Hypothesis Testing and Avoiding False-Positive Conclusions

Philippe P. Hujoel
Department of Dental Public Health Sciences, School of Dentistry BOX 357475
University of Washington, Seattle, WA 98195.
SHORT TITLE: False-positive conclusions.
In part supported by NIDCR #U01DE016752
Introduction

Hormone replacement therapy was suggested to reduce cardiovascular disease risk\(^1\). Several vitamins were suggested to prevent cancer\(^2\). Vaccination was suggested to lower HIV infection risk\(^5,6\). These positive associations were derived from animal experiments, ecological studies, observational studies, and even randomized controlled trials (RCTs). All these associations were ultimately proven to be false-positives by means of pivotal RCTs. Publicity around these false-positives findings had nefarious consequences. From a public health perspective, scientific publications and press reports on the false-positive findings influenced physicians and laymen alike and ultimately led to hundreds of thousands of premature deaths. From a research perspective, the false-positives findings led to a loss of billions of dollars in pivotal randomized trials that should not have been initiated, and often multiple decades of lost time in the search for cures. It may not be surprising that controlling the rate of false-positives has been considered an important issue in clinical trial design for decades\(^7,8\).

The saga of alpha-tocopherol (a form of vitamin E) demonstrates how even large-scale randomized controlled trials (RCTs) can generate false-positive findings. One of the promising hypotheses of the 20th century was that alpha-tocopherol could reduce lung cancer mortality. The results of the pivotal RCTs on this hypothesis were disappointing; alpha-tocopherol increased overall mortality risk\(^9\), increased the risks for serious adverse events such as heart failure\(^10\), and did not impact lung cancer mortality rates\(^2\). Nonetheless, secondary data analyses on two of these randomized trials - data explorations without pre-specified hypotheses – led to the observation that alpha-tocopherol was associated with a decreased risk for prostate cancer\(^2,11\). These positive results were considered sufficiently promising to initiate “the largest individually randomized cancer prevention trial ever conducted”, the SELECT trial\(^12\).

The results of the SELECT trial\(^13\) may well be the ultimate confirmation of Yusuf’s 1991 report: positive findings from post-hoc generated hypotheses in RCTs are unreliable; the equivalent of betting on the horse after the race is over\(^14\). If anything, the results of the SELECT trial indicated that alpha-tocopherol increased rather than decreased prostate mortality risk\(^13\).

The excess of false-positive results in the literature is in part caused by fitting causal explanations to observed chance associations. There is a difference between testing a hypothesis against data that will be collected versus the manufacturing of a hypothesis tailored to the data that have been analyzed. Formulating a specific hypothesis prior to data collection is a quintessential feature of scientific experimentation. In medical and dental research, it provides investigators with the opportunity to predict how patients with pre-defined characteristics will respond to two or more interventions with respect to a pre-specified clinical outcome. These predictions can then be compared to the observations collected during the clinical experiment, and lead to reliable inference. Fitting a causal explanation post hoc is a treacherous exercise, as a large number of hypotheses can be consistent with an observed association, only one of which is accurate.

False-positive findings might be plaguing dental research just as much as medical research. Positive findings in randomized controlled trials suggested that chlorhexidine would lower caries risk\(^15\), that periodontal therapy would prevent low birth weight\(^16\), or that anti-inflammatory would control periodontitis\(^17\). These findings came into question when larger, better-controlled randomized trials were conducted\(^18,19\). This chapter will review why the machinery of statistical hypothesis testing can
lead to a preponderance of false-positive results and provide some approaches to reduce the incidence of false-positives.
How can hypothesis testing lead to an overabundance of false-positive leads?

A false-positive conclusion is defined here as a claim that an experimental intervention is superior to a control intervention, while in truth the interventions have the same effectiveness, or more insidiously, that the control is superior to the experimental. Two common reasons for false-positive are: (1) changing the pre-trial hypothesis during or after the conduct of the trial, and (2) cherishing positive results more than the negative results. The price paid for ignoring negative results and fishing expeditions to find positive results may be financially and professionally lucrative in the short-term, but have adverse long-term consequences; a long road towards an expensive definitive trial whose ultimate results fail to confirm the early positive findings and the imprinting of false health beliefs among medical and lay alike.

The publication biases towards positive findings – at the expense of negative findings -have been extensively reported on. There may be a continuing bias of journal editors and reviewers, commercial companies, and investigators against publishing negative results. Publishing exciting small research findings may generate future commercial and government funding, publicity and the attention of reporters, academic promotion, and increases in stock prices. In contrast, there are typically few incentives and several disincentives to publish negative results. The only time negative results appear of commercial interest is in equivalence trials where marketing advantages may exist.

There are many medical examples where pivotal trials go unpublished because of pressure by commercial interests or editor bias. The powerful biases against publishing negative results may be present in dentistry as much as in medicine. A negative finding in a large government-funded trial on oral hygiene and dental caries was rejected for publication in a premier dental journal. A negative finding on chlorhexidine rinsing and tooth loss was rejected in a top-ranked dental journal even though the study was one of the first government sponsored studies to have an independent data and safety monitoring board, had a sample size on over 1100 patients with an average of 4.4 years of follow-up, and that had multiple safeguards against generating false-positive finding. Yet, this same journal publishes positive findings from exploratory studies.

Given the all around unpalatable nature of negative results, it should not be surprising that many approaches have been devised to generate positive results. One approach to generate positive conclusions is to explore various patient subgroups, various endpoints, and various statistical techniques. For instance, analyses can explore the effects of treatment in patients with more severe dental disease, or dental patients with good oral hygiene, or other subgroups that lead to the desired results. Or, if the investigated treatments do not differ on the a priori defined endpoint, one can explore different endpoint definitions or switch to a different endpoint altogether. For instance, pre-trial it may have been decided to consider a pocket reduction of >2 mm as a treatment success. This definition of success can be changed post-trial to a pocket depth reduction of >1 mm, a pocket depth reduction of >3 mm, or if neither leads to the desired result, to an evaluation of various cut-offs on probing attachment levels, radiographic bone levels, bleeding, or combinations of these. The possibilities to generate false-positives are almost infinite.

Statistical techniques themselves can be bent to provide positive results. Different statistical techniques can be explored with respect to their impact on p-values; parametric or non-parametric, one-sided versus two-sided tests, with and without adjustment for baseline covariates, site-based or patient-based, and if site—based analyses are selected, various correlation structures can be assessed. Such
cherry-picking of statistical analyses can often switch marginally significant results to the magical $p < 0.05$ conclusion.

Actual flaws or deceptions—intended or not—in the design or data analysis\(^7\) can similarly lead to false-positive findings. Controls groups can be contrived, randomization can be tampered with\(^{24}\), and inappropriate statistical methods can be employed to try to squeeze significance out of a data set. Contrived control groups typically revolve around tampering with the effectiveness of the control group. For instance, instead of using an experienced surgeon for the control intervention, the control procedure could be performed by a lesser qualified person. Or, instead of using the standard dose of a pain killer, or an antibiotic, the dose could be halved thereby increasing the likelihood that the experimental treatment will be effective. Several articles have reviewed the use of inappropriate statistical methods in dental research have noted the continued high prevalence of such practices\(^{25}\). Motives for using inappropriate statistical technique can be as simple as picking the method that provides statistical significance, regardless of appropriateness.

When the above approaches are mixed and matched the possibilities to generate false-positive findings are endless. Reducing false-positive findings can be achieved in part by the a priori specification of nine elements: (1) the subject characteristics, (2) the experimental treatment, (3) the control treatment, (4) the primary outcome by which the effectiveness treatment will be judged, (5) the magnitude of the difference between treatments considered clinically significant ($\Delta$), (6) the type I error rate ($\alpha$), (7) the type II error rate ($\beta$), (8) the likelihood of finding a positive association ($\pi$), and (9) the analytic method. Any post hoc deviation from these nine elements increases the risk for false-positive conclusions.
Quantifying the false-positive rates; the theory

The type I error rate ($\alpha$), the type II error rate ($\beta$), and the "a priori" likelihood that the null hypothesis is false ($\pi$) determine the false-positive error rates.

The type I error rate ($\alpha$) is the probability of rejecting the null hypothesis, when it is true. Note that $\alpha$ is a probability which is conditioned on the information that the null hypothesis is true. Thus, for instance, a type I error rate of 5% indicates that, if the null hypothesis is true, the chance for rejecting the null hypothesis is less than 5%. The type I error rate is a conditional probability. One does not know when conducting an experiment whether the null hypothesis is true or false. The experiment is designed to modify our belief that the null hypothesis is true.

The type II error rate ($\beta$) is the probability of accepting the null hypothesis, when it is false. The type II error rate is also a conditional probability. The type II error rate is conditioned on the knowledge that the null hypothesis is false and that there exists a given difference between the experimental and the control group. For instance, a type II error rate of 20% indicates that if the null hypothesis is false and if there exists a given difference between the investigated treatments, there is less than 20% chance to accept the null hypothesis. The power of an experiment is one minus the type II error rate (1 - $\beta$).

The "a priori" likelihood that the null hypothesis is false ($\pi$) is the third factor that determines the likelihood of false-positive and false-negative conclusions. In order to provide an intuitive sense on $\pi$, two extreme examples may provide a useful illustration. First, assume you are the all-supreme ruler of knowledge and that you know of a type of dental or medical disease for which no effective treatment exists. For instance, maybe there is a type of leukemia which leads to bleeding gingival tissues for which no type of locally administered periodontal treatment is effective. All treatments, when compared to the control treatment, are ineffective against this type of leukemia-induced periodontal disease. Under this scenario, $\pi$ is 0. We can do a million randomized trials, all trials will lead to the same conclusion; the experimental treatment does not work. In such a situation, all positive-conclusions reached in clinical trials will be false-positives. Regardless of whether the type I error rate is 1% or 5%, and whether the type II error rate is 80% or 99%, the rate of false-positive conclusions will be 100%.

A second extreme example is where all experimental treatments, when compared to the control treatment, are effective. The local mechanical treatments, the drug treatments, the combination of local and drugs treatments are all effective at dealing with a particular dental disease, when compared to no treatment. Under this circumstance, $\pi$ equals 1. All treatments, when compared to the control will be truly effective. Under this situation there will be no false-positive conclusions and all negative conclusions are false-negatives. All treatments identified as effective will truly be effective. Regardless of whether the type I error rate is 1% or 5%, or whether the type II error rate is 80% or 99%, the false-positive error rate will be 0%.

In real-life, $\pi$ falls somewhere between these two extremes. The history of clinical trial results in the different chronic disease areas provides some estimate of the likelihood that the experimental treatment is effective. In general, statisticians with experience in randomized controlled trials for chronic human diseases (e.g., cancer, AIDS, rheumatoid arthritis) generally consider $\pi$ to be small. For
cancer trials, the *a priori* probability that a novel treatment is better than a standard treatment (π) has been reported to be in the range of 0.1% to 20%.

In dental research, it would be useful to have some discussion on the range of plausible π values for the family of dental interventions. The history of randomized controlled trial results on a particular treatment can provide an estimate of π. A first small trial may be suggestive that the experimental treatment works. Our estimate of π may go up (π>0). We do more clinical trials, and if all clinical trial results remain positive our estimate of π may keep increasing. Translating our degree of belief in an actual value of π is subjective, but required to estimate the likelihood of obtaining false-positive conclusions.

This estimate of π can affect whether the conduct of randomized controlled trials is considered ethically feasible. For a non-fatal disease where effective treatment(s) exists, a clinician may be uncomfortable to randomize patients unless there is a 50-50 chance that the new experimental treatment is as effective as the standard. In this situation, an assumption of π = 0.5 may be desired to convince an ethical review board that a randomized trial is appropriate.

On the other hand, for a fatal disease a clinician may be comfortable to randomize patients to a novel experimental treatment even when he believes the experimental treatment has a less than 1% chance of being effective. The more desperate the clinical situation becomes, the larger the willingness of both patient and clinician to try anything (even treatments where π is close to zero).

The false-positive and false-negative error rates depend on the type I error rate (α), the type II error rate (β), and the *a priori* likelihood that experimental treatment is better than control treatment (π). A 2×2 table illustrates these relationships for *n* experimental treatments (Table 1). Among the *n* experimental treatments, π *x* *n* treatments will be truly effective and n *x* (1-π) treatments will be truly non-effective. For the π *x* *n* truly effective treatments, a certain proportion will be incorrectly classified as non-effective. This proportion is determined by the power of the experiment. If the power of the experiment is 80% (type II error rate of 20%), 80% of the π *x* *n* treatments will be correctly classified as being effective, and 20% of the π *x* *n* treatments will be incorrectly classified as non-effective.

Analogously, there will be n *x* (1-π) experimental treatments that are truly non-effective. The proportion of these that will be correctly classified depends on α (the type I error rate). Of the n *x* (1-π) non-effective treatments, n *x* (1-π) *x* 0.05 will be incorrectly classified as effective if the type I error rate is 5%, and n *x* (1-π) *x* 0.95 will be correctly classified as non-effective (Table 1).

Table 1: The relationship between α (say 0.05), β and π and the false-positive and false-negative error rates.

<table>
<thead>
<tr>
<th>p-value</th>
<th>True Relationship between experimental and control treatment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 0.05</td>
<td>π <em>n</em>(1-β) + n α (1-π)</td>
<td>π <em>n</em>(1-β) + n α (1-π)</td>
</tr>
<tr>
<td>&gt; 0.05</td>
<td>π <em>n</em>β + n (1-π) <em>α</em></td>
<td>π <em>n</em>β + n (1-π) (1-α)</td>
</tr>
</tbody>
</table>

- * Correct conclusion on the effectiveness/non-effectiveness of the treatment
- & False-positive conclusion
- $ False-negative conclusion
Based on the type I error rate (α), the type II error rate (β), and the a priori likelihood that experimental treatment is better than the control treatment (π), the false-positive probability equals \[
\frac{(1-\pi)\alpha}{1-\pi} \frac{\alpha}{1-\pi + \pi(1-\beta)}
\] (1)
and the false-negative probability equals \[
\frac{\pi\beta}{(1-\pi)(1-\pi + \pi(1-\beta))}
\] \(2\), 28.

A common problem is to misinterpret a p-value less than 0.05 as a reflection that the chance of a false-positive conclusion is less than 5%. This is not the case. Under most circumstances, the likelihood for false-positive conclusions is higher than the type I error rate of 5%. The lower the probability that treatment differences exist (i.e., the smaller \(\pi\)), the more challenging it is to find effective treatment for the chronic disease under investigation, and the higher the likelihood for false-positive conclusions. Before proceeding with suggestions on how to decrease the false-positive rates in dental research, an example of the false-positive rate for a typical small periodontal trial may be useful.

Quantifying false-positive rates; a periodontal example
In most periodontal trials, the classical significance level (α) of 0.05 is selected as the type I error rate. The typical exploratory randomized periodontal trial in the not so distant past had 15 subjects \(29,30\), have variability can be abstracted from the published periodontal literature (Table 2), and considered 0.2 mm pocket depth difference clinically significant\(21\). For the plaque index, a 15% difference (mean plaque index=1.0) was considered clinically significant. For bleeding on probing, a 15% difference (30% reduction in bleeding on probing) was considered clinically significant. For simplicity’s sake, a parallel arm design is assumed (Table 2).

Table 2: Typical power (1- β) of exploratory trials for studies with a sample size of 15 subjects (30 patients total) lasting 6 months or more

<table>
<thead>
<tr>
<th>Endpoints</th>
<th>SD</th>
<th>(\Delta)</th>
<th>Power 1-β</th>
<th>(\Pi=0.1%)</th>
<th>(\Pi=1%)</th>
<th>(\Pi=10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean pocket depth 5-6 mm</td>
<td>0.6</td>
<td>0.2 mm</td>
<td>15%</td>
<td>&gt;99%</td>
<td>97%</td>
<td>75%</td>
</tr>
<tr>
<td>Mean Pocket depth ≥ 7mm</td>
<td>1.0</td>
<td>0.2 mm</td>
<td>9%</td>
<td>&gt;99%</td>
<td>98%</td>
<td>83%</td>
</tr>
<tr>
<td>Mean Probing Attachment Level 5-6 mm</td>
<td>0.7</td>
<td>0.2 mm</td>
<td>12%</td>
<td>&gt;99%</td>
<td>98%</td>
<td>79%</td>
</tr>
<tr>
<td>Mean Probing Attachment Level ≥ 7mm</td>
<td>1.1</td>
<td>0.2 mm</td>
<td>8%</td>
<td>&gt;99%</td>
<td>98%</td>
<td>85%</td>
</tr>
<tr>
<td>Mean Recession 5-6 mm</td>
<td>0.7</td>
<td>0.2 mm</td>
<td>12%</td>
<td>&gt;99%</td>
<td>98%</td>
<td>79%</td>
</tr>
<tr>
<td>Mean Recession ≥ 7mm</td>
<td>0.8</td>
<td>0.5 mm</td>
<td>11%</td>
<td>&gt;99%</td>
<td>98%</td>
<td>80%</td>
</tr>
<tr>
<td>Sillness and Loe Plaque Index</td>
<td>0.62</td>
<td>15%</td>
<td>11%</td>
<td>&gt;99%</td>
<td>98%</td>
<td>80%</td>
</tr>
<tr>
<td>Bleeding on Probing</td>
<td>0.29</td>
<td>15%</td>
<td>28%</td>
<td>&gt;99%</td>
<td>98%</td>
<td>62%</td>
</tr>
</tbody>
</table>

\(\Delta\) = clinically significant difference

Based on \(\alpha\) and \(\beta\), one can calculate the false-positive error rate for a small periodontal trial (formula 1). For \(\pi = 0.1\%\), a larger than 99% false-positive error rate is present. When \(\pi\) is raised from 0.1% to 1%, 2-3% of the significant treatment effects reported reflect truly effective treatments, and 97% of the reported significant results are false-positives. When \(\pi\) is 10%, the false-positive rate varies between 62% and 83%.

These false-positive error rates are valid if there is one a priori defined hypothesis. The number of reported hypotheses tested in randomized periodontal trials is typically 15 30. The number of unreported
statistical tests is probably higher. Such multiple testing impacts false-positive error rates. For 15 independent tests, the probability of making at least one type I error is not 5% but 54%, the type II experimentwise error rate is not 20% but close to 0%. As the number of statistical tests increases without controlling for the experiment wise error rate, the type I error rate (α) approaches one, and the power approaches 1. In the limiting case of α=1 and β=0, the false-positive rate equals 1 - π, the a priori probability of no difference between experimental and control treatments. Thus, if the probability that the novel treatment is more effective than the standard is 5%, then 95% of all claims of significance will be false-positive claims.
Minimizing false-positive conclusions; π, registries, surrogates, randomization

The goal in clinical trials is to minimize the risk of false-positive conclusions. In order to achieve this goal, there should be one a priori defined hypothesis. The situation where there is more than one a priori hypothesis raises the complexity substantially and will not be discussed here. The following steps can reduce the likelihood of false-positive conclusions in the published literature.

Reporting π and its impact on α and β
Specifying the likelihood that a dental treatment is effective (π) is useful in terms of determining α and β and also in directing research funding. This process is currently an area every effort is done “to prove” that π is high, and any effort at refutation may be considered counterproductive in the sense that it reduces the changes of funding for a clinical trial. Ranges of priors could be determined for the likelihood that anti-infected approaches reduce caries risk (π < 1%), that antibiotics reduce tooth loss in patients with periodontitis (π < 5%), or that regenerative periodontal treatments reduce tooth loss.

In the beginning, controversy may exist on values of π but as evidence becomes available, controversy may decrease. For instance, why give a π < 1% for anti-infected caries treatments? Despite labeling caries as an infectious disease for decades and a multitude of clinical trials on anti-infected approaches, none have provided unequivocal evidence of effectiveness. If the estimate of π keeps decreasing with each additional randomized controlled trial, it has two consequences. First, upcoming pivotal trials on anti-infected approaches for dental caries need to specify a very small α and β to decrease the large risk for false-positives (given that π < 1%). Second, as π keeps decreasing, research funding may increasingly focus on non anti-infected approaches such as sealants and fluorides where the likelihood for identifying effective treatments (i.e., π) is larger. Similarly, that rationale for suggesting that antibiotics are unlikely to decrease tooth loss in periodontitis patients is based on the evidence that in cohort studies antibiotics appear to increase tooth loss, and that systematic reviews of the short-term clinical trials are inconsistent and report at best small changes in surrogate endpoints. Or, finally, take the example of regenerative periodontal products. If a truly effective treatment for periodontal regeneration existed there would not be such a therapeutic diversity and the half-life of the regenerative products that reached the market place would be longer. For regenerative periodontal trials π may be less than 1% suggesting that α and β should be specified as small as possible to avoid the false-positive findings.

Prevalent diseases with low morbidity/mortality require trials with small α and β
The prevalence and the morbidity and mortality of the disease determines the level of concern for false-positive findings and side-effects. For a rare and fatal disease such as Creutzfeldt-Jakob disease for which no effective treatment exists there is minimal concern regarding side-effects of promising treatments. Both common and serious side-effects and a high chance for false-positives are acceptable given that the alternative to treatment is certain death. On the other hand, for a widely prevalent benign disease such as the common cold there is almost no tolerance for side-effects, even if they are extremely rare. Even a side-effect as rare as Reye’s syndrome with an incidence of less than 1.1 per million can become unacceptable, even when treatment is effective.

Since dental conditions such as gingivitis are – in terms of clinical significance – closer to a common cold, than Creutzfeldt-Jakob disease, there should be more certainty regarding the safety and effectiveness of the treatments. Dental treatments such as tooth pastes or fillings are used almost worldwide,
sometimes resulting in almost life-long exposures. As a result, ineffective products (e.g., products on the market as a result of false-positive findings) and rare side-effects can have a substantial adverse impact in terms of wasted health care resources and common adverse events.

In dental research, the selection of \( \alpha \) and \( \beta \) should be guided by the anticipated prevalence of the population exposure to the investigated treatment (or market penetration) and the seriousness of the condition under investigation. A clinical trial on a novel toothpaste should specify much a smaller \( \alpha \) and \( \beta \) (e.g., \( \alpha=0.001 \) and \( \beta = 0.05 \)) than a clinical trial on a novel treatment for pemphigoid (e.g., \( \alpha=0.05 \) and \( \beta = 0.20 \)).

**Proper Randomization and intent-to-treat**

Randomization is a delicate process that can easily be tampered with, and allegedly was in medical trials where clinicians have been reported trying to circumvent the process of randomization \(^{24}\). The process of randomization may leave much to be desired in periodontal trials where only 7% of the studies provide evidence on allocation concealment \(^{36}\).

An important corollary to randomization is the intent-to-treat principle. The most perfect randomization scheme is useless if the analysis is based on a set different than the randomized set. For dental trials, which rely extensively on surrogate endpoints, every randomized patient and site should be accounted for in the analyses. It is used to be common that once a patient became edentulous, or when patient lost a tooth, the patient or tooth would be dropped from the analysis. There was the mistaken belief that the split-mouth design would protect against such biases \(^{37}\). Both a proper randomization and an intent-to-treat analysis are critical as deviations prohibit reliable information on treatment effectiveness.

**Trial registration and negative results; is there a solution?**

A 1997 review came to the conclusion that non-steroidal anti-inflammatory medications (NSAIDS) have been “unequivocally” shown to have “primary therapeutic efficacy for periodontitis in humans \(^{38}\)”. A casual PUBMED search may further identify small exploratory trials suggesting that such treatments are “a useful adjunct in the treatment of rapidly progressive periodontitis”\(^{39}\). Some clinicians may have been tempted to extrapolate such reports and prescribed rofecoxib based on the positive findings of in-vitro evidence \(^{40}\). Yet, what was not published in the literature was that at least two large multi-center pivotal trials were conducted whose results were to the best of our knowledge not published. As a result, the published evidence consists of the small exploratory trials and the enthusiastic expert reviews, the unpublished evidence consists of the large multi-center studies with negative findings whose results disappeared into a black hole.

Such selective publication bias endangers a fair assessment of harms and benefits. Periodontal patients prescribed NSAIDS can expect a four-fold increased risk of serious gastro-intestinal complications \(^{41}\). For every 57 periodontal patients prescribed rofecoxib for up to 3 years, one patient can be expected to have a fatal or non-fatal myocardial infarction or stroke or a death from an unknown cause \(^{42}\) as a direct result of taking rofecoxib. Such risks may be acceptable to some clinicians if indeed anti-inflammatory medications are effective against periodontitis. However, the non-publication of pivotal trial results raises substantial doubt on the question of effectiveness.
The ongoing efforts at clinical trial registration and repositories for clinical trial results go a long way at solving such problems. However, even in medicine it is realized that the current systems is still far from perfect \(^4\) and that the black hole for negative results is ever present. In dental research, where trials are conducted across the globe in locations ranging from China to Middle-America to Eastern European countries, the potential for black holes appear at least as large as in medicine. Unless government organization such as the Food and Drug Administration, professional organizations such as the American Dental Association, and leading dental journals insists on pre-trial registration of a protocol and on the mandatory reporting of negative results, false-positives will continue to crowd out negative findings, and a warped reality will remain present in the published literature. Until solutions for this challenging problem are found, what is not published may remain more informative than what is published.

**Size of the treatment effect and surrogates; how it relates to hypothesis testing.**

Both the type of endpoint and the size of the treatment effect impacts on when to use statistical hypothesis testing and on what to specify for \(\alpha\) and \(\beta\) \(^4\). Four situations are differentiated and listed in order of decreasing clinical importance.

**Clinical importance level 1:** Bone marrow transplantation for leukemia, or tooth extraction for resolving the pain of acute pulpitis, are examples where treatment has a large impact on a clinically relevant outcome. Rigorous clinical trial design, hypothesis formulation, or issues such as \(\alpha\) and \(\beta\) are irrelevant for establishing effectiveness. However, the less morbidity and mortality involved with the disease under investigation, the more important safety issues are. The use of amalgam in dentistry offers one example were safety, and not effectiveness, was the primary reason to conduct randomized controlled trials \(^4\).

**Clinical importance level 2:** Reliable detection of small treatment effects on true endpoints requires the conduct of randomized controlled trials, formal hypothesis formulation and specification of \(\alpha\) and \(\beta\). Such an approach minimizes the risk for false-positive conclusions. Trials designed on true endpoints are typically large in sample size allowing for the reliable detection of side-effects, a key issue for dental products with large market penetration.

Findings on surrogate endpoints are of a lesser clinical importance than findings on true endpoints. As a result, there is a more stringent need to have evidence on the absence of long-term adverse patient outcomes. Results on surrogate endpoints are classified depending on the size of the effect.

**Clinical Importance level 3:** A 10+mm reduction in probing depth or a 90% reduction in incidence of caries lesions that extend into the dentin are examples of large effects on surrogates. No hypothesis formulation, RCTs, or specification of \(\alpha\) and \(\beta\) are required to reliably detect such effects. The unforeseen long-term consequences on morbidity or mortality are of concern and cohort studies or case-control studies are needed to ensure that the large surrogate effects translate into a large patient benefits. For instance, while bone marrow transplants resulted in large amount of periodontal regeneration, longer follow-up indicated that up to 50% of such teeth were lost due to root resorption.

**Clinical Importance level 4:** The conduct and analysis of the randomized controlled trials designed to detect small effects on surrogate endpoints should be most rigorous and specify small \(\alpha\) and \(\beta\) as even tiny biases can increase the risk in false-positive conclusions. Use of sophisticated measurements techniques may allow for the design of studies where subtle treatment effects can be detected with
small sample sizes. Such studies however cannot reliably detect side-effects. Specifying small \( \alpha \) and \( \beta \) will not only allow to reduce the rate of false-positive, but also to detect side-effects more reliably.
Conclusion

Hypothesis testing and randomization are two key elements that provide a probabilistic basis for calculating false-positive and false-negative error rates. Reliable inference on the safety and the effectiveness of treatments is possible if the hypothesis is specified prior to the data collection, if the outcome is a true endpoint, if the likelihood for identifying effective treatments is large, and if all clinical trial results are reported, not just the positive ones. Potential solutions to reduce the rate of false positive conclusions includes providing a range for π prior to the initiation of the trial, a requirement to register clinical trials and their primary hypothesis prior to the conduct of the study, an ability to access all registered clinical trial results, specifying small α and β, and usage of true rather than surrogate endpoints. The ongoing changes in the drug and device approval process in the medical arena may ultimately lead to the implementation of some these issues in dental research and thereby lead to a reduction in false-positive results.

References


