Review

Biological Psychiatry

Statistical and Methodological Considerations for the Interpretation of Intranasal Oxytocin Studies

Hasse Walum, Irwin D. Waldman, and Larry J. Young

ABSTRACT

Over the last decade, oxytocin (OT) has received focus in numerous studies associating intranasal administration of this peptide with various aspects of human social behavior. These studies in humans are inspired by animal research, especially in rodents, showing that central manipulations of the OT system affect behavioral phenotypes related to social cognition, including parental behavior, social bonding, and individual recognition. Taken together, these studies in humans appear to provide compelling, but sometimes bewildering, evidence for the role of OT in influencing a vast array of complex social cognitive processes in humans. In this article, we investigate to what extent the human intranasal OT literature lends support to the hypothesis that intranasal OT consistently influences a wide spectrum of social behavior in humans. We do this by considering statistical features of studies within this field, including factors like statistical power, prestudy odds, and bias. Our conclusion is that intranasal OT studies are generally underpowered and that there is a high probability that most of the published intranasal OT findings do not represent true effects. Thus, the remarkable reports that intranasal OT influences a large number of human social behaviors should be viewed with healthy skepticism, and we make recommendations to improve the reliability of human OT studies in the future.

Keywords: Bias, Effect size, Neuroendocrinology, Positive predictive value, Social cognition, Statistical power

http://dx.doi.org/10.1016/j.biopsych.2015.06.016

Oxytocin (OT) has been the subject of intensive investigation for several decades due to its pivotal role in reproductive physiology. More recently, attention has turned to its role in regulating complex social behavior, including parental care, social bonding, and social cognition in general (1–6).

Much of the excitement regarding OT over the past decade has been driven by a remarkable proliferation of research suggesting that intranasal OT (IN-OT) administration influences various aspects of human social behavior (7). These studies appear to provide compelling, but sometimes bewildering, evidence for the role of OT in influencing complex social cognitive processes in humans. If all of the conclusions from human OT research were true, one might characterize OT as the elixir of the social brain. Yet, we know from the nature of the scientific process that all findings that are statistically significant do not represent true effects.

Our goal here is to discuss quantitatively some statistical and methodological limitations that should moderate our interpretation of the vast literature on the effects of IN-OT on human social behavior. These limitations are not specific to IN-OT research, but we are particularly concerned that there is a certain degree of irrational exuberance emerging from this field that could be detrimental to the field when initial reports are not replicated. We feel that researchers and the media should maintain an appropriate level of skepticism and regard individual reports not as fact, but as evidence to be considered in the context of the limitations presented here. Our discussion is focused on evidence-based concepts and we consider statistical and methodological issues of IN-OT studies, including factors like statistical power, prestudy odds, and bias. We conclude that the literature on the effects of IN-OT on human behavior should be interpreted cautiously and provide some recommendations to improve reliability of IN-OT data and moving OT research forward.

THE STATISTICAL POWER OF BEHAVIORAL IN-OT STUDIES IN HUMANS

Statistical power is the probability that a test will be able to reject the null hypothesis considering a true relation with a given effect size. True effect size values are, however, difficult, if not impossible, to acquire. This problem can to some extent be avoided by using effect size estimates from meta-analyses of relevant prior studies. Even though summary effects from meta-analyses can be inflated due to various sources of bias (8), these analyses provide the best estimates of the true effect size.

To date, three meta-analyses of the effects of IN-OT on human behavior have been published. Van IJzendoorn and Bakersmans-Kranenburg (9) investigated the effect of IN-OT on facial emotion recognition (13 effect sizes, total $n = 408$), trust to in-group (8 effect sizes, total $n = 317$), and trust to

ISSN: 0006-3223

out-group (10 effect sizes, total $n = 505$). Shahrestani *et al.* (10) conducted a meta-analysis of studies examining the effect of OT on recognition of basic emotions (7 effect sizes, total $n = 381$). Bakermans-Kranenburg and Van IJzendoorn (11) studied the effect of IN-OT in clinical trials (19 effect sizes, total $n = 304$). These studies yielded summarized effect sizes ranging from $d = .21$ to $d = .48$. We reanalyzed the data from these meta-analyses by calculating the average effect size for healthy subjects included in the studies in the meta-analyses, weighted by sample size. This resulted in a mean effect size of $d = .28$. The median sample size for the individual studies in these meta-analyses was 49 individuals. For simplicity, when determining the individual sample sizes, we multiplied the n by 2 for studies adopting a within-subject design. Using the effect size estimates and median sample size from these studies, we calculated the average power, assuming an alpha level of 5%, using G*Power software (12). For verification, power calculations were also performed using simulation in R (3.1.1; The R Foundation for Statistical Computing; http://www.R-project.org) (13) (Figure 1) and yielded very similar results compared with G*Power. Our results indicate that the average study investigating the effect of IN-OT in healthy subjects has a statistical power of 16%. For clinical trials, the median sample size is 26 individuals and the effect size is $d = .32$, resulting in a statistical power of only 12%. If the studies included in the meta-analyses are representative of the field, statistical power values of 16% for studies investigating IN-OT effects in healthy subjects and 12% in clinical trials are certainly very low but not very different from studies in neuroscience in general (average power = 21%) (14). In Figure 1, we show the achieved statistical power for different effect sizes plotted against the range of sample sizes for the studies included in the meta-analyses ($n = 4$ to 112). As seen in Figure 1, IN-OT studies in humans are underpowered. For all sample sizes and effect sizes, the power is lower than 80% (often considered the standard for minimal adequate statistical power). Even in the situation in which the true effect size is .48 (the largest observed) and the sample size also is the largest observed within studies included in the meta-analyses (i.e., $n = 112$), the statistical power is no higher than 70%.

As mentioned above, the median sample size for studies in healthy subjects and in clinical trials was 49 and 26 individuals, respectively. Studies within this sample size range can only reliably detect large effect sizes ($d = .81$ to $d = 1.14$) with 80% power. To achieve 80% power given the summarized effect size of .28 for healthy subjects, a sample size of 352 individuals would be needed. Similarly, for the average effect size in clinical trials ($d = .32$), 310 individuals would be needed to achieve 80% power.

Why is it a problem that IN-OT studies are underpowered? If the statistical power is only 12% to 16%, this implies that the false-negative rate is between 84% and 88%. In other words, replication attempts of true positive findings would fail up to 88% of the time. Failure to replicate calls into question the validity of the initial finding. This is obviously very problematic since the majority of replication attempts using samples of roughly the same size as the original studies within the field of IN-OT in humans will fail for statistical reasons alone and could significantly influence funding and regulatory agencies making decisions regarding clinical applications of IN-OT. Thus, individual reports should be interpreted in the context of the totality of the evidence, such as in meta-analyses.

Further, in situations when an underpowered study detects a true effect, the estimate of this effect size is likely to be highly exaggerated, a phenomenon often referred to as the winner's curse (15). In Figure 2, we show the effect size inflation for the same effect sizes as in Figure 1, plotted against the sample size range. Clearly, IN-OT studies considerably overestimate the true effect size. In cases where the sample size is below 40 individuals, the inflation is very large, but even when $n = 100$, the overestimation of the effect is by no means negligible. Inflation of this extent makes it difficult to determine adequate sample size for replication studies and could imply overconfidence in positive findings.

One could argue that the most important problem to avoid in science is false-positives and that this is largely accomplished by adopting a relatively conservative alpha level of 5%. As Ioannidis (16) showed by statistical modeling, this is not the case. The proportion of reported positive findings that are actually true can be described as the positive predictive value (PPV) (16) and is further discussed below.

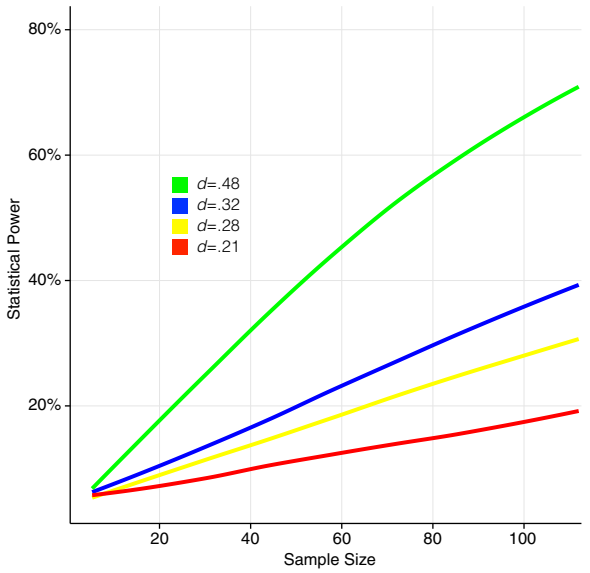

Figure 1. Statistical power as a function of effect size and sample size. The figure shows the relationship between sample size and statistical power for four different effect sizes. Power calculations were performed using simulations in R (3.1.1). In the simulations, half of the sample was drawn from a standard normal distribution and the other half from a second normal distribution with a mean representing the investigated effect size. This procedure was repeated 1000 times per effect size and sample size. Power was determined as the proportion of these 1000 experiments rejecting the null hypothesis (using one-way analysis of variance), with the alpha level set to .05. The effects sizes presented in the figure represent the largest ($d = .48$) and smallest ($d = .21$) effects sizes within the field of intranasal oxytocin studies in humans, as well as the mean effect size for healthy subjects ($d = .28$) and clinical trials ($d = .32$). It is clear that the studies within this field are underpowered, since for all effect sizes and sample sizes the statistical power is below 80%, the standard for minimal adequate statistical power.