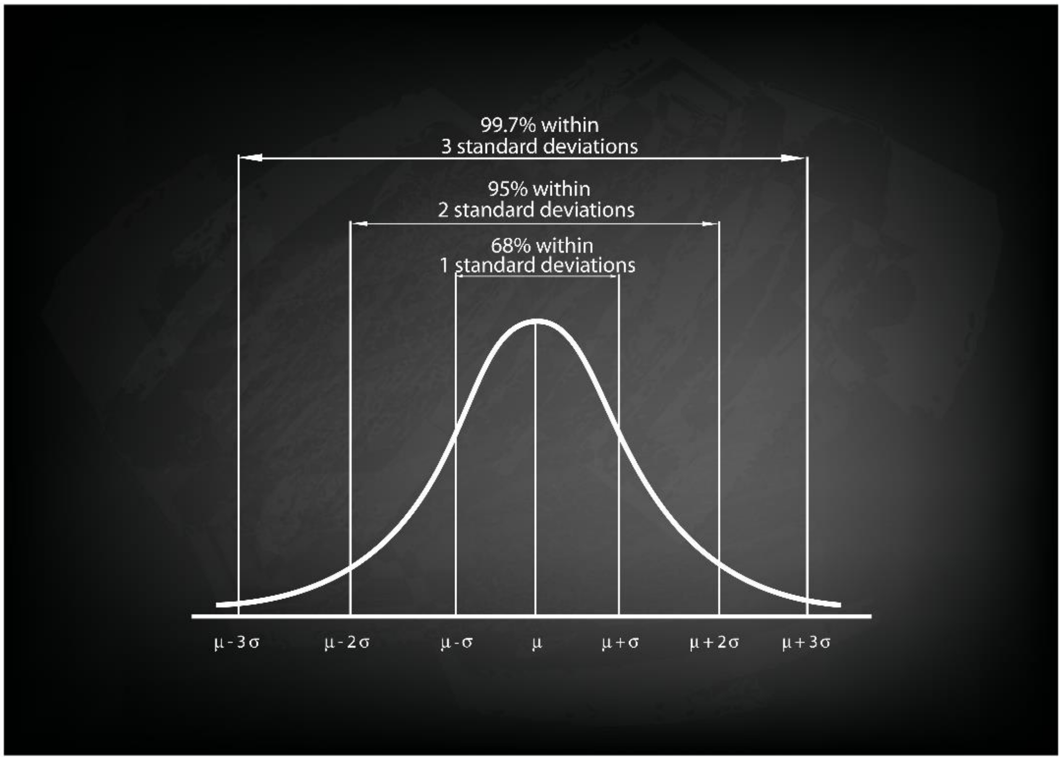
Medical Statistics for the Non-Mathematician

By

Stuart M Caplen, MD

Reading medical literature requires an understanding of statistics, which are used in every experimental trial. Unfortunately, the mathematical underpinning of statistical analysis involves rather complicated equations and theories. This article will discuss some of the definitions and concepts needed to understand the statistics used in the medical literature without delving into the more complicated mathematics, essentially medical statistics for the non-mathematician.

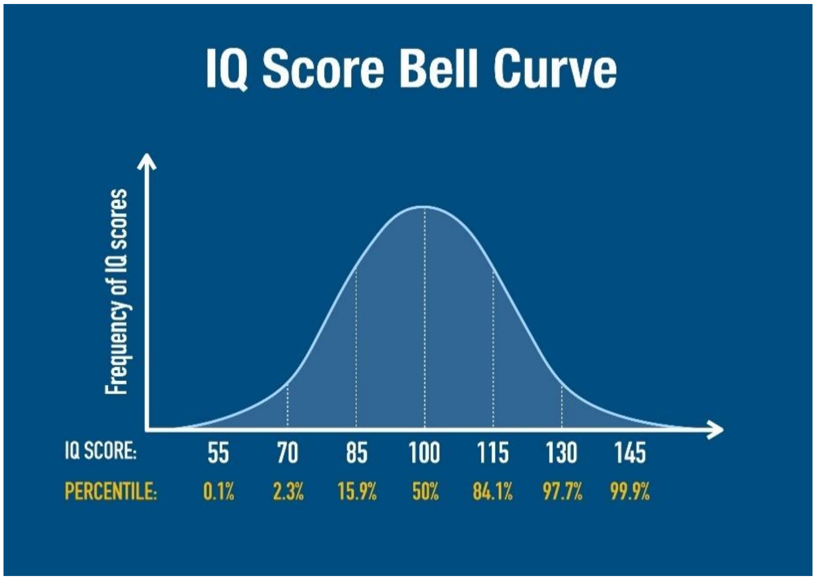
Statistical Terms Used in the Medical Literature



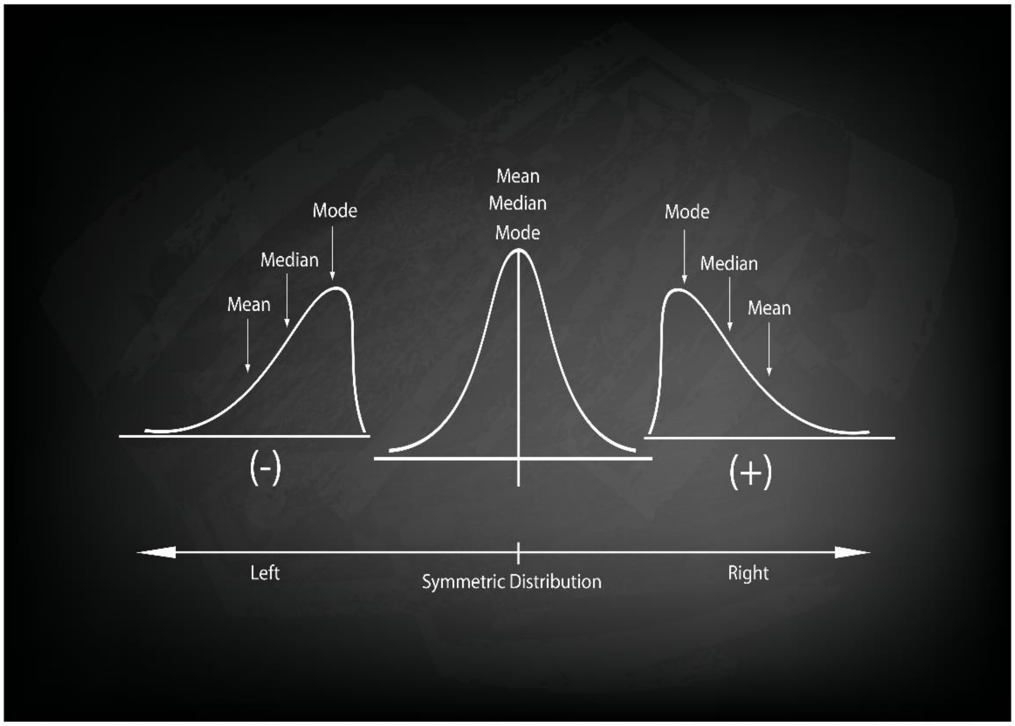
Normal bell-shaped data distribution - μ = population mean, σ = standard deviation

Normal Distributions

The graph above represents a normal bell-shaped curve distribution. It represents average measurements from a sample with most of the values being near the center of the curve. This type of data curve is seen very frequently when performing statistical measurements and is very useful in statistical analysis. A normal distribution could represent diverse data sets such as the intelligent quotients (IQ) in a population or the average number of times asthmatics use their inhaler when they have an asthma attack. The center line labelled μ is the mean or average. The standard deviation (σ) is a measure of how dispersed the data is in relation to the mean. A low standard deviation means the data are clustered around the mean, and a large standard deviation indicates the data points are more spread out. The curve represents how frequently a data value is measured with most results near the mean and fewer results occurring further out from the mean. 95% of the measured values fall within two standard deviations above or below the mean. A result that is outside of two standard deviations from the mean has a much higher chance of being significantly different from the general population or sample being studied. In the graph below representing IQ, 100 is the average or mean in the population with most samples of people centered around that number. A sample of people with an average IQ over 145 would be much more likely to have a statistically significant different IQ score from the 100 IQ mean than a sample of people with an average IQ of 105.



There are a number of different statistical tests that can be used for analyzing data in both normal distributions and distributions of data not in the normal bell-shaped curve, which are beyond the scope of this article.



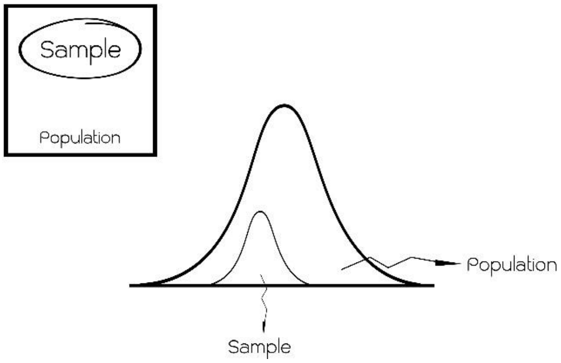
Central Tendencies of Data

There are 3 ways of describing a central tendency of data; mean, median, and mode. Mean is the average of all the data, median is the number in the middle of the data, and mode is the number that occurs most frequently in the data. As can be seen in the figure above, when the data map forms a symmetric bell-shaped curve all three measures of central tendency are the same, but if the data curve is skewed the three values will be different. The median has the advantage of not being affected as much by outlier values as the mean. The mode is not used very much in the medical literature.

In the number series 1, 3, 5, 7, 49, the mean is 13 and the median is 5, which is less affected by the outlier value of 49 than the mean. Experimental results can be affected by which measure is chosen by the author(s).

Statistical Definitions and Measurements

A **population** is the total group of interest, and a **sample** is a smaller subset of the population that will be examined to try to make conclusions about the larger population. There are formulas that can be used to determine how big the sample size must be to accurately represent the whole population.



Prevalence is the proportion of a population that have a condition during a specific time frame. **Incidence** is the proportion of a population who newly develop the condition during that time frame. If in a town of 100 people, 40 in total had contracted a disease and 10 had become newly ill in the last year, the prevalence would be 40% and the incidence 10%.

In **randomized trials**, participants are randomly assigned to receive or not receive treatment. These are generally better designed trials with potentially less bias than non-randomized ones. They avoid much of the **selection bias** that would come from participants with certain characteristics being chosen preferentially to be in one group that might affect the results, and also the **volunteer or self-selection bias** that can occur when individuals who volunteer for a trial differ in clinical characteristics from those who do not. However, even in randomized trials there should be an analysis of the characteristics of the different groups or people being tested as one group by chance be sicker or older than the other and the outcomes may be due to that difference rather than from the therapy being tested. [1]

Prospective studies designed before data collection or treatment will generally have less sources of bias and confounding factors than **retrospective studies**, which typically rely on chart review of previously collected results.[2]

Accuracy is how close a measurement is to an accepted value or gold standard. **Precision** is how reproducible a result or measurement is, even if it the result is incorrect or not accurate. Having high precision and accuracy together produces the best results. If a man weighing 150 pounds weighs himself on a scale and the scale indicates 150 pounds and when repeated reads 150 pounds again, the scale is both accurate and precise. If instead the scale read 165 pounds and repeated weighing attempts also read 165 pounds, the scale would have precision but be inaccurate.

Sensitivity (true positive rate) is the ability of a test to correctly identify subjects with a specific condition (true positives) out of all people with the condition (true positives + false negatives). **Specificity** (true negative rate) is the ability of a test to correctly identify all the subjects without the condition (true negatives) out of all people without the condition (true negatives + false positives).[3]

In a meta-analysis of CT scanning to diagnose appendicitis the overall sensitivity was 95% and the specificity was 94%. This means that in this analysis CT scanning correctly diagnosed 95% of the patients with appendicitis, but missed 5%. CT scan correctly diagnosed 94% of those who did not have appendicitis and diagnosed 6% incorrectly as having appendicitis when they did not.[4]

Positive predictive value is the probability that patients with a positive test result truly have the disease (true positives/(true positives + false positives)). **Negative predictive value** is the probability that subjects with a negative test result do not have the condition (true negatives/(true negatives + false negatives)). Prevalence of disease plays a significant role in predictive values. The lower the prevalence of a disease the more false positives there will be in testing which can be important clinically. If a test is studied in a high prevalence of disease sample, there may only be a few false positives. If the test is then used as a screening test in a general population with a low prevalence of disease there may be a much higher percentage of false positives that may require additional testing to rule out the disease.[3] This was one of the issues when HIV screening tests were used to test the general population; there were more false positives which required confirmatory testing compared to using the test in a targeted high-risk high prevalence segment of the population.[5]

When trials are repeated it is not uncommon to get slightly different results. A 95% **confidence interval (CI)** is a set of values that if an experiment is repeated the new result would have a 95% chance of being found between those two values.[6] It also represents the set of values from the sample being tested between which there is a 95% chance the true population mean lies. Larger sample sizes (larger trials) and bigger differences between the two groups being compared will generally have smaller confidence intervals and a higher confidence that the result is precise. An example of a CI would be a mean trial result of 78 with a 95% CI of 45 to 82.

The Odds ratio (OR) is a measure of association or correlation between a variable and an outcome. It tells you how much the presence or absence of a variable has an effect on the presence or absence of an outcome. The OR is used to figure out if a particular exposure (such as asbestos or sun exposure) is a risk factor for a particular outcome (such as mesothelioma or melanoma).

An odds ratio of more than 1 means that there are higher odds of an outcome happening with exposure to the variable. The larger number the odds ratio is, the stronger the association between the two events. An odds ratio of less than 1 means that there are lower odds of an outcome occurring with exposure to the variable. An odds ratio of exactly 1 means that presence of the variable does not affect the odds of the outcome occurring.[7,8] Odd ratios are typically calculated with 95% confidence intervals. If the spread of values of an OR confidence interval includes the number 1 as a possible result, such as the 95% CI = 0.8-1.3, the calculated OR is not likely to be statistically significant.[7,9]

An example of a positive OR is that asbestos exposure has an odds ratio of 3.7 (95% CI = 1.7 to 7.8) for the development of mesothelioma, meaning those exposed to asbestos in one study had an average 3.7 times the risk of developing mesothelioma than those unexposed, although the true value based on the CI could be anywhere from 1.7 to 7.8 times the mesothelioma risk with 95% certainty.[10]

One needs to be careful in interpreting odds ratios as the correlation or association of two events does not necessarily mean that one actually caused the other.

A **Hazard ratio (HR)** is a measure of the effect of an intervention over time and is most commonly used for analysis of survival. It is calculated by dividing the hazard in the treatment group by the hazard in the control group. Hazard is defined as the probability that an individual would experience an event such as death or relapse after receiving the treatment being studied. A **HR** of 0.5 means that at any particular time, half as many patients in the treatment group are experiencing an event compared to the control group. A **HR** of 1 indicates event rates are the same in both groups. A **HR** of 2 means that at any particular time twice as many patients in the treatment group are experiencing an event compared to the control group. A hazard ratio is also reported with a 95% CI.[11] In a real-life example, a recent study compared mortality rates in subjects who received four doses of COVID vaccine versus those who received three doses. The reported hazard ratio was 0.22 (95% CI = 0.17-0.28) meaning that patients in the four-dose group were 78% less likely to die over the course of the 40-day period of the study than those who received three doses. The actual value at a 95% confidence level would lie between 72% to 83% less likely to die in the four-dose group.[12] Notably, the hazard ratio takes into account the timing of death (or other events) and not just the overall survival by the end of the study period.

Relative risk reduction (RRR) is the percent reduction in a measured outcome between the experimental and control groups. This measure is used in medical literature, but is not a very good way to compare outcomes as it can amplify small differences and make insignificant findings appear more significant. The RRR doesn't reflect the actual risk of a measured outcome, but can be used by authors to make weak results look better. For example, if 2% of the placebo group die and only 1% of the treatment group die the RRR would be ((2-1)/2)=0.5 or 50%. In the above example, if the numbers were instead a 40% death rate in the placebo group and 20% in the treatment group, the RRR will still remain 50% ((40-20)/40) even though 20% more people will be saved by the treatment compared to only 1% in the first example.[13,14]

Absolute Risk Reduction (ARR) is a better measure than RRR and is the arithmetic difference between the event rates in the two groups. In the 2%/1% example above, the ARR is 0.01 (1%). In the 40%/20% example above the ARR is 0.20 (20%) and better reflects the difference in incidence rates than the RRR.[13,14]

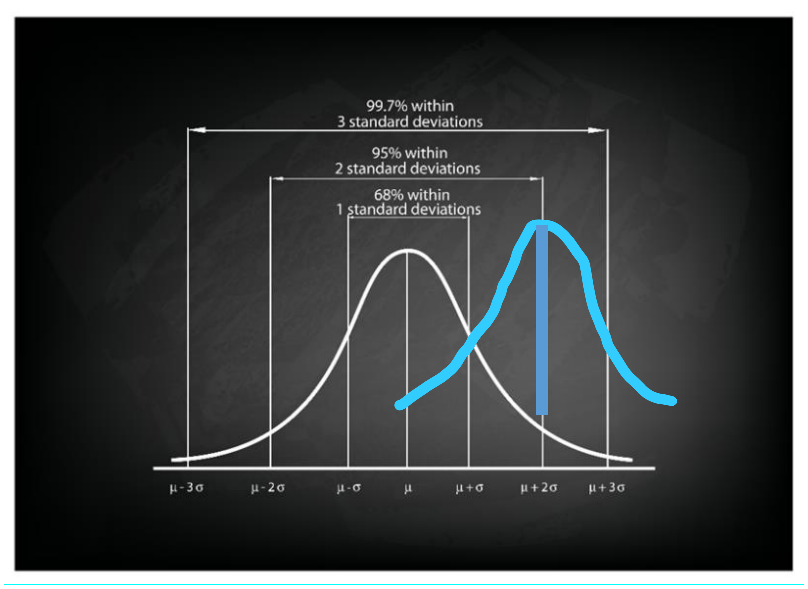
Number needed to treat (NNT) is the number of patients needed to treat to prevent one additional bad outcome. As an example, for people with known heart disease who took statins for 5 years, 83 patients would have to be treated to save one life and 125 treated to prevent one stroke.[14] NNT is calculated by the equation 1/ARR.[13,14]
Number needed to harm (NNH) is the number of people needed when treated for one to develop a harmful side effect. As an example, for every 50 patients who take statins for five years, one patient would be expected to develop diabetes, so the NNH would be 50.[15]

Statistically Significant Differences

Most experiments start with the **null hypothesis** which is the baseline assumption that there is no significant difference between the two populations or samples being tested, such as subjects in a treatment group compared to a placebo group. Only if the data demonstrates a significant difference between the two groups is the null hypothesis rejected and a difference between the groups acknowledged, such as the tested drug successfully treating the disease compared to a placebo.

Although **p values** or **probability values** are used in just about every scientific trial, there are some confusing issues in understanding their significance**.** P values represent the likelihood that the difference in results of two groups is by random chance. P values range from 0 to 1. The lower the p value, the more statistically significant the difference in results are and the more likely that the two groups are different (treatment versus placebo for example). The larger the p value, the less likely any differences found between the two groups are statistically significant and the more likely that the two groups are similar and any differences found are due to chance.

Results of two groups that are separated by two standard deviations or more are typically going to be found to be statistically different by p value and not the result of random chance. Picture it as two bell-shaped curves of data superimposed on each other. The further apart the means of those curves are, the more likely the two groups will be found to be significantly different, as seen in the figure below. The actual calculations used to determine statistical significance are fairly complicated. T- or Z-tests and tables can be used to decide if results are statistically significant but the use of them is beyond the scope of this article.



Comparison of two data curves and mean values

The p value represents the likelihood of getting the given results if the null hypothesis was correct, and there is no difference between the two groups being testing. By convention, for most trials, a p value greater than 0.05 supports the null hypothesis and there was at least a 5% probability of getting the difference in results simply by chance. A p value less than 0.05 indicates that there is likely a significant difference in the populations or samples, such as a drug that significantly improves mortality over placebo and there was a less than 5% probability of getting the results simply by chance. There is an inherent error rate in conclusions of statistical significance using p values as the standard determination of statistical significance. Choice of the 0.05 p value level for significance is an arbitrary choice, and some investigators may decide to use a lower number as the point of significance to reject the null hypothesis. In genetics research, for example, reduction of false-positive risk is sometimes achieved by setting p values at 0.00000001 or lower.[16] Trials with a large number of participants or with large differences between the groups being compared may also decide to use a lower point of significance than 0.05 as the standard.

One of the caveats to remember is that just because a result has been shown to have statistical significance or a p value of less than 0.05 does not mean that it is true. By itself, the p value is not necessarily a good test of a hypothesis or evaluation of a clinical model and is not a substitute for scientific reasoning. Scientific conclusions and decision-making should be based on more than whether a p value falls below an arbitrary threshold.[16,17]

One should also understand that statistical significance is different than clinical significance or importance. A positive finding in a trial may be statistically significant but have no real clinical effect. An example might be a cancer drug that extends survival by three days that is found to be statistically significant compared to placebo but might not really represent a significant clinical improvement.

Statistical Errors

It is rare to have a result that is 100% certain and there are errors that may occur because of this. Using a p value of .05 means that there is still up to a 5% chance that the conclusions of statistical significance may be incorrect. A **type I error** results in a **false positive** leading to finding a significant difference between two populations or samples when one does not truly exist. A type I error occurs when a null hypothesis is rejected, or a significant difference between the groups is found, when there was really no difference between the groups and they were statistically equivalent. Type I errors increase as p values used to define significance increase, so there is more chance of a type I error when 0.05 is used as the point of significance than when 0.01 is used

A **type II error** is when there was a significant difference between the tested samples but it was not recognized. The null hypothesis was accepted when it should have been rejected creating a **false negative**. Type II errors tend to increase as the p value set for significance decreases, so there is more chance of a type II error when 0.01 is used as the point of significance than when 0.05 is used.

Non-Inferiority Trials

As opposed to standard experimental trials, there are also **non-inferiority trials** where a new treatment is compared to a known standard one. Non-inferiority trials may be used in cases where a treatment for a disease already exists and it would not be ethical to compare a new drug to a placebo. In those trials, the null hypothesis is reversed from normal and states that the two treatments are different and the new treatment is inferior to the old one. If the the two treatments are found to be in equivalence, the null hypothesis of inferiority is rejected and non-inferiority of the new treatment compared to the old one is supported.[18]

Conclusion

Medical statistics help make sense out of data, and an understanding of commonly used statistical terms will help a reader better interpret medical literature results. However, clinical and scientific judgment is still required as a result that has statistical significance may not actually be true or have clinical significance. Statistical analysis has allowed medical science to move forward, but it is not perfect and the medical community as a whole needs to both understand and strive to improve the use of statistical analysis in the medical literature.

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

------Mark Twain attribution to Benjamin Disraeli.[19]

Author’s note: If you wish to learn about biases and statistical problems in the medical literature leading to incorrect conclusions, the topic is discussed in a FibonacciMD blog article for which *AMA PRA Category 1 Credit(s)™* are available; **Potential Bias and Incorrect Results In the Medical Literature, and Why You Shouldn’t Believe Everything You Read,** at this link https://www.fibonaccimd.com/post/potential-bias-and-incorrect-results-in-the-medical-literature.

References

[1]Tripepi G et al. Selection Bias and Information Bias in Clinical Research. Nephron Clin Pract 2010. Retrieved from: https://www.karger.com/Article/Fulltext/312871#

[2]Prospective vs. Retrospective Studies. Statsdirect. 2021. Retrieved from: https://www.statsdirect.com/help/Default.htm#basics/prospective.htm

[3]Understanding medical tests: sensitivity, specificity, and positive predictive value. HealthNewsReview.org. 2022. Retrieved from: https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value/

[4]Rud B et al. Computed tomography for diagnosis of acute appendicitis in adults. Cochrane Database Syst Rev. 2019 Nov 19;2019(11). Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6953397/

[5] False-Positive HIV Test Results. CDC. May 2018. Retrieved from: https://www.cdc.gov/hiv/pdf/testing/cdc-hiv-factsheet-false-positive-test-results.pdf

[6]McLeod S, What are Confidence Intervals in Statistics? Simply Psychology. June 10, 2019, updated 2021. Retrieved from: https://www.simplypsychology.org/confidence-interval.html

[7]Szumilas M. Explaining Odds Ratios. J Can Acad Child Adolesc Psychiatry. 2015 Winter; 24(1): 58. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/#

[8]Stephanie Glen. "Odds Ratio Calculation and Interpretation". StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/probability-main-index/odds-ratio/

[9]Tenny S, Hoffman MR. Odds Ratio. [Updated 2021 May 30]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK431098/

[10]Pintos J et al. Risk of mesothelioma and occupational exposure to asbestos and man-made vitreous fibers: evidence from two case-control studies in Montreal, Canada. J Occup Environ Med. 2009 Oct;51(10):1177-84. Retrieved from: https://pubmed.ncbi.nlm.nih.gov/19749604/

[11]Albarqouni L. Tutorial about Hazard Ratios. Students 4 Best Evidence. 5th April 2016. Retrieved from: https://s4be.cochrane.org/blog/2016/04/05/tutorial-hazard-ratios/

[12]Arbel R et al. Second Booster Vaccine and Covid-19 Mortality in Adults 60 to 100 Years Old. A pre-publication study. Research Square. March 24th, 2022. Retrieved from: https://assets.researchsquare.com/files/rs-1478439/v1/24514bba-2c9d-4add-9d8f-321f610ed199.pdf?c=1648141784

[13]Flaherty RJ. A Simple Method for Evaluating the Clinical Literature. Fam Pract Manag. 2004 May;11(5):47-52. Retrieved from: https://www.aafp.org/fpm/2004/0500/p47.html

[14]Understanding absolute and relative risk reduction. The University of Western Australia. https://www.meddent.uwa.edu.au/__data/assets/pdf_file/0005/2670593/Risk_reduction_guide0.2.pdf

[15]Statins Given for 5 Years for Heart Disease Prevention (With Known Heart Disease). The NNT. Updated: November 2, 2013. Retrieved from: https://www.thennt.com/nnt/statins-for-heart-disease-prevention-with-known-heart-disease/

[16]Andrade C. The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives. Indian J Psychol Med. 2019;41(3):210-215. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6532382/#ref8

[17]The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 09 Jun 2016. Retrieved from: https://www.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108?needAccess=true

[18]Hahn S. Understanding noninferiority trials. Korean J Pediatr. 2012;55(11):403-407. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3510268/

[19]Twainquotes.com, Directory of Mark Twain's maxims, quotations, and various opinions, Retrieved from: http://www.twainquotes.com/Statistics.html

IMIT takes pride in its work, and the information published on the IMIT Platform is believed to be accurate and reliable. The IMIT Platform is provided strictly for informational purposes, and IMIT recommends that any medical, diagnostic, or other advice be obtained from a medical professional. Read full disclaimer.