Statistics as a Tool in the Physician's Black Bag

by Robert W. Baer, PhD

ATSU | A.T. Still University
FIRST IN WHOLE PERSON HEALTHCARE

Each time we encounter new medical information presented using unfamiliar statistical terminology, we have an unprecedented opportunity to self-educate.

Robert W. Baer, PhD, is Professor of Physiology at the Kirksville College of Osteopathic Medicine, A.T. Still University of Health Sciences, Kirksville, Missouri. In addition to physiology, he teaches evidence-based medicine to first year medical students and biostatistics to graduate students.

Abstract

The era of evidence-based practice began in the 1990s with the hope patient outcomes would be improved by eliminating clinical bias and clinically unsound practices. Clinical guidelines which had been previously written by panels of experts were slowly replaced by careful analysis of existing clinical studies and more rigorous designs of new clinical studies based on sound scientific and statistical principles. This still leaves the practitioner with the responsibility of understanding what the evidence is showing them. This article reviews the statistical thinking that underlies the evidence-based literature. We will review some evolutionary changes to statistical analysis being advocated by statisticians and discuss some nuances related to the use of statistics describing diagnosis and treatment in clinical settings.

Introduction

Years ago, I was asked to do some statistical consulting for a prominent surgeon who was writing up his pre-clinical experiments for a new *in utero* procedure. I looked at his dataset and began to explain how he could go about analyzing the data. After only a couple of sentences, he interrupted me and said, "I don't want to know all that, just give me a p-value." I took his dataset, did the statistical analysis for him, and supplied him with the desired p-value. The next time I saw this surgeon, I had moved across the country, and he was on the national evening news explaining the tremendous impact of his new procedure. Neither he nor the news anchor even once mentioned statistics or p-values.

The moral of this story might appear to be that physicians don't need to understand statistics to help their patients. In fact, this surgeon understood that he would need to do some statistics that produced a p-value if he was going to get his animal study published and convince his colleagues that he could safely perform the operation on human patients. In this article, we will discuss some big picture statistical concepts, including some unappreciated aspects of the p-value, in the hopes that they can help you conduct a more effective, evidence-based clinical practice. We will start with some basics, turn to some recent statistical trends, and finish with common statistics used in clinical diagnosis and treatment.

Descriptive Statistics and Confidence Intervals

If one reads the clinical literature, one is likely to encounter confidence intervals. A 95% confidence interval (95% CI) is

an estimate of how confident we are about where the middle value of the population falls. It does not tell us about the spread of values within the population. It is really a statement about the validity of the statistical technique we have used to determine the center of the population. The confidence interval is directly related to the spread of values in the population and inversely related to the square root of the sample size. The more confident we wish to be about having included the real center of the population, the wider the confidence interval will need to be. Confidence intervals are based on standard error and not standard deviation. They depend not only on both sample size and sample spread but also on sampling theory.

Al Kibria[1] tells us that based on the NHANES database[2] the population mean diastolic blood pressure is 70.7 mmHg with a 95% CI of (70.4, 71.0) mmHg. This interval can be easily calculated from the reported mean (70.7 mmHg) and the reported standard error (0.2 mmHg). The reason this 95% confidence interval is so tight is that the data came from 16,103 subjects, and thus, the standard error is small. This confidence interval helps us fix the middle value of the population with a known certainty.

Confidence Intervals Don't Describe Clinical Variability

Let's use the diastolic blood pressure example again, but this time to see what range of diastolic pressures, we might expect in 68% of patients. If diastolic blood pressure were to be normally distributed, 68% of the pressures would fall between one standard deviation below and one standard deviation above the mean diastolic blood pressure. Notice that we are using standard deviation and not standard error. Standard deviation can be calculated by multiplying standard error by the square root of sample size. Thus, from Al Kibria[1] we get a standard deviation of $\sqrt{16103} \cdot 0.2 = 25.4$ mmHg. Based on our normality assumption, 68% of all individuals would have blood pressures of 70.7±25.4 mmHg or between 45.3 and 96.1 mmHg. This interval is much wider than the 95% CI of (70.4, 71.0) mmHg that we calculated above. It makes the point that 95% confidence intervals have nothing to do with the range of values that we might expect in our patient population. In summary, confidence intervals depend on standard error and tell us where the central value of the population might lie. Standard deviation is the variable of interest if we wish to know what range of values we might expect to see across a heterogeneous population of patients.

Null Hypothesis Statistical Testing

Most of the current clinical and biomedical literature makes use of what is referred to as Null Hypothesis Statistical Testing (NHST). In the NHST framework, the investigator proposes a straw man null hypothesis that drug X does not improve heart attack outcomes in hypertensive patients with the hope that he can "disprove" this hypothesis. If the null hypothesis is disproved, the alternative hypothesis is invoked, and the investigator concludes that drug X is effective. When using the NHST framework, we have the potential of making two types of mistakes. We can wrongly conclude that the null hypothesis is false when it is true (type I error), or we can conclude wrongly that the null hypothesis is true when it is false (type II error).

In practice, investigators determine some maximal acceptable probability of wrongly concluding that drug X is effective called the alpha level, A common choice is alpha = 0.05. In the NHST framework, if the actual p-value is less than the alpha level, the investigator will declare a "statistically significant" difference.

Our p-values are conditional probabilities based on the premise that the null hypothesis is true.[3] Investigators are often tempted to claim that a finding is "more statistically significant" if they get a $p< 0.001$ instead of a $p< 0.05$. This is an inappropriate conclusion. In both cases their p-value falls below the alpha cutoff. Effectively, the investigator has concluded that the null hypothesis untrue in both cases. This also means that any conclusion based on the premise of the null hypothesis being true is meaningless.

Despite the ubiquitous presence of NHST and p-values in the clinical and biomedical literature, the American Statistical Association (ASA) in 2016 issued a statement condemning or at least cautioning against its exclusive use.[4] This statement followed long-standing, independent criticisms of the NHST approach.[5–7] The statement enumerated six principles. The interested reader is invited to read their entire statement, but three principles particularly worth repeating here are:[4]

- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

- A p–value, or statistical significance, does not measure the size of an effect or the importance of a result.

We have previously discussed the first bullet point by noting that p-values are conditional probabilities about the data, given the model/hypothesis framing the experiment, and not a statement about the model/hypothesis itself. The second bullet reminds us that science is an iterative process of building a model that represents reality. Often our model is neither correct nor incorrect. Rather our model is just incomplete. This bullet reminds us that it is important to keep in mind nuances and gray areas when integrating results from new studies. The third bullet reminds us that how much something changes can be as important or more important than whether it changes.

Effect Size

The 2016 American Statistical Association stopped short of banishing the term "statistically significant." Three years later, an editorial in the American Statistician did just that.[8] The reporting of p-values was not discouraged, but the editorial emphasized the necessity of providing important contextual information. Effect sizes have been forwarded as a key component of providing context. An effect size is simply the magnitude of the change that occurs with an intervention.[9] In statistical practice, effect size measures are often normalized to some measure of variability. Examples include Cohen's d and Pearson's correlation coefficient but there are many others. The precise calculations are not important because online calculators are readily available.[10] What is important is how effect sizes can provide context to clinical studies.

If you were considering a new drug that could lower blood pressure and were told that the drug would lower a patient's mean blood pressure by 1 mmHg if taken for a year, you would likely not be impressed. If the drug lowered mean blood pressure by 20 mmHg, 10 mmHg, or even 5 mmHg you might keep reading depending on your patient's needs. The problem with p-values is that you can have situations where results are statistically significant but not clinically significant, or conversely, you can find situations where results are clinically significant but not statistically significant. Statisticians have begun to suggest novel approaches for handling these differences which include enhanced variations on the p-value.[11–13]

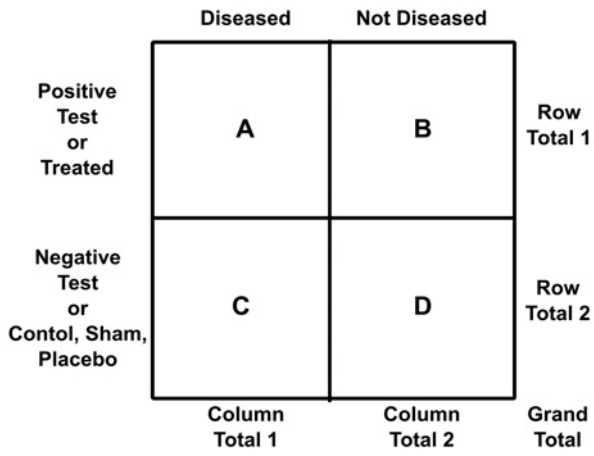

Figure 1. Using 2 x 2 tables. A standard 2 x 2 layout can be used to calculate statistics related to diagnostic testing or treatment success. The first set of row labels is used with diagnostic testing statistics. The second set of row labels is used with treatment statistics.

Diagnostic Testing and Statistics

Sensitivity and specificity are the two terms most physicians associate with diagnostic testing. Sensitivity measures the fraction of diseased patients that have a positive test. Specificity measures the fraction of not-diseased patients that have a negative test. Values for sensitivity and specificity are typically generated by diagnostic test manufactures. Categorization of diseased and not diseased patients to generate sensitivities and specificities is generally based on a gold standard test which may be expensive, invasive, or only available later in the disease progression. It is critical to understand that sensitivity only contains information about diseased patients and specificity only contains information about not diseased patients. You may wonder why we use the term "not diseased" rather than the term "healthy." It is because with diagnostic testing we are focused on a specific disease state, and the fact that a patient does not have the disease of focus does not imply that they are entirely healthy.

The calculation of sensitivity and specificity is easily appreciated from examining Figure 1 where the top row represents number of diseased and not diseased patients with positive tests, and the second row represents the number of diseased and not diseased patients with a negative test. Sensitivity is the number of patients in cell A divided by column total 1. Specificity is simply the number of patients in cell D divided by column total 2. Both are fractions that can vary between 0 and 1, but they are sometimes

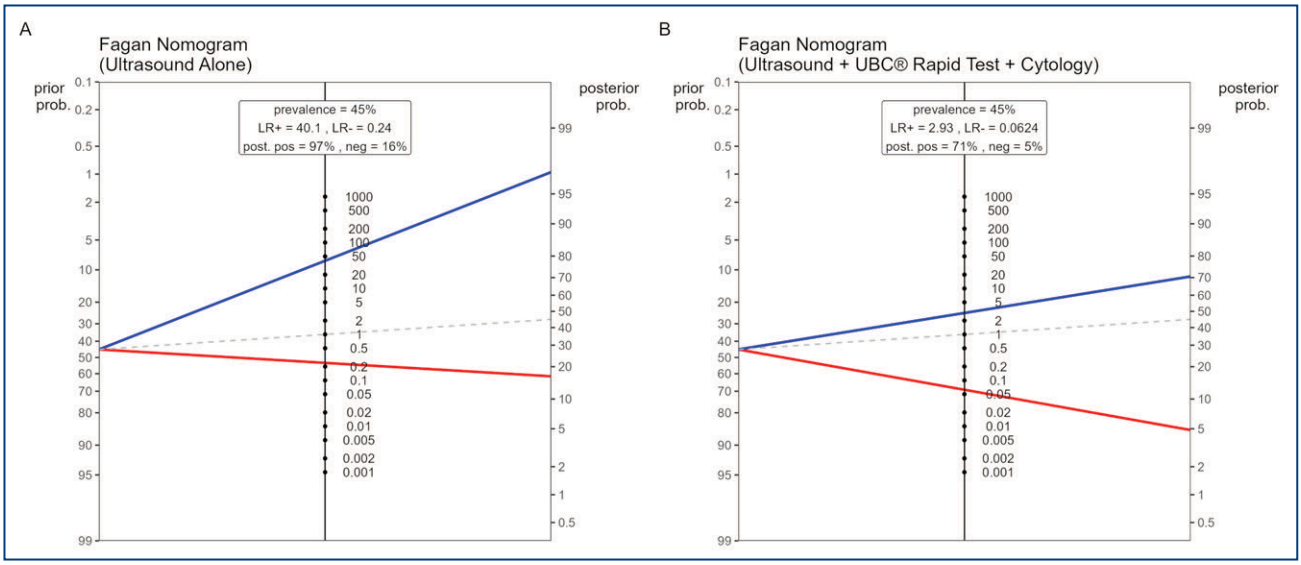

Figure 2. Fagan nomograms. These nomograms convert the pretest probability of a given diagnosis into posttest probability based on the test result. A line connecting the pretest probability to the LR+ if the test is positive or to the LR- if the test is negative is projected to the right side of the graph and the posttest probability is read. The examples show data from Barak et al.17 to compare the diagnosis of bladder cancer using ultrasound alone (panel A) or using ultrasound in combination with UBC® rapid test and urine cytology (panel B). Blue lines show effect of a positive test result on diagnostic certainty. Red lines show the effect of a negative test result on diagnostic certainty. The gray dotted lines represent reference probability that is unchanged by testing (LR = 1).

expressed as percentages by multiplying the fractions by 100. An excellent but somewhat dated resource from the American College of Physicians provides a rich description of sensitivity and specificity values for numerous common diagnostic tests.[14] Medical students are often taught that a test with a high specificity and a positive result is useful for ruling in a disease (SpPIn), and a test with a high sensitivity and a negative result is useful for ruling out a disease (SnNout). Nevertheless, sensitivity and specificity values are difficult to apply in clinical practice because they don't directly relate test results to the likelihood that a patient has a disease on the differential.

In fact, what physicians really care about the fraction of positive tests that are true positives rather than false positives, called positive predictive value (PPV). PPV can be calculated by dividing cell A by the row total 1 in Figure 1. Similarly, physicians are interested in the fraction of negative tests that are true negatives, called negative predictive value (NPV). NPV can be calculated as cell D divided by row total 2 in Figure 1.

Although PPV and NPV are closer to values of use to a physician, they are quite static numbers and are not easily applied to a physicians thinking about an individual patient. Generically, a physician takes a history and physical and, in conjunction with local prevalence, develops an initial probability that a patient has a certain diagnosis. Diagnostic tests are then run to help increase or decrease that diagnostic suspicion. This can be turned in to a semiquantitative process that considers both the initial probability and the strength of the diagnostic test. This process is based on Bayesian statistics[15] and involves the calculation of positive and negative likelihood ratios.[16] The positive likelihood ratio (LR+) is calculated as Sensitivity/(1-Specificity). The negative likelihood ratio(LR-) is calculated as (1-Sensitivity)/ Specificity.

The process of determining the posttest probability of a diagnosis involves converting the pretest probability into a pretest odds, multiplying by the appropriate likelihood ratio for the actual test result, and then converting the post-test odds back into a posttest probability. This can be done graphically using a Fagan nomogram (Figure 2). A straight line from the pretest probability and extended through the appropriate likelihood ratio intersects the right axis at the posttest probability.

As an example, consider a recent study that compared non-invasive diagnostic testing strategies for diagnosing bladder cancer. [17] In one strategy they used ultrasound alone. In the second strategy they used a 3-test combination of ultrasound, UBC® rapid test, and urine cytology.[17] In the study setting,

the pretest probability (local prevalence) of having bladder cancer was about 45%. The sensitivity and specificity for ultrasonography alone were 91.3% and 72.2%, respectively. This translates to an LR+ of 3.28 and an LR- of 0.12, respectively. The sensitivity and specificity for a 3-test combination of ultrasonography, UBC®rapid test, and urine cytology were 95.8% and 67.3% respectively. This corresponds to LR+ of 2.93 and LR- of 0.062.

The posttest probability of bladder cancer for positive and negative tests using each diagnostic regimen can be determined using a Fagan nomogram (Figure 2). The upper blue lines show us that bladder cancer is more likely in the face of a positive test, and the lower red lines show us that bladder cancer is less likely when the test is negative. The two diagnostic testing strategies differ with respect to how much the probability changes with each test result. With ultrasound alone, a positive test greatly increases the probability that our patient has bladder cancer, but a negative test only mildly decreases that probability. With the set of 3 non-invasive tests, a positive result yields a more modest increase in bladder cancer probability, but a negative result decreases the cancer probability more effectively than ultrasound alone. As the authors point out, the 3-test diagnostic strategy has advantages in this scenario because it can be more effective at delaying or avoiding unnecessary surgery which would be the next step in the face of an uncertain diagnosis. Plotting Fagan diagrams for common scenarios in your own clinical practice has the potential to improve the efficiency and reliability of your own diagnostic strategies.

Therapeutic Effectiveness and Statistics

Most clinical studies, including clinical trials, deal with the risk of having a disease. Risk is essentially the probability that a person will become ill or die within a specified period of time.[18] Generally, a new therapy is compared to a control group consisting of the existing standard of care, a placebo treatment, a sham treatment, or even an untreated group. The goal of the study is to see if the new therapy mitigates risk more effectively than the control. We can use the lower set of row labels in Figure 1 to understand some of the terminology that is expressed in making these comparison.

The risk in the treated group is often called the experimental event rate (EER) and can be calculated by dividing the number of subjects that end up in cell A by row total 1. Similarly, the risk in the control group (CER) can be calculated by dividing the number of subjects in cell C by row total 2. Two key measures of treatment effectiveness are often derived from these numbers: absolute risk reduction (ARR) and relative risk reduction (RRR). Absolute risk reduction is the difference between the risk in control subjects and the risk in treated subjects, i.e., ARR = CER – EER.[19] Relative risk reduction normalizes this difference to the control event rate, i.e., RRR = (CER – EER)/CER.[19]

When reading the medical literature, it is essential to distinguish between relative and absolute risks. Generally, absolute risk reduction is more meaningful, but relative risk reduction is often reported by pharmaceutical manufacturers who wish to make a new therapy look more effective. As an example, Marabotti[20] summarized absolute and relative risk reductions for COVID vaccine from a number of manufacturers. Absolute risk reduction ranges from 0.54% to 1.19%, whereas relative risk reduction ranges from 66.9% to 95%. To the uninitiated, it is easier to make a positive case for vaccine use by sharing the relative risk numbers than by sharing the absolute risk reduction numbers. To understand why the numbers are so different we need to look more closely.

To make ARR easier to understand, one can take its reciprocal to calculate the so-called number needed to treat, i.e., NNT = 100/ARR when ARR is in percent.[21] The number needed to treat is the number of individuals that must be treated over the duration to have one less adverse incident. The NNT for COVID vaccines summarized by Marabotti[20] range from 82 to 185 for various vaccines. For comparison, NNT = 100 for antihypertensive therapy to prevent fatal or non-fatal myocardial infarction.[22] Duration of treatment is important and needs to be considered when comparing values. NNT tends to go down with longer treatment times because there are more "potential events" that can be mitigated by treatment. For the antihypertensive studies, treatment times are typically looking at risk across a 5-year time span. The COVID studies are of shorter duration, and it is to be expected that the ARR will increase and the NNT will decrease for COVID vaccination as the post-pandemic era continues to lengthen.

One benefit of relative risk reduction values is that they tend to stay relatively constant across different practice settings.[23,24] For example, the relative risk reduction for an antihypertensive agent may be similar in a family practice setting and a cardiology setting. This is not true for absolute risk reduction. The implication is that the NNT determined from a study carried out in a cardiology practice may overestimate the effectiveness of that same drug in a family medicine setting where the NNT will be higher. This is because the baseline risk of an adverse cardiac event is lower in the family practice setting. If we examine the relationship between ARR and RRR we see that ARR = CER·RRR. In a practice setting where control event rate is low, absolute risk reduction will also be low, and number needed to treat will be high. Conversely, if we know RRR and have a good feel for our local control event rate, we can estimate ARR and NNT for our own practice setting even though we only have data derived in a specialty practice setting. This discussion highlights that we need to carefully pay attention to practice setting and duration of treatment information before applying risk measures to our own practice needs.

Statistics and Evidence-Based Practice

We have talked about some, but clearly not all, of the statistical concepts that are beneficial for understanding the clinical literature. When evidence-based medicine came into its own during the 1990s many pictured that the individual practitioner would be applying evidence at the bedside. We have made some progress toward that goal, but analyzing evidence is a job that will likely continue to be left to the professional statistician, just like clinical lab results are left to the laboratory technologists. However, the individual practitioner is left with taking the statistical information provided by these specialists and incorporating it into daily patient care. Statistics does not need to be understood all at once. Each time we encounter new medical information presented using unfamiliar statistical terminology, we have an unprecedented opportunity to self-educate. New AI tools will make this even easier as they mature. Hopefully, this life-long learning will contribute to ever more positive patient outcomes.

References

1. Al Kibria GM. Racial/ethnic disparities in prevalence, treatment, and control of hypertension among US adults following application of the 2017 American College of Cardiology/American Heart Association guideline. Preventive Medicine Reports. 2019;14:100850. doi:10.1016/j.pmedr.2019.100850
2. Centers for disease Control. NHANES Questionnaires, Datasets, and Related Documentation. Accessed December 12, 2024. https://wwwn.cdc.gov/nchs/nhanes/
3. Gigerenzer G. Mindless statistics. The Journal of Socio-Economics. 2004;33(5):587-606. doi:10.1016/j.socec.2004.09.033
4. Wasserstein RL, Lazar NA. The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108
5. Goodman SN. A comment on replication, P -values and evidence. Statistics in Medicine. 1992;11(7):875-879. doi:10.1002/sim.4780110705
6. Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician. 2006;60(4):328-331. doi:10.1198/000313006X152649
7. Johansson T. Hail the impossible: p -values, evidence, and likelihood. Scandinavian J Psychology. 2011;52(2):113-125. doi:10.1111/j.1467-9450.2010.00852.x
8. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond " p < 0.05." The American Statistician. 2019;73(sup1):1-19. doi:10.1080/00031305.2019.1583913
9. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews. 2007;82(4):591-605. doi:10.1111/j.1469-185X.2007.00027.x
10. Wilson DB. Practical Meta Analysis Effect Size Calculator – Campbell Collaboration. 2023. Accessed December 7, 2024. https://www.campbellcollaboration.org/calculator/equations
11. Betensky RA. The p -Value Requires Context, Not a Threshold. The American Statistician. 2019;73(sup1):115-117. doi:10.1080/00031305.2018.1529624
12. Blume JD, D'Agostino McGowan L, Dupont WD, Greevy RA. Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. Smalheiser NR, ed. PLoS ONE. 2018;13(3):e0188299. doi:10.1371/journal.pone.0188299
13. Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD. An Introduction to Second-Generation p -Values. The American Statistician. 2019;73(sup1):157-167. doi:10.1080/00031305.2018.1537893
14. Black ER, Bordley DR, Tape TG, Panzer RJ, eds. Diagnostic Strategies for Common Medical Problems. 2nd ed. American College of Physicians; 1999.
15. Kruschke JK. Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Academic Press; 2011.
16. Sackett DL, ed. Clinical Epidemiology: A Basic Science for Clinical Medicine. 2. ed. Little, Brown and Comp; 1991.
17. Barak V, Itzkovich D, Einarsson R, Gofrit O, Pode D. Non-invasive Detection of Bladder Cancer by UBC Rapid Test, Ultrasonography and Cytology. Anticancer Res. 2020;40(7):3967-3972. doi:10.21873/anticanres.14389
18. Porta MS, Greenland S, Hernán M, Silva I dos S, Last JM, International Epidemiological Association, eds. A Dictionary of Epidemiology. Six edition. Oxford University Press; 2014.
19. Straus SE, Glasziou P, Richardson WS, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. Fifth edition. Elsevier; 2019.
20. Marabotti C. Efficacy and effectiveness of covid-19 vaccine - absolute vs. relative risk reduction. Expert Review of Vaccines. 2022;21(7):873-875. doi:10.1080/14760584.2022.2067531
21. Furukawa TA, Leucht S. How to Obtain NNT from Cohen's d: Comparison of Two Methods. Biondi-Zoccai G, ed. PLoS ONE. 2011;6(4):e19070. doi:10.1371/journal.pone.0019070
22. McCormack J. Anti-Hypertensives to Prevent Death, Heart Attacks, and Strokes. TheNNT. Accessed December 5, 2024. https://thennt.com/nnt/anti-hypertensives-to-prevent-death-heart-attacks-and-strokes/
23. BMJ Knowledge Centre. Understanding statistics: risk | BMJ Best Practice. Accessed December 6, 2024. https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/how-to-calculate-risk/
24. Irwig L, Irwig J, Trevena L, Sweet M, Tandberg R. Smart Health Choices: Making Sense of Health Advice. Hammersmith Books Limited; 2013.

Disclosure

No financial disclosures reported. Artificial intelligence was not used in the study, research, preparation, or writing of this manuscript. **MM**