants before randomization. An independent data and safety monitoring committee monitored the trial. The authors vouch for the accuracy and completeness of the data and for the fidelity of the trial to the protocol.

We conducted this multicenter, randomized, controlled, parallel-group, unmasked trial at 41 hospitals participating in the Maternal–Fetal Medicine Units Network of the Eunice Kennedy Shriver National Institute of Child Health and Human Development.

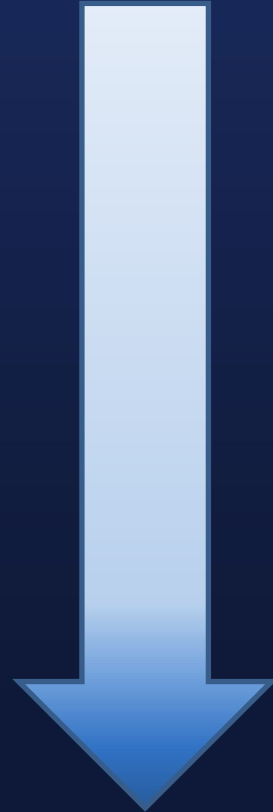
**WAIT!**

This study isn't blinded!!!

# Study Design

- Case report
- Case series
- Cross sectional study
- Case-Control
- Cohort Study
- Randomized Control

■ Trial



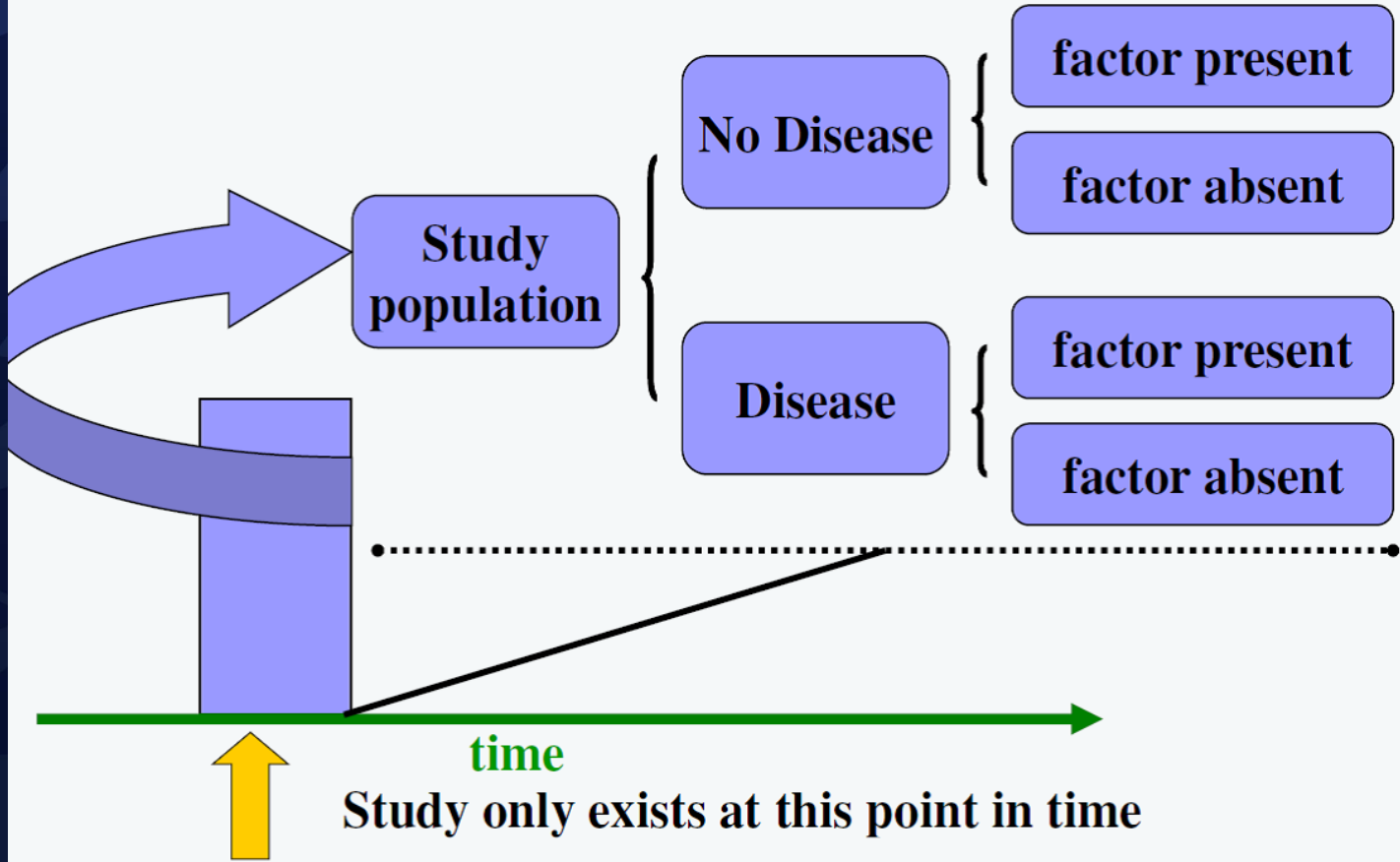
LOW

Strength of Association  
(cause and effect)

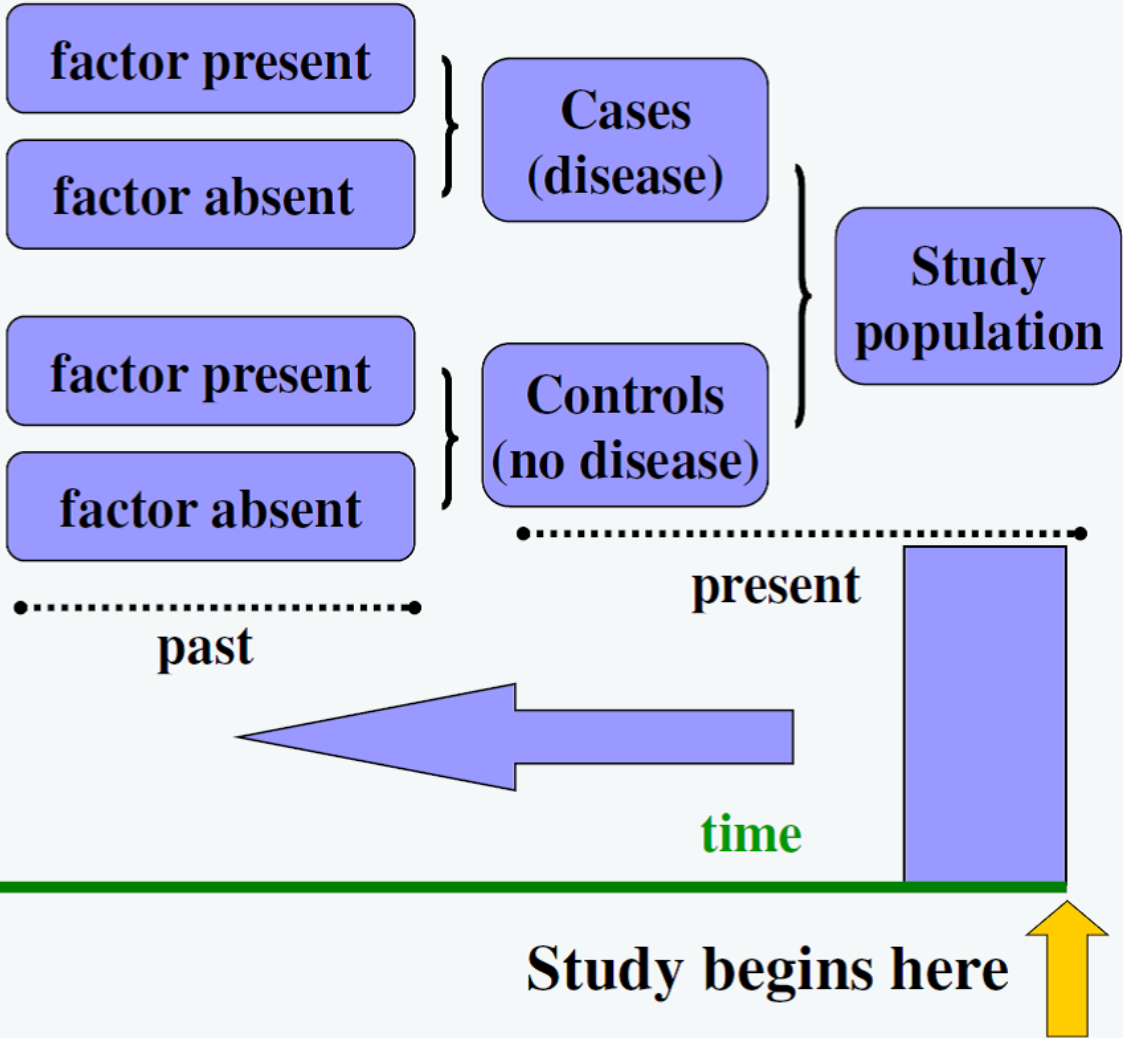
HIGH



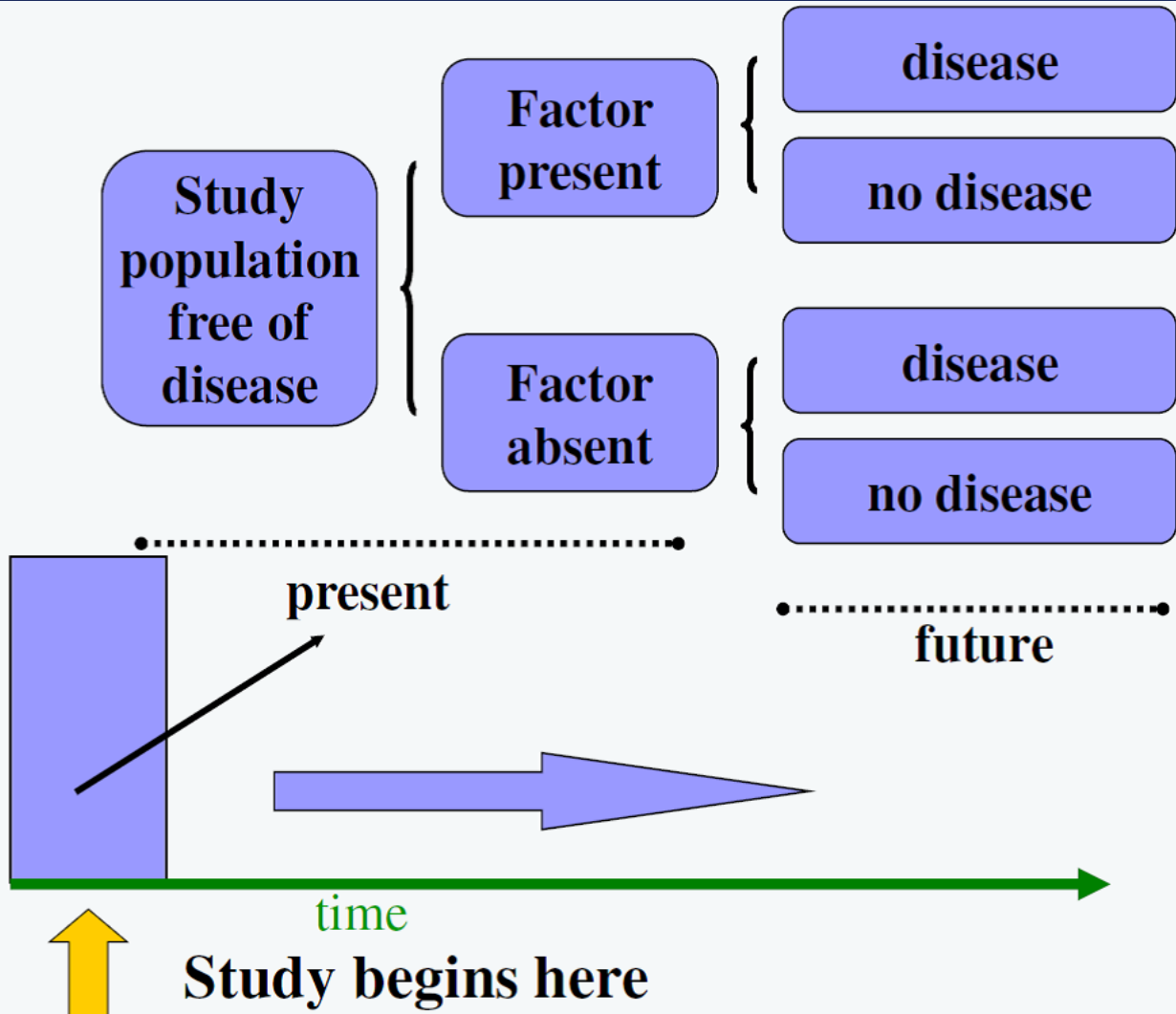
# Cross-sectional Design



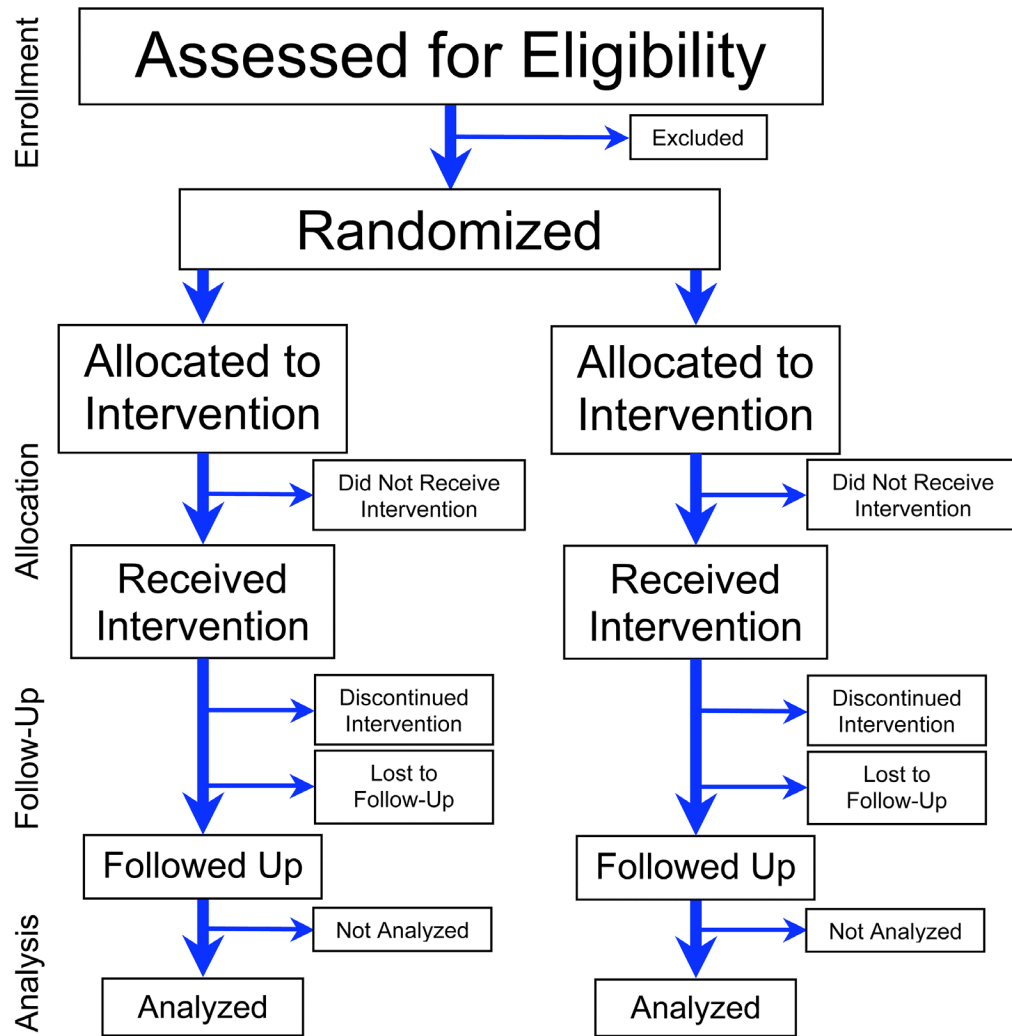
# Case-Control Design



# Cohort Design



# Randomized Control Trial



# Bias

- **Confounding Bias**
- **Selection Bias**
- **Measurement Bias**
- **Screening Bias**
- **Reader Bias**



# IRB/Funding/COI



Human subjects protection  
(plus it's required)

Hands off the data if you are  
supplying the cash => BIG  
TIME BIAS!



# ARRIVE Trial - Randomization

Women in the induction group were assigned to undergo induction of labor at 39 weeks 0 days to 39 weeks 4 days. Women in the expectant-management group were asked to forego elective delivery before 40 weeks 5 days and to have delivery initiated no later than 42 weeks 2 days. A specific induction protocol was not mandated for women who underwent induction in either group. Other protocol guidelines are provided in the Supplementary Appendix.

Women in the induction group were assigned to undergo induction of labor at 39 weeks 0 days to 39 weeks 4 days. Women in the expectant-management group were asked to forego elective delivery before 40 weeks 5 days and to have delivery initiated no later than 42 weeks 2 days.

# Primary Outcome

- This is the whole paper 😊
- Do I care (IVF pregnancy rate, versus)? Do I really care (live term birth)?
- Is the primary outcome a secondary outcome for something that I really care about, and you are being lazy or don't have the data? (EBL versus transfusion, OR time versus hospital days, etc.)
- Sample size and power is based on the primary outcome! Secondary outcomes are almost never powered!!!



# Secondary Outcomes

- Often underpowered (**beta error**) – often what you really care about (rare outcomes)
- There can be a few, or there can be a many
- If there are many, need to also worry about **alpha error**
- Here is where subgroup analysis and sensitivity analysis gets introduced – sometimes this is good and thoughtful, other times this is just fishing for a p-value

# ARRIVE Trial – Primary Outcome

## TRIAL OUTCOMES

The primary outcome was a composite of perinatal death or severe neonatal complications and consisted of one or more of the following during the antepartum or intrapartum period or during the delivery hospitalization: perinatal death, the need for respiratory support within 72 hours after birth, Apgar score of 3 or less at 5 minutes, hypoxic–ischemic encephalopathy,<sup>17</sup> seizure, infec-

The primary outcome was a composite of perinatal death or severe neonatal complications and consisted of one or more of the following during the antepartum or intrapartum period or during the delivery hospitalization...

# ARRIVE Trial – Secondary Outcome

tion (confirmed sepsis or pneumonia), meconium aspiration syndrome, birth trauma (bone fracture, neurologic injury, or retinal hemorrhage), intracranial or subgaleal hemorrhage, or hypotension requiring vasopressor support. The principal prespecified maternal outcome (the main secondary outcome) was cesarean delivery.

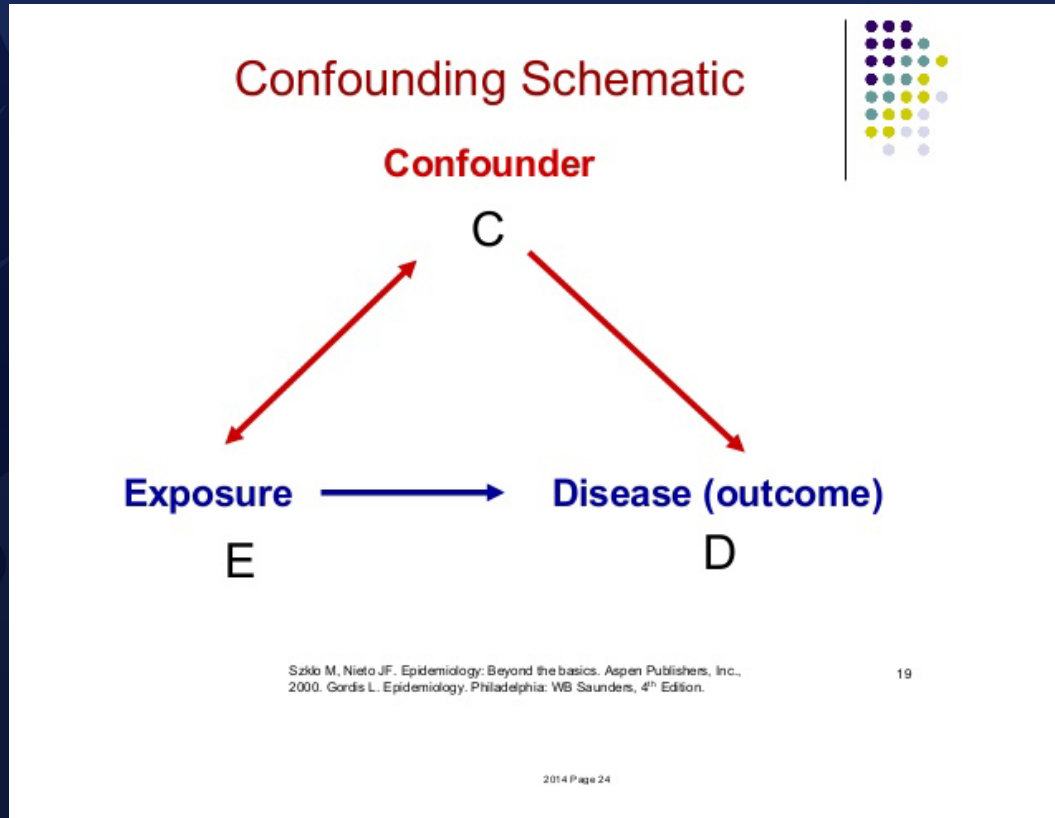
The principal prespecified maternal outcome (the main secondary outcome) was cesarean delivery.

# Variables

- Dependent variable = outcome variable
  - i.e., Response to treatment
- Independent variable = variables that have an impact on the dependent variable
  - i.e., Risk of cervical dysplasia
    - HPV status high risk vs. low risk
    - # sexual partners
  - Interaction terms
    - There is interaction between HPV status and # of sexual partners



# Confounders – Watch Out!



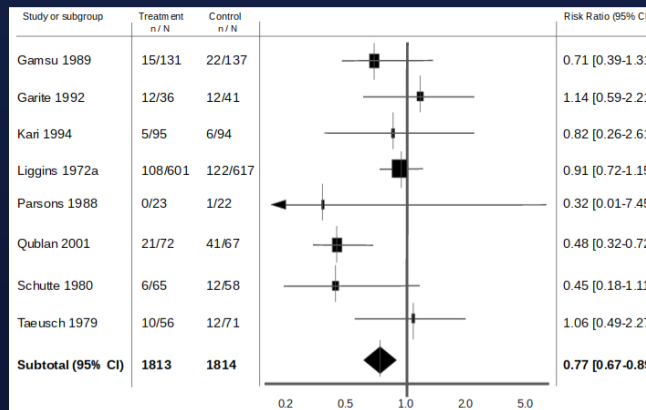
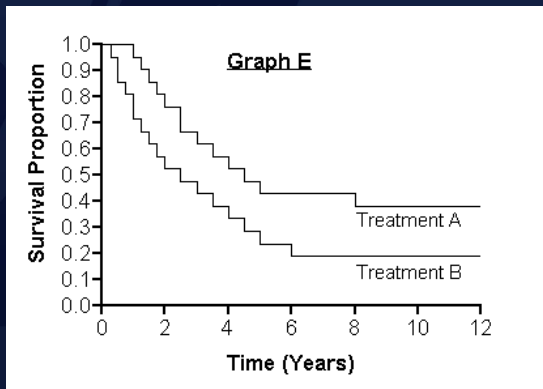
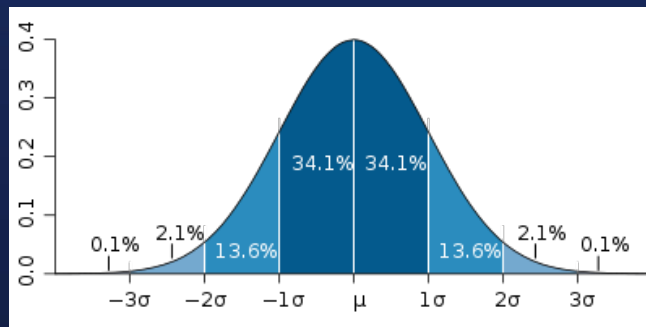
# Is the method that the authors used a reasonable approach to answer the question?

- A common flaw in experimental design is that the research methodology fails to test the hypothesis
- The internal validity of a study refers to the study's quality and is based on the adequacy of the research methodology
- A well-designed study attempts to minimize bias and confounding factors
- Did the authors conduct an intention-to-treat analysis?



# Statistics

		The "Truth"	
		Yes	No
Test Result	Yes	(A) True +	(B) False +
	No	(C) False -	(D) True -



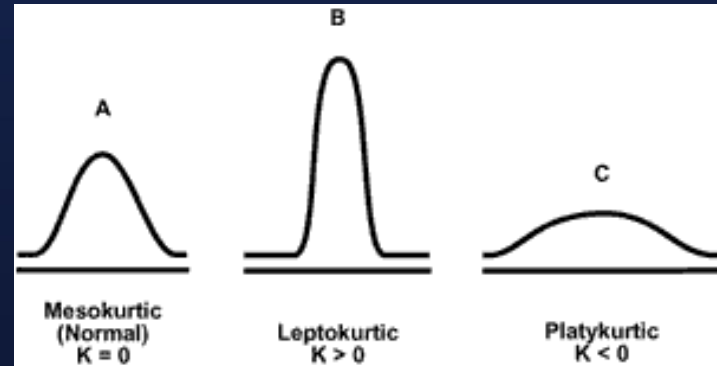
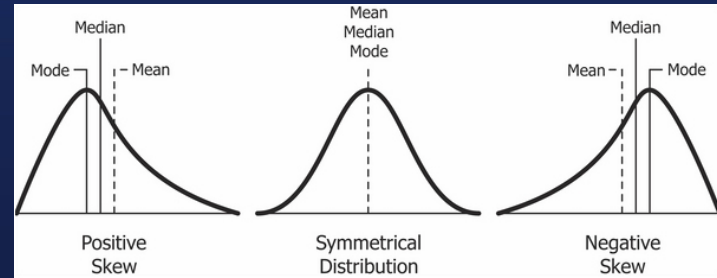
# Types of Data

- **Categorical or Qualitative Variables**
  - Nominal: race, gender, ACOG district
  - Ordinal: small, medium, large, extra large
- **Numerical or Quantitative Variables**
  - Discrete: 1 – 10 pain scale
  - Continuous: temperature, EBL



# Descriptive Statistics

- Distribution
  - Mean: average
  - Median: middle of a range
  - Mode: most common
- Dispersion
  - Range / Quartiles
  - Standard deviation / Variance
- Parametric / Non-Parametric
  - Normal
  - Skewness – asymmetry middle
  - Kurtosis – asymmetry tails



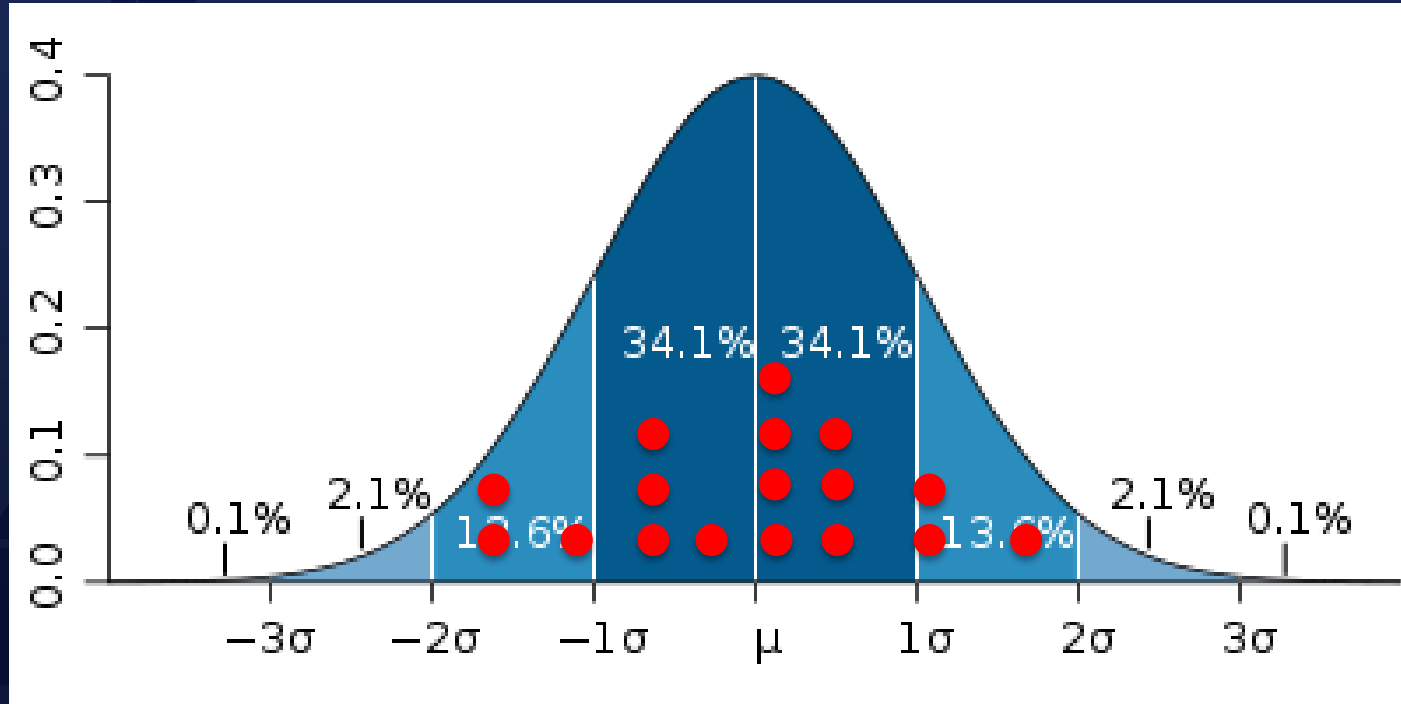
# Sample Size

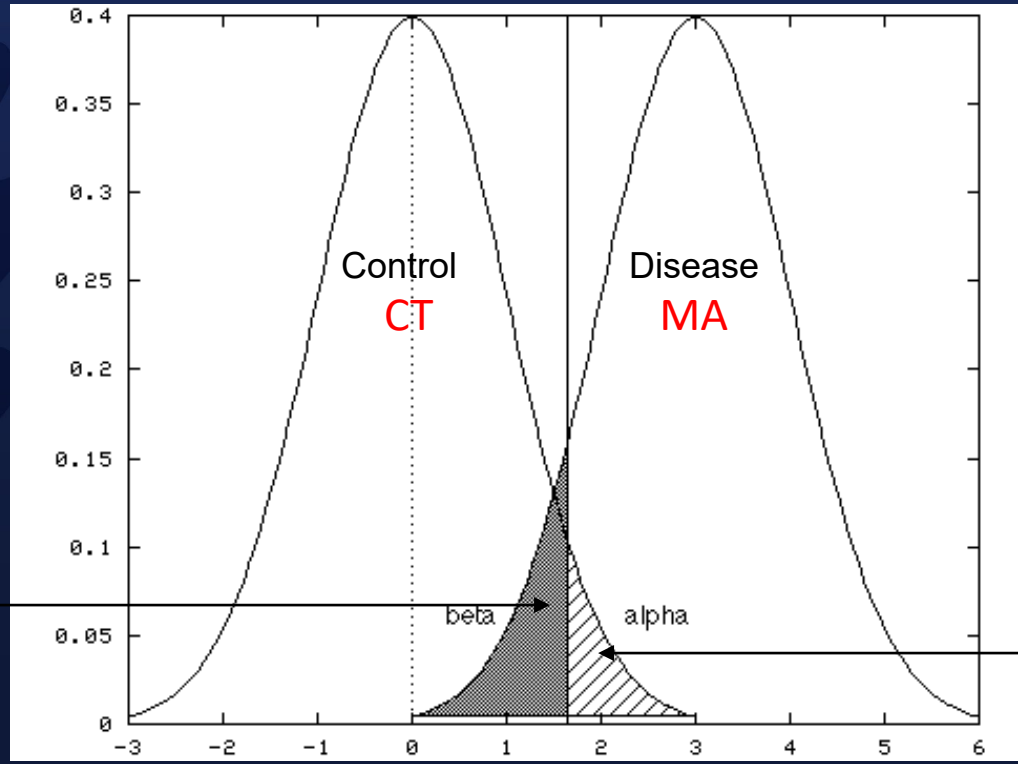
- 3 criteria are specified to determine the appropriate sample size:
  - the level of precision (standard deviation)
  - the level of confidence or risk (confidence interval)
  - the degree of variability in the attributes being measured (how much each measurement varies from the mean)

# Level of Precision

- The *level of precision*, sometimes called *sampling error*, is the range in which the true value of the population is estimated to be.
- We base our calculation *on the standard deviation of our sample*. The greater the sample standard deviation, the greater the standard error (and the sampling error). The standard error is also related to the sample size. The greater your sample size, the smaller the standard error.

# Standard Deviation





False negative

False positive

# ARRIVE Trial – Sample Size Calculation

## STATISTICAL ANALYSIS

The expected rate of the primary perinatal outcome in the expectant-management group was estimated to be 3.5%.<sup>18</sup> We calculated that enrollment of 6000 women would provide a power of at least 85% to detect a 40% lower rate of the primary outcome in the induction group than in the expectant-management group, at a two-sided type I error rate of 5%. This power analysis incorporated the assumption that for 7.5% of the women, management would not be consistent with the protocol of the assigned strategy.

The expected rate of the primary perinatal outcome in the expectant-management group was estimated to be 3.5%.<sup>18</sup> We calculated that enrollment of 6000 women would provide a power of at least 85% to detect a 40% lower rate of the primary outcome in the induction group than in the expectant-management group, at a two-sided type I error rate of 5%.

# Inferential Statistics

- Qualitative
  - Chi Square / Exact Tests
- Quantitative
  - T Test / Mann-Whitney U
  - ANOVA / Kruskal-Wallis
- Regression
  - Linear
  - Logistic
- Time to Event
  - Survival Curve
  - Cox Proportional Hazard

# Summary Table of Statistical Tests

Measurement Scale	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	$X^2$ or bi-nomial	$X^2$	Macnarmar's $X^2$	$X^2$	Cochran's Q	
Rank or Ordinal		Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friendman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r
		Factorial (2 way) ANOVA				

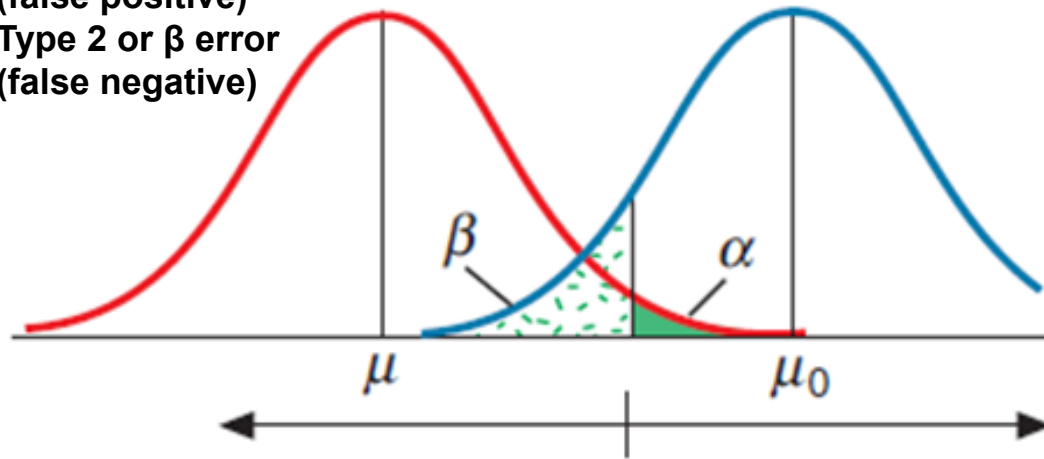
(Adapted from M. [Plonsky](http://www4.uwsp.edu/psych/stat/indexTests.htm) at [www4.uwsp.edu/psych/stat/indexTests.htm](http://www4.uwsp.edu/psych/stat/indexTests.htm))



# Power

- The probability that a study will detect the phenomenon studied when it exists is called “power”.
- Power depends on group variability, size of the sample, the true nature of the phenomenon being observed, and the level of significance.
- A good clinical study should inform the calculated power of the sample, so the reader can evaluate “non-statistically significant” results.

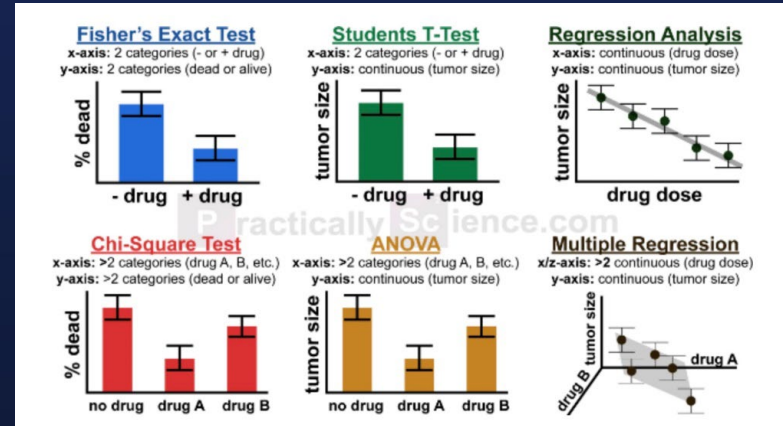
Type 1 or  $\alpha$  error  
(false positive)  
Type 2 or  $\beta$  error  
(false negative)



$H_0$  is accepted     $H_0$  is rejected

# Questions to Ask about the Stats


- Are these the right statistical tests for the data?
- Can I understand the tests and the output they will give me?
- Are you cheating or mathing the numbers to hide something?
- Do you know what you are doing?



**There seems to be a pervasive notion that  
"you can prove anything with statistics."**

**"There are three kinds of lies: lies, damned  
lies, and statistics"**





# RESULTS

(aka, show me what you found)



# Results

- What data are presented?
- Do the data follow from the investigators' methods?
- Is it clear where the data came from?
- Is it clear how the data was obtained?
- Are all the data presented, and are all groups accounted for?
- If all the subjects or groups are not accounted for, how did the authors address this issue?
- Did the investigators perform an intent-to-treat analysis?
- What do the results show?
- Could these results have been from chance?



# ARRIVE Trial – Results

**Table 2. Primary Perinatal Outcome and Components.\***

Outcome	Induction Group (N = 3059)	Expectant- Management Group (N = 3037)	Relative Risk (95% CI) <sup>†</sup>	P Value <sup>‡</sup>
	<i>no. (%)</i>			
Primary composite outcome	132 (4.3)	164 (5.4)	0.80 (0.64–1.00)	0.049
<b>Maternal</b>				
Cesarean delivery — no. (%)	569 (18.6)	674 (22.2)	0.84 (0.76–0.93)	<0.001 <sup>‡</sup>



# DISCUSSION

(aka, add context to the results)





# Discussion

- Was the hypothesis verified?
- Did they summarize the main research findings, the unique aspects of the study and the conclusions that can be drawn?
- Did they explain how and why these results were obtained, along with their significance.
- Did they review other studies relating to their investigation and explain what, if any, different differences exist among their findings and those reported in the literature?



# ARRIVE Trial - Conclusion

- In summary, we found that elective labor induction at 39 weeks of gestation did not result in a greater frequency of perinatal adverse outcomes than expectant management and resulted in fewer instances of cesarean delivery. These results suggest that policies aimed at the avoidance of elective labor induction among low-risk nulliparous women at 39 weeks of gestation are unlikely to reduce the rate of cesarean delivery on a population level; the trial provides information that can be incorporated into discussions that rely on principles of shared decision making.



# Questions about the discussion

- What conclusions did the authors draw from the data? **Would I draw the same conclusions?**
- Are the authors' conclusions based on the methods and data?
- Do the results from the data disagree with the authors' conclusions? If so, going back to the Results section to see where the discrepancy in interpretation occurred may be helpful.
- Do the results and conclusion apply to the patients in my practice?
- How does the study advance knowledge?
- Do the authors acknowledge limitations of the study? Are there additional limitations that should be included?
- Do the authors adequately account for any unexpected results?





# References

## How to become good at peer review: A guide for young scientists

📅 December 13, 2013   👤 Jennifer Raff

<https://violentmetaphors.com/2013/12/13/how-to-become-good-at-peer-review-a-guide-for-young-scientists/>

## Peer Review of a Manuscript Submission A How-To Guide for Effective and Efficient Commentary

Allen LA, Ho PM. Peer Review of a Manuscript Submission: A How-To Guide for Effective and Efficient Commentary. *Circ Heart Fail*. 2017 Dec;10(12):e004766.