

Math is Your Friend: A Consumer's Primer to Understanding Epidemiology

Robert S. Van Howe

Abstract

Mathematics is a tool, but like any tool it can be used correctly or used incorrectly. I will explore how numbers are used and manipulated in the conversations about the health impact of circumcision. The difference between relative risk reduction and absolute risk reduction will be delineated. The derivation of number needed to treat and number needed to harm will be discussed, as well as how the number needed to treat for a urinary tract infection miraculously dropped from 195 to 4. The outcomes of three randomized clinical trials of circumcision's impact on HIV incidence showed remarkably similar results: were the similarities too remarkable? Examples of the abuse of mathematics in the circumcision debate will be presented.

Introduction

We use mathematics to quantify things. This can take the form of measuring or counting. We use mathematics to describe shapes and trajectories using mathematical formulas. We also use math to make comparisons, such as absolute differences and ratios.

Measuring with Confidence

Statistics are based on inference in which "a conclusion is reached on the basis of evidence and reasoning." While the most accurate way of measuring an attribute in a population is to measure every member of the population, this is far too expensive and time consuming. Instead, we measure an attribute in a representative sample of the population and then extrapolate from the representative sample to make an estimate for the entire population. While such estimates are not completely accurate, you can calculate how accurate you believe the estimate to be. For example if we were to measure the height of everyone in a classroom of students, not everyone would be the same height. We would expect some variation. The degree of variation can be estimated by calculating a standard deviation.

When we derive an estimate based on a representative sample, we can calculate how much we trust our estimate in the form of a standard error (the standard deviation divided by the square root of the number of people sampled). As the number of people we sample from the population increases, the standard error decreases. As the standard error decreases, the trust in our estimate increases.

If we wanted to compare two populations, for example the height of the males versus the height of the females in a classroom, there is likely to be a fair amount of overlap of the measured heights of the individuals from both sexes. As our sample size increases, the standard error for each sex decreases and the overlap of how much we trust our estimates for the average height for the two sexes will decrease as well. There comes a point at which we can say with a degree of certainty that the average height of the males and females are different.

Incidence versus Prevalence

One of the most common sources of confusion seen in epidemiology is the difference between incidence and prevalence. Incidence is the number of new cases over a specified period of time. For example, the 2012 American Academy of Pediatric Task Force on Circumcision noted that the incidence of penile cancer was 0.58 per 100,000 *person-years*.¹ This is not the same as saying that 0.58 out of 100,000 people have penile cancer but rather how quickly new cases accumulate. By contrast, prevalence is the number of active cases of an illness within a population. This is typically expressed as 2.4 per 1000 *people* or 0.24%.

One way to illustrate the difference between prevalence and incidence is to look at HIV infections. Globally, the number of new cases of HIV infections peaked at the end of the 1990s.² Since the number of new cases per year has decreased, the incidence is decreasing. With the implementation of antiretroviral therapy, people infected with HIV have been living longer. Consequently, there are more people with new infections than there are people dying from established infections. Consequently, there is a higher percentage of people within the population who are living with HIV and the prevalence is increasing.

Lifetime Risk

One of the statistics that is bantered about is the lifetime risk of acquiring certain illnesses. This cannot be calculated from prevalence because illnesses can come and go, afflict different people for different lengths of time, result in early death, or present at different ages. We can however calculate lifetime risk from incidence estimates. Since incidence estimates are age-adjusted, the lifetime risk is approximately the yearly risk multiplied by the average lifespan, which is 72 years.¹ So for penile cancer in the United States, the lifetime risk would be 0.0000058×72 or 0.0004176 (The precise formula gives an answer of 0.000417512).

Lifetime risk is usually not expressed in this fashion because no one wants to count the number of zeroes following the decimal point, but as the inverse ($1/x$) of this number. In this case, the inverse is expressed as a one in 2395 lifetime risk. To put this in perspective the lifetime risk of breast cancer in women is one in eight. By comparison, penile cancer is a rare illness.

Number Needed to Treat

This can be taken a step further. The 2012 American Academy of Pediatrics Task Force report noted that you needed to circumcise 909 males for that one case of penile cancer. This estimate came from a discussion section of an article³ citing a 1980 opinion piece that assumed that it was impossible for circumcised men to get penile cancer.⁴ We now know that is nowhere near the truth. They also noted that a review article put this number at 322,000.⁵ The review article confused incidence with lifetime risk and failed to multiply it by 72 as discussed above. Neither number is correct. Interestingly, the Task Force had all the numbers at its disposal to make a rough estimate of the number needed to treat but failed to recognize this opportunity or act on it.

Let's do the math they were unwilling to do. The lifetime risk, as we noted above, is 0.0004176. The Task Force report noted that the relative risk reduction for penile cancer by circumcision was between 1.5 and 2.3. If you take the lifetime risk of penile cancer and reduce it by a factor of 2.3 you get 0.0001815, which would be the expected lifetime risk for penile cancer in circumcised men. The absolute risk reduction would be the difference between the two rates: 0.0004176 minus 0.0001815 or 0.0002360. The number needed to treat is the inverse (1/x) of the absolute risk reduction or 4237. This means that 4237 infant males would need to be circumcised in order to prevent one case of penile cancer, which usually strikes on average at 80 years of age. If, however, the relative risk reduction is 1.5, the number needed to treat is 7184.

Cost Effectiveness

So how much does it cost to prevent one case of penile cancer using infant circumcision? If it takes 7184 circumcisions to prevent one case of penile cancer and each circumcision costs an average of \$285 paid at the time of the procedure,⁶ the cost would be the product of these two numbers or \$2,047,440. But the story does not end there. The money for the circumcision was spent at the time the male was circumcised, but penile cancer usually does not develop until about 80 years of age. So, for 80 years the opportunity of having that cash spent at the time of the procedure has been lost. These opportunity costs add up over 80 years. For example, if that money were put out at 3% interest for 80 years, the opportunity costs would be \$21,786,584. If the money were to earn 5% interest for 80 years, the costs of preventing one case of penile cancer would be \$101,474,076. This may explain why the American Academy of Pediatrics Task Force elected not to do the calculations.ⁱⁱ

What Not to Do

There are a number of ways to play inappropriately with numbers. I'll give a couple of examples. One is to take an estimate from a select population and then apply to the population in general. For example, if there is a 19% rate of repeat urinary tract infection in intact boys, this does not translate into a one in 5.2 lifetime risk of urinary tract infections for all intact boys. This is the lifetime risk for boys to have a *repeat* urinary tract infection. Since only about 1% of intact boys ever get a urinary tract infection, the risk in the general population would be 0.01 times 0.19 or 1 in 526. It would appear that the estimate is only off by a factor of 100.

As noted earlier, mathematics can be used to make comparisons: a measure of the similarity or dissimilarities between two groups. We can compare averages, we can compare rates, and we can compare percentages. You cannot make a comparison when there is no comparison group. This is what Edgar Schoen and his colleagues tried to do in their study on circumcision and penile cancer. They published a *case series* of 213 men noted to have penile cancer. Of the men with invasive penile cancer, 2 were circumcised and 87 were not.^v Of those with carcinoma *in situ* 16 were circumcised and 102 were not. The study concluded the "relative risk for IPC[invasive penile cancer] for uncircumcised men to circumcised men is 22:1."⁹ A case series like this does not have a control group. Without a comparison group, it is impossible to make this claim. The number of cancers tallied by circumcision status would need to be compared to the

rates of circumcision of the general population for men the same age and ethnicity. Without a comparison group, there's no way of knowing or calculating the relative risk. In other words, you can't have a fraction without a denominator.

Schoen is also guilty of making false comparisons. He noted that during a 55-year span of history more than 50,000 men in United States would have been expected to be diagnosed with penile cancer. During that same span of history, there were only 10 case reports of penile cancer in circumcised men. Consequently, he estimated that the ratio of penile cancer in intact men to circumcised men was 5000 to 1.¹⁰ This is absurd. For it to be a true comparison, you would either need to compare only the cases of penile cancer in intact man that were reported in the medical literature to every case of penile cancer in circumcised men in the medical literature. Alternatively, he would need to have a method of determining the total number of circumcised men who developed penile cancer within the population during that span of history. Clearly not every case gets published in the medical literature, and there may be hundreds or thousands of cases that go unreported for each case that is reported. It is intellectually dishonest to compare a nearly complete tally in a population to one that is highly likely to be incomplete.

I mentioned absolute risk reduction earlier as the difference in the percentages between comparison groups calculated by subtraction. By comparison, the relative risk reduction is the ratio of the percentages in the comparison groups that has the outcome of interest. For example, if in the control group 2% have the outcome of interest over a year and only 1% in the treatment group does, then the relative risk reduction would be 1% divided by 2% or 50% reduction. The absolute risk reduction would be 2% minus 1% or 1%. The number needed to treat would be the inverse (1/x) of this number or 100.

When looking at the number of men who became HIV positive in the three randomized controlled trial in Africa (Figure 1), you might get the impression of a substantial difference between the treatment group and the control group.

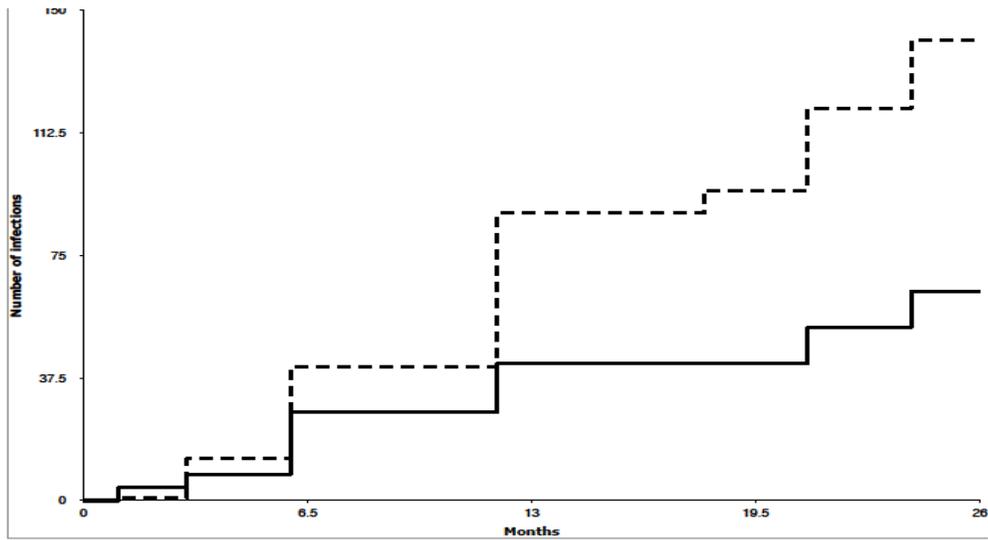


Figure 1 The cumulative number of men infected with HIV in the three African circumcision trials over time (treatment group solid line, control group dashed line.)^{21,23,24}

But putting their findings into a proper perspective and having a y-axis that goes from 0 to 100% (Figure 2) shows a very different story that highlights how small the absolute risk reduction was in these trials. The 60% *relative* risk reduction, which sounds like a huge difference, is actually a 1.3% *absolute* risk reduction. Most statisticians note researchers often use the relative risk reduction to bolster hyperbole. If the report of a study's finding only mentions the relative risk reduction and fails to mention the absolute risk reduction or the number needed to treat, they are probably trying to draw attention away from that fact that their findings may not be clinically important or relevant.

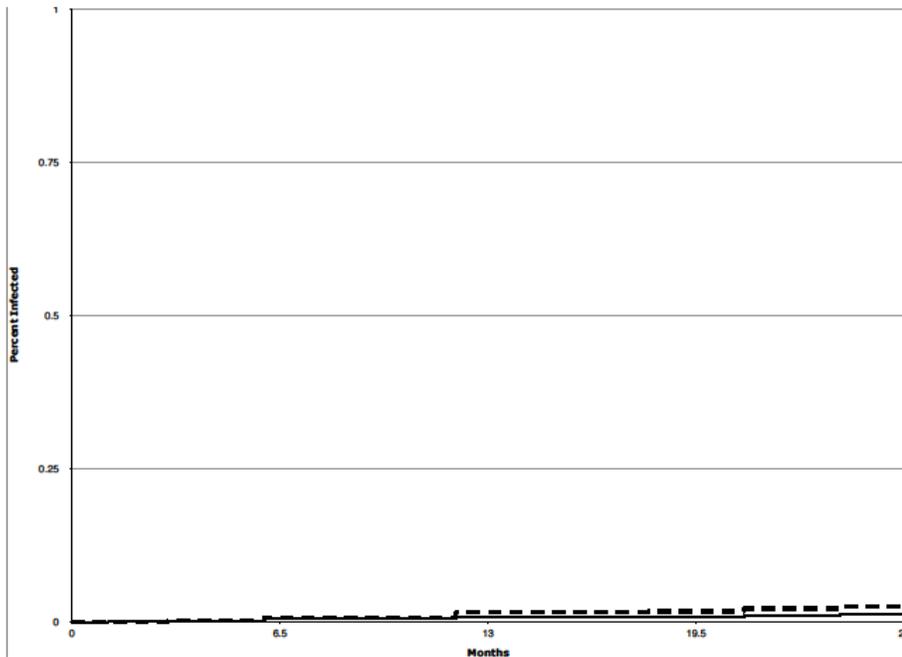


Figure 2 The cumulative percentage of men infected with HIV in the three African circumcision trials over time (treatment group solid line, control group dashed line.)^{21,23,24}

Number Needed to Harm

The flip side of the number needed to treat is the number needed to harm. It is calculated in a similar method. The percentages of those harmed in the two groups are subtracted from each other and the inverse (1/x) is the number needed to harm. For example, in the randomized clinical trial of circumcision for men infected with HIV, 18% of men in the treatment group had female sexual partners who became infected with HIV, while only 12% in the control group had a similar outcome.¹¹ The inverse of the difference is 17. So for every 17 men who were circumcised you would expect one additional female partner to become HIV infected that would have not become infected if the procedure had not taken place.

Keeping It Real

When using numbers to describe the world around us, we have to keep things real. For example, the other day a colleague of mine stated “95% of all statistics are made up.” I pointed out to him that he overplayed the statement. But employing “95%” he used a number that was a bit too extreme and made his statement less plausible. If instead he was to say 35%, or maybe even 55%, that would have been within the range in which the average listener would believe he was being truthful, thus making it plausible. Some

circumcision enthusiasts do not understand this. For example Morris and Wiswell have declared that the number needed to circumcise to prevent one urinary tract infection is 4.¹² This is a dramatic drop from previous previously published figures of 111 and 195.^{13,14} Such a statement quickly registers a high value on the male bovine fecal matter detector. To have a number needed to treat of four, the absolute risk reduction would need to be 25%. This would translate into a difference between a urinary tract infection rate of 75% and 50%, or a difference between 26% and 1%. Such a difference is implausible. In a similar fashion, Morris had repeatedly stated that the benefits of circumcision outnumber the risks by a factor of 100 to 1.¹⁵⁻¹⁸ Even if the complication rate of circumcision was as low as 3%, which may be low, based on what Morris has put forth, one would expect 300% of circumcised men to obtain a benefit. This is clearly absurd. If they want to be taken seriously when confabulating new “facts,” circumcision enthusiasts would do well to keep their “facts” within the realm of the plausible.

Relative Risk Ratios and Odds Ratio

Relative ratios and odds ratios can be calculated from numbers that appear in a 2 X 2 table (Figure3).

2 X 2 Table	Outcome Positive	Outcome Negative
Trait Positive	A	B
Trait Negative	C	D

Figure 3 The 2 X 2 table.

Relative risk ratios are the ratio of percentages.

$$(A \div (A + B)) \div (C \div (C + D))$$

They are reported primarily in prospective cohort studies and some representative population surveys. Conceptually a comparison of percentages makes more sense to us, but the mathematical properties of odds ratios, such as their use in logistic progression, make reporting odds ratio more common.

If the odds of having a disease in those with the trait of interest is A/B and the odds of those without the trait having the disease is C/D, the ratio is $(A/B) \div (C/D)$. With a little bit of manipulation the formula can be converted to:

$$(A \times D) \div (B \times C)$$

Some people refer to the odds ratio as the ratio of the cross products. As with any ratio, the normal value is one. With outcomes of low-frequency the relative risk ratio and the odds ratio will be similar. Often times ratios are reported with their 95% confidence interval, which means that there is 95% chance that the true value for the population is within this range. If the confidence interval includes 1 then the p-value is greater than 0.05.

P-values

What is the p-value? We use inference and sampling of the population to make a guess as to the true difference between two groups within the population we are interested in. We measure a variance to estimate how confident we are of our guess. The p-value is the probability that our estimate of a difference, given its variance, would be found if no difference actually existed between the two groups. A p-value of 0.05 (one in twenty) is usually used as the threshold for statistical significance. This originated with Ronald A. Fisher, founder of modern statistics, who arbitrarily decided that if the probability of getting the result by chance alone, *in repeated experiments*, was less than one in 20, then the results were probably significant. In recent years, p-values are being phased out and replaced by confidence intervals, which provide much more information. If a 95% confidence interval includes the value of one, then the p-value is >0.05 and the result is not statistically significant.

Fragility Index

One of the major shortcomings of the p-value is that we use it in an almost arbitrary way to determine whether the results of a study are worthy of publication. If the p-value is less than 0.05, it is often determined to be worthy of publication, but if it is not, often times the manuscript is either not submitted for publication or it is rejected during the review process. This creates two problems. One is publication bias, which I will discuss briefly later in this paper. The other is that weak, non-robust findings are propped up as “significant” merely because they reach the 0.05 threshold. Since the cut off of 0.05 is often just barely attained, there needs to be a measure of how close a study is to the brink of statistical significant. To address the robustness of published findings, the group at McMaster University in Hamilton Ontario have developed the “fragility index.”¹⁹ To determine the fragility index, one starts with the 2 x 2 table in which the numbers in each category with and without the disease are tabulated. Then incrementally a positive outcome in the one group is added while a positive outcome in the other group is

subtracted. The number of additions/subtractions needed to make the results of a study no longer statistically significant is the fragility index. In their survey of randomized controlled trials, most studies had a fragility index of eight or higher. A study in which the fragility index was three or less would indicate that the results were not very robust. Amazingly, in their survey of the medical literature they found several studies in which the results were reported to be statistically significant when they actually were not.ⁱⁱⁱ For example, in the Ugandan circumcision/HIV study,²¹ when those lost-to-follow-up are eliminated, the p-values are shown in Table 1.

Because such a small percentage of the men enrolled in this trial became HIV-infected the fragility index is quite low. Remarkable that with a swing of only a few positive outcomes the p-value goes from 0.0043 to 0.0638. Even more remarkable is that the study with results that are far from robust was stopped early because of the low p-value measured at the interim analysis. Perhaps the fragility index should be a part of early termination decisions.

Another example comes from a study by Warner et al. of the men coming to their STD clinic with known heterosexual HIV exposure.²² In this study, intact men were at significantly greater risk, while men from the general population were at equal risk regardless of circumcision status. In this select population the fragility index was 1. (35/309/11/39 p=.0302; 36/308/10/39 p=.0596).

Missing in Action

Another important part of any trial is how many people are lost to follow-up. In the African randomized clinical trials, for every new HIV infection there were between 3.48 and 7.37 subjects who were lost to follow-up.^{21,23,24} Since the outcomes of these men are unknown, it is impossible to determine whether the outcomes in these men would have swayed the data in the trials, especially considering how few men became HIV infected during the trials. There are several ways to look at how to accommodate those that were lost to follow up. One accepted approach of sensitivity analysis is to use the most extreme possible outcomes with those who are lost to follow-up. In the African randomized clinical trials, one could assume that all of those who were lost to follow up in the treatment group became infected while none of those lost to follow-up from the control group did. Doing so completely flips the results. In the South African trial, the relative risk would be 2.52, 1.81–3.51; in Uganda the relative risk will be 5.60, 3.14–7.70; and in Kenya the relative risk would be 3.13, 2.25–4.35 (risk of infection in treatment (circumcised) group versus control (intact group)). This indicates how vulnerable these trials were to the problem of missing data.

The problem with missing data and those who are lost to follow-up is that these men are usually different than those who completed the trial. For example, in the Kenyan trial the men who were lost to follow-up were more likely to have serological evidence of herpes and report more sexual partners at the beginning of the trial. (Robert Bailey, personal communication, Washington, DC, 2012) The other problem is that significantly more men randomized to the control group were lost to follow-up (OR=1.27, 95%CI=1.11-1.45). One can speculate as to why this occurred. My personal opinion is that the men

interested in participating in these trials were looking for an opportunity to obtain a free circumcision, which in these African countries would cost them what the average worker would earn over several months. A high percentage of the men in these trials were unemployed. Once randomized to waiting two years for a free circumcision, men in the control group may have seen little point in continuing to participate in the trial.

The high ratio of those lost to follow-up to those who became HIV-infected is an important wildcard in these trials that has not been adequately addressed by the researchers who performed the trials. In rating the quality of studies, typically randomized controlled trials are the gold standard; however, when there is a high rate of lost to follow-up the quality rating of the randomized trial falls below that seen in observational studies of moderate to good quality. This is one of several factors why the African randomized clinical trials should be considered to be of poor quality.

Misclassification Bias

Studies can also be affected by the misclassification bias in which subjects who are circumcised are identified as being intact and those that are intact are being identified as circumcised. For example, in 1985 researchers at the Centers for Disease Control and Prevention found that in Atlanta 15.7% of circumcisions were not recorded on the face sheet of the hospital chart. Data are gleaned from the face sheets and accumulated into databases. Some studies rely on these databases to identify which males are circumcised. This finding from Atlanta translates into the circumcision rates being higher than what the databases would indicate. The largest studies to look at a link between circumcision and urinary tract infections were performed by the United States Army using their databases to make these calculations. If 15.7% of circumcisions were not recorded on the face sheet of the charts, this would be consistent with 78.1% of infants identified as being intact by the database actually being circumcised. If this is the case, then the US Army data from 1974 through 1983 would have the odds ratio of UTIs in intact versus circumcised infants reduced from 10.27, (95%CI=8.69–12.15) to 4.13 (95%CI=3.34-5.11), and the data from 1984 through 1988 from 9.43 (95%CI=7.48-11.90) to 4.07 (95%CI=3.28-5.05), which would make these results consistent with what has been found in other studies that did not rely on databases. The reason that misclassification bias had such a large influence on the US Army data can be attributed to the small minority of boys who were left intact. When the same technique is applied to the Canadian study by To et al,¹⁴ the impact is much smaller because about half of the boys in that study were intact.

Misclassification can also occur when a circumcised man says that he is intact or an intact man states he is circumcised. One of the most outrageous examples of this is in a study from Mexico in which 95 men stated they were circumcised when only 8.3% were found to be circumcised on physical examination. The authors of the study published their findings using self-report as their method of determining circumcision status.^{25,26}

Meta-analysis

Meta-analysis is a process in which results from multiple studies are combined to determine a summary estimate. The calculations methods are fairly straightforward as

long as you don't use the method employed by Wiswell and Hachey.²⁷ The Wiswell method is to take the 2 x 2 tables from all the trials and tabulate all of the subjects in each category and then use these tabulations to calculate an odds ratio and confidence interval. For example, in Wiswell's meta-analysis of circumcision and UTIs published in 1993, his method resulted in a summary odds ratio of 12.0 (95%CI=10.6-13.6). If an appropriate calculation method (Mantel-Haenszel) is used, the random-effects summary odds ratio is 9.75 (95%CI=5.99-15.86). This is a much lower odds ratio with markedly wider 95% confidence interval. One problem with the Wiswell method is it can result in Simpsons paradox in which the opposite results can be obtained than would be obtained if proper calculation methods were used. To perform the calculations properly, the results need to be stratified by study and evidence of between-study heterogeneity factored into calculating the confidence interval.

The biggest problem with meta-analyses is that the results can be predetermined by the inclusion criteria, so the accuracy of the summary odds ratio is only as good as the studies that were included in the analysis.

Sometimes studies are missing because of publication bias. If the results of a study are not statistically significant it is less likely to be submitted for publication, and/or less likely to be published, and less likely to be included in a meta-analysis. Consequently, if all studies on a particular topic were published, we would expect as many small studies with large positive findings as small studies with large negative findings. A number of statistical methods have been developed to estimate whether there is such a balance of publications in the medical literature.²⁸ If only positive studies are published, this would skew the summary odds ratio to be higher than it should be. There have been some attempts to adjust for publication bias in estimating a summary odds ratio.^{28,29}

Meta-regression

Meta-regression takes meta-analysis one step further. It creates a regression model in which individual studies are treated as individual subjects within a study.³⁰ Basically, it creates a regression model of the logarithm of the odds ratio from individual studies based on characteristics that can be measured in each study. For example, if you have a number of trials using a medication to determine the impact it has on risk for heart attacks but the dose of the medication was not consistent from trial, a simple meta-analysis would determine the overall treatment effect for the medication. With meta-regression, a value is assigned to each study for the dose of medication given in that particular study. Consequently, if there were a large number of studies using a low dosage showing little or no effect, but a few studies using a higher dose showing a positive effect, meta-regressive would be able to demonstrate that the treatment effect was dose-dependent. Without the meta-regression, you may have reached the conclusion that the medication was not useful and missed that the impact was dose-dependent.

This technique has come in handy when evaluating studies that looked for an association between genital human papillomavirus infections and circumcision status. With meta-regression, it could be demonstrated that studies with either misclassification

or sampling bias reported odds ratios of genital HPV in intact men compared to circumcised men that were significantly greater than the odds ratios reported in studies that did not exhibit these forms of bias.^{31,32} As a consequence, rather than excluding these studies, they can be included in the analysis and the differential impact of the significant forms of bias can be adjusted for. While certain circumcision enthusiasts, who apparently do not understand this rather simple regression technique, have voiced their dissatisfaction because the calculations exposed the impact of their biased studies,³³ meta-regression is becoming a widely used tool in systematic reviews.

Bonferroni Adjustments

Whenever a test is performed, there is a likelihood of the test providing a falsely positive result. When we set the p-value at 0.05 as the threshold for statistical significance, we are accepting a one-in-20 likelihood that the positive finding will occur when there is actually no difference between the groups of interest. If you increase the number of independent tests performed and use $p=0.05$ as the rate of falsely positive results you were willing to accept, then the likelihood that one or more tests will produce a falsely positive test increases. For example, if you do 20 comparisons in which there is no actual difference, the likelihood that at least one test of the 20 by chance alone will be positive is 0.64. To determine the p-value to apply to each individual tests that would give a p-value of 0.05 for not having any falsely positive test can be calculated using a Bonferroni adjustment ($0.05/n$ where n is the number of tests performed).

This influential adjustment should only be used in limited circumstances as it was designed to apply to decision-making scenarios rather than for evidence evaluation. It should only be applied selectively when multiple hypotheses are being tested or multiple tests are being performed without a pre-established hypothesis.³⁴

Our friends have been applying this adjustment where it is not appropriate, and not applying it where it may be appropriate. For example, Waskett and Morris³⁵ suggest that Bonferroni adjustments should be applied to the data in the study of penile fine-touch receptors by Sorrells et al.³⁶ Waskett and Morris demonstrate their ignorance of statistical methods on several levels in their letter. First, they erroneously believed that the table in the study that presented the fine-touch threshold data from the 19 locations on the penis represented the results of the model chosen for the study. Then they erroneously assumed that we had made multiple separate comparisons of fine-touch thresholds between intact and circumcised men in determining the final result of the trial. What they failed to recognize is that the trial had a single hypothesis based on a pilot study that the fine-touch thresholds were lower (more sensitive) in the glans of the intact penis than the circumcised penis. The model, a marginal mixed model that allows for multiple measurements in individual participants and corrects for between-subject variability, was used to evaluate the pilot data and was chosen to be the model to analyze the data from study before data collection began. By using a marginal mixed model, only one statistical test was performed. With only one hypothesis being tested and only one test being performed, the n in the Bonferroni adjustment would be 1 and the threshold for statistical significance would be 0.05.^{iv}

If one were to follow Waskett and Morris's lead, then the Bonferroni adjustment should have been applied to all of the observational studies of circumcision and HIV infection in which data on other risk factors were collected. The p-value should have been reduced for each of the factors, and very few of these studies could report a statistically significant finding when it came to circumcision. The take-home message is that the Bonferroni adjustment should be used sparingly, and not only when someone does not like the results of a study.

It does apply when a study has divided its population into multiple substrata without an *a priori* hypothesis of how the primary association would differ between subgroups in an effort to prevent researchers from statum shopping. An example of this occurred with Castellsagué et al.³⁷ in which they reported their results for monogamous women whose male sexual partners has six or more lifetime sexual partners. Since there are four different strata, the acceptable threshold p-value would be 0.0125.

Probability and Conspiracy Theories

Sansom et al. noted that the findings in the African randomized clinical trials were "remarkably" consistent.³⁸ The question is whether the findings were *too* consistent. The number of intact men in each of the trials was within and two of each other while the number of infected circumcised men for each of the three trials was within a couple of each other. What is the likelihood of such tight clustering of results happening by chance alone?

The three trials in which approximately 70 men became infected with HIV is similar to flipping a coin 70 times counting the number of times it came up heads, doing 70 more coin flips, tallying the number of heads, and then doing 70 more coin flips. To answer the question, we need to calculate the probability of flipping a coin 70 times in three separate trials and have the tally in each of the three trials being either 34, 35, or 36. While it is possible to flip 70 heads in a trial of 70 flips, it is very unlikely. By the same token, the most likely outcome would be flipping heads exactly 35 times given a 50-50 probability of each flip coming up heads. But other possibilities exist. If you do the math, the probability of flipping exactly 35 heads is 9.503%. The odds of flipping exactly 34 heads is 9.239% as is the probability of flipping exactly 36 heads. The probability of flipping exactly either 34, 35, or 36 heads would be the sum of these three probabilities or 27.981%. That is the probability of coming up with either 34, 35, or 36 heads in only one trial. The question is what is the likelihood of doing it in three trials in a row? To calculate this you need to have the probability of each of the trials multiplied by each other. In this example, you would take 27.981% times 27.91% times 27.981%, which would yield a probability of 2.191%. So it is quite unlikely to have three trials in a row with such tightly clustered results. In applying this to the results of the randomized trials, in which the ratio was 2 to 1, all of the possible clusters of that tightness, not just the most likely cluster, would need to be included. When all the possible clusters of having the number of men who were turned positive for HIV in the study be within two of each other are included, the probability of having this level of "consistency" in the results of three trials of the approximate 3.3%.

Such an unlikely tight cluster of results is enough to get the tail and the tongues of the conspiracy theorists wagging. I am not saying that the results of the randomized clinical trials were contrived to produce nearly identical results, but probability theory tells us that such consistent results from three trials that were supposedly unrelated is quite improbable.

Table 1: Calculation of the Fragility Index for the Ugandan Male Circumcision Trial²¹

FI*	Intact+ve	Intact-ve	Cd+ve	Cd-ve	p-value
0	45	1933	22	2010	.0043
1	44	1934	23	2009	.0092
2	43	1935	24	2008	.0186
3	42	1936	25	2007	.0355
4	41	1937	26	2006	.0638

FI=Fragility Index

Notes

ⁱ The actual formula is $LR=1-(\exp(-\text{incidence} \times \text{years}))$ which for illnesses with a low incidence gives a nearly identical result.

ⁱⁱ When making a plea to justify circumcision as a preventive for prostate cancer, Morris, Waskett, and Bailis conveniently failed to consider the opportunity costs in their calculations.^{7,8}

ⁱⁱⁱ In the cervical cancer and circumcision literature for example the study by Agarwal et al. reported a statistically significant finding when in fact it wasn't. The two-sided Fisher p-test yield $p=0.0733$, while the study reported $p=0.045$.²⁰

^{iv} Far from being an "erroneous statistical analysis," as Waskett and Morris claim, the marginal mixed model is a powerful statistical model when multiple or repeated measures are made on individual participants. Their lack of statistical expertise undermines any credibility for their wild, uninformed claims.

References

- ¹ Task Force on Circumcision. Male circumcision. *Pediatrics* 2012; 130: e756-85.
- ² Chin J. *The AIDS pandemic: the collision of epidemiology with political correctness*. Oxford: Radcliffe Publishing; 2007.
- ³ Christakis DA, Harvey E, Zerr DM, Feudtner C, Wright JA, Connell FA. A trade-off analysis of routine newborn circumcision. *Pediatrics* 2000; 105: 246-9.
- ⁴ Kochen M, McCurdy S. Circumcision and the risk of cancer of the penis. A life-table analysis. *Am J Dis Child* 1980; 134: 484-6.
- ⁵ Learman LA. Neonatal circumcision: a dispassionate analysis. *Clin Obstet Gynecol* 1999; 42: 849-59.
- ⁶ Hart-Cooper GD, Tao G, Stock JA, Hoover KW. Circumcision of privately insured males aged 0 to 18 years in the United States. *Pediatrics* 2014; 134: 950-6.
- ⁷ Morris BJ, Waskett J, Bailis SA. Case number and the financial impact of circumcision in reducing prostate cancer. *BJU Int* 2007; 100: 5-6

- ⁸ Van Howe RS. Case number and financial impact of circumcision in reducing prostate cancer [letter]. *BJU Int* 2007; 100: 1193-4.
- ⁹ Schoen EJ, Oehrli M, Colby CJ, Machin G. The highly protective effect of newborn circumcision against invasive penile cancer. *Pediatrics* 2000; 105(3): e36.
- ¹⁰ Schoen EJ. The relationship between circumcision and cancer of the penis. *CA Cancer J Clin* 1991; 41: 306-9.
- ¹¹ Wawer MJ, Makumbi K, Kigozi G, Serwadda D, Watya S, Nalugoda F, et al. Circumcision in HIV-infected men and its effect on HIV transmission to female partners in Rakai, Uganda: a randomised controlled trial. *Lancet* 2009; 374: 229-37.
- ¹² Morris BJ, Wiswell TE. Circumcision and lifetime risk of urinary tract infection: a systematic review and meta-analysis. *J Urol* 2013; 189: 2118-24.
- ¹³ Singh-Grewal D, Macdessi J, Craig J. Circumcision for the prevention of urinary tract infection in boys: A systematic review of randomized trials and observational studies. *Arch Dis Child* 2005; 90: 838-8.
- ¹⁴ To T, Agha M, Dick PT, Feldman W. Cohort study on circumcision of newborn boys and subsequent risk of urinary-tract infection. *Lancet* 1998; 352: 1813-6.
- ¹⁵ Morris BJ, Bailey RC, Klausner JD, Leibowitz A, Wamai RG, Waskett JH, et al. Review: A critical evaluation of arguments opposing male circumcision for HIV prevention in developed countries. *AIDS Care* 2012; 24: 1565-75.
- ¹⁶ Morris BJ, Bailis SA, Castellsague X, Wiswell TE, Halperin DT. RACP's policy statement on male circumcision is ill-conceived. *Aust N Z J Public Health* 2006; 30: 16-22.
- ¹⁷ Morris B. Why circumcision is a biomedical imperative for the 21(st) century. *Bioessays* 2007; 29: 1147-58.
- ¹⁸ Morris BJ, Wodak AD, Mindel A, Schreiber L, Duggan KA, Dilley A, et al. Infant male circumcision: an evidence-based policy statement. *Open J Prev Med* 2012; 2(1): 79-92.
- ¹⁹ Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. *J Clin Epidemiol* 2014; 67: 622-8.
- ²⁰ Agarwal SS, Sehgal A, Sardana S, Kumar A, Luthra UK. Role of male behavior in cervical carcinogenesis among women with one lifetime sexual partner. *Cancer* 1993; 72: 1666-9.
- ²¹ Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, et al. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. *Lancet* 2007; 369: 657-66.
- ²² Warner L, Ghanem KG, Newman DR, Macaluso M, Sullivan PS, Erbelding EJ. Male circumcision and risk of HIV infection among heterosexual African American men attending Baltimore sexually transmitted disease clinics. *J Infect Dis* 2009; 199: 59-65.
- ²³ Bailey RC, Moses S, Parker CB, Agot K, Krieger JN, Williams CFM, et al. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* 2007; 369: 643-56.
- ²⁴ Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: The ANRS 1265 Trial. *PLoS Med* 2005; 2(11): e298.

- ²⁵ Lajous M, Mueller N, Cruz-Valdez A, Aguilar LV, Franceschi S, Hernandez-Avila M, et al. Determinants of prevalence, acquisition, and persistence of human papillomavirus in healthy Mexican military men. *Cancer Epidemiol Biomarkers Prev* 2005; 14: 1710-6.
- ²⁶ Van Howe RS, Cold CJ. Human papillomavirus link to circumcision is misleading. *Cancer Epidemiol Biomarkers Prev* 2006; 15: 405.
- ²⁷ Wiswell TE, Hachey WE. Urinary tract infections and the uncircumcised state: an update. *Clin Pediatr Phila* 1993; 32: 130-4.
- ²⁸ van Assen MALM, van Aert RCM, Wicherts JM. Meta-analysis using effect size distributions of only statistical significant studies. *Psychol Methods* 2014; epub ahead of print.
- ²⁹ Duvall S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56: 455-63.
- ³⁰ van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002; 21: 589-624.
- ³¹ Van Howe RS. Human papillomavirus and circumcision: A meta-analysis. *J Infect* 2007; 54: 490-6.
- ³² Van Howe RS. Sexually transmitted infections and male circumcision: a systematic review and meta-analysis. *ISRN Urol* 2013: 109846.
- ³³ Castellsagué X, Albero G, Clèries R, Bosch FX. HPV and circumcision: a biased, inaccurate and misleading meta-analysis. *J Infect* 2007; 55: 91-3.
- ³⁴ Perneger TV. What's wrong with Bonferroni adjustments. *Br Med J* 1998; 316: 1236-8.
- ³⁵ Waskett JH, Morris BJ. Fine-touch pressure thresholds in the adult penis [letter]. *BJU Int* 2007; 99: 1551-2.
- ³⁶ Sorrells ML, Snyder JL, Reiss MD, Eden C, Milos MF, Wilcox N, et al. Fine-touch pressure thresholds in the adult penis. *BJU Int* 2007; 99: 864-9.
- ³⁷ Castellsagué X, Bosch FX, Muñoz N, Meijer CJLM, Shah KV, de Sanjosé S, et al. Male circumcision, penile human papillomavirus infection, and cervical cancer in female partners. *N Engl J Med* 2002; 346: 1105-12.
- ³⁸ Sansom SL, Prabhu VS, Hutchinson AB, An Q, Hall I, Shrestha RK, et al. Cost-effectiveness of newborn circumcision in reducing lifetime HIV risk among U.S. males. *PLoS One* 2010; 5(1): e8723.

Robert S. Van Howe, MD, MS, FAAP, is Professor and Interim Chairman of Pediatrics at the Central Michigan University College of Medicine. He has lectured and been published internationally on the topic of circumcision and has been a consultant to the American Academy of Pediatrics, the Centers for Disease Control and Prevention, and the World Health Organization. Much of his research has focused on secondary research, including meta-analysis and cost-utility analysis. Saginaw, Michigan, USA.