HHS Public Access

Author manuscript

Biol Psychiatry. Author manuscript; available in PMC 2017 February 01.

Published in final edited form as:

Biol Psychiatry. 2016 February 1; 79(3): 251–257. doi:10.1016/j.biopsych.2015.06.016.

Statistical and methodological considerations for the interpretation of intranasal oxytocin studies

Hasse Walum[1,2,3,4,*], **Irwin D. Waldman**[2,4], and **Larry J. Young**[1,2,3]

[1]Silvio O. Conte Center for Oxytocin and Social Cognition

[2]Center for Translational Social Neuroscience

[3]Yerkes National Primate Research Center, Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia, United States

[4]Department of Psychology, Emory University, Atlanta, Georgia, USA

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 paper 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, pre-study 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

Social cognition; Neuroendocrinology; Statistical Power; Bias; Positive predictive value; Effect size

[*]Corresponding author: Hasse Walum, 954 Gatewood Rd. Yerkes National Primate Research Center, Emory University, Atlanta GA 30329. hasse.walum@emory.edu.

Conflicts of interest

LJY has applied for a patent (US20120108510 - Methods of improving behavioral therapies) for combining melanocortin agonists with behavioral therapies to enhance social cognition in psychiatric disorders. HW and IW declare no biomedical financial interests or potential conflicts of interest.

Introduction

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, pre-study 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 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 out-group (10 effect sizes, total $N = 505$) (9). Shahrestani et al. conducted a meta-analysis of studies examining the effect of OT on recognition of basic emotions (7 effect sizes, total $N = 381$)(10). Bakermans-Kranenburg and Van IJzendoorn studied the effect of IN-OT in clinical trials (19 effect sizes, total $N=304$)(11). These studies yielded summarized effect sizes ranging from $d=0.21$ to $d=0.48$. 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*=0.28. 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 two 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) (13) (Figure 1), and yielded very similar results compared to 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*=0.32, 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 0.48 (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 were 49 and 26 individuals, respectively. Studies within this sample size range can only reliably detect large effect sizes (*d*=0.81 to *d*=1.14) with 80% power. In order to achieve 80% power given the summarized effect size of 0.28 for healthy subjects, a sample size of 352 individuals would be needed. Similarly, for the average effect size in clinical trials (*d*=0.32), 310 individuals are 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.

The positive predictive value of behavioral IN-OT studies in humans

The formula for calculating the PPV using information on power ($1-\beta$), the pre-study odds (R; described below) and the alpha level (α) is: $PPV = ((1-\beta) \times R)/((1-\beta) \times R + \alpha)$.

Although rather exact values for both power (calculated above) and alpha level (most commonly set to 5%) can be put into this formula, picking a reasonable value for R is more problematic. Within any research field both true and false hypotheses can be made. Thus R represents the ratio of the number of true relationships to the number of false relationships. For example, if the effects researchers within a field look for actually exist half of the time, this corresponds to $R = 1$ (1/(2−1)). Pre-study odds estimates can be viewed as informed predictions of the likelihood of the postulated hypotheses within a field being true.

Button et al. have argued that large-scale phase III clinical trials represent the case when the value of R will be the highest (17). The argument is that these studies are relatively low risk and represent the end product of a long process of biomedical research. Data suggest that the R for phase III clinical trials is approximately 1, meaning that in about 50% of the time the drugs that make it to these trials are more effective than the current “gold standard” treatment (18).

We do not argue that we can determine with much certainty the true value of R for behavioral IN-OT studies in humans. It seems reasonable however to assume that it is considerably lower than for phase III clinical trials. Although the idea that manipulating OT in humans could influence social behavior is supported by rigorous animal research, we argue that the pre-study odds of studies investigating the effect of IN-OT still would be low, primarily due to limitations in deep brain penetration of OT when administered intranasally (2, 19).

Also, the publication culture promotes novel, and often surprising, findings and this will motivate researchers to postulate improbable hypotheses (20). Regarding the IN-OT field this seems to translate into a wide spectrum of investigated phenotypes beyond what can reasonably be predicted based on prior animal research. When a research field is, to a large degree, exploratory, the pre-study odds decrease.

In Figure 3 we show how PPV estimates differ depending on statistical power and pre-study odds. The value for the alpha level is kept at 5%. Three values, 12%, 16% and 80%, are used for power, representing the average power of studies of IN-OT in healthy subjects, in clinical trials and the standard for adequacy, respectively. We picked a range of R values

from 0 (all effects are null effects) to 1 (equal to the pre study odds of phase III clinical trials). As shown in Figure 3 when *R* is low the probability of reported positive findings being true is low. Importantly, low statistical power has a large negative impact on PPV. The negative influence of decreasing pre-study odds on PPV is strong in the case when the power is as low as the average power of behavioral studies of IN-OT in humans. If the value of *R* for IN-OT studies is approximately 0.10, comparable to what has suggested for experimental psychology (21) and exploratory epidemiology (16), the PPV is no higher than 24%, given a statistical power of 16%. However discouraging these estimated values might seem they assume no influence of bias, meaning that the actual PPV values could be considerably lower.

Bias in behavioral IN-OT studies in humans

When reading the literature on behavioral IN-OT studies in humans it is obvious that most papers report positive findings, which is in line with a study by Fanelli showing that more than 80% of scientific publications in various sciences report positive results (22). Considering the low statistical power within the field of IN-OT we would expect that approximately 80% of all attempts to detect a *true* effect would fail. But it seems very unlikely that all hypotheses about how IN-OT affects human behavior are true, as described in the previous section. If about 10% of all postulated hypotheses are correct, similar to what is expected for experimental psychology in general (21), we would expect, in the absence of bias, that only about 2% of the investigated effects would turn out to be statistically significant.

We investigated to what extent low power is reflected in behavioral IN-OT studies by examining the proportion of positive effects published in 2014 (described in Supplemental information). Twenty-nine out of the investigated 33 papers (88%) reported at least one positive finding (uncorrected p-value below 0.05). However, 17% (62 out of 357) of all tested effects were statistically significant when calculating the proportion of positive findings over the total amount of tests across all studies. If this number represents the true proportion of successful experiments, this would mean that almost all investigated hypotheses within these studies are true. An alternative, perhaps more plausible, explanation is that there could be a significant amount of unpublished negative or inconclusive results, a phenomenon referred to as the “file-drawer effect” or publication bias (23). Bias of this kind could have serious consequences such as failure to replicate findings (24), and there are also reasons to believe that publication bias is more likely to affect low powered studies (25). The meta-analyses mentioned in this paper all tested for evidence of publication bias in the IN-OT literature, with mixed results (9–11). However, publication bias tests might be problematic for IN-OT studies due to small sample sizes resulting in insufficient statistical power and sample size variability for the tests to yield reliable and significant results (26).

In addition to publication bias, the excess of statistically significant findings may be explained partly by the use of other questionable research practices. As shown by Simmons et al. (2011), Type I errors are easily inflated when researchers allow themselves to employ undisclosed analytic flexibility regarding choice of statistical model, definition of variables, and the rationale for exclusion of outliers (27). Such questionable practices are common in

psychology (28), and likely pervasive in other disciplines as well. Use of multiple analyses on the same data set and selective reports of statistical methods used, or insufficient correction for multiple comparisons, can cause inflation in effects (29). Reporting bias and multiple comparisons issues seem to be problems for studies of IN-OT. For example, out of the 33 IN-OT papers we investigated, only 3 mention that any correction for multiple comparisons was performed, and in these cases it is unclear how corrections were applied. The average number of tests performed in the 33 studies is 11, but it should be mentioned that we have no way to assess to what extent these tests are independent. If many outcome variables are investigated and selective reporting is present, or correction for multiple testing is not adopted, the likelihood of any study finding statistically significant results will be determined by the number of dependent variables rather than an actual underlying effect. We illustrate the impact of this kind of bias in Figure 4 by showing the effect of number of uncorrected multiple comparisons on PPV for power estimates of 12% and 16%. The PPV values in Figure 4 was estimated using the formula presented by Ioannidis (16) (PPV = $R(1-\beta^n)/(R+1-(1-\alpha)^n-R\beta^n)$, where n equals the number of comparisons. For example, if data from 10 independent tests are collected and only reported for one of these, the statistical power is 16% for all tests and the R-value is 0.10, then the PPV goes down from 24% to 17%.

Within the field of behavioral IN-OT studies in humans, there are several other examples of questionable practices including the use of statistical methods, such as unjustified use of one-tailed tests and unexplained exclusion of outliers. In addition, it has been shown that the more popular or “trendy” a scientific field is the less likely it is to generate true findings (16) and the extent to which a study overestimates true effects is positively correlated with the impact factor of a journal (30). Studying behavioral effects of OT is at the moment a hot research topic, attracting many new research groups, and studies within this field have been published in very high impact journals such as *Science* and *Nature*.

Taken together we think that it is fair to say that the field of behavioral IN-OT studies in humans is prone to several types of bias. If we consider both the potentially low PPV and the influence of different types of bias described above, it is possible that most published “positive” findings within the field actually are false positives, and thus do not represent true effects.

Conclusions and Recommendations

Our analyses demonstrate that IN-OT studies are generally considerably underpowered. This leads to a high probability that the reported effects of IN-OT are overestimated. Also, underpowered studies are prone to other types of biases, such as the use of questionable research practices. The combination of low power and low pre-study odds results in low PPV estimates. Taken together this suggests that most of the reported “positive” findings regarding how OT affects human behavior are likely to be false positives. From a statistical point of view this problem might not be any more serious for this field than other domains of psychology (20, 21, 31) or neuroscience in general (14). However, treatments involving OT have clinical promise and studies investigating the effects of this peptide not only receive attention from the scientific community, but are often mentioned in the media and are of

interest to the general public. Nasal sprays purportedly containing OT are available on the internet for a variety of indications and parents are increasingly seeking IN-OT as a therapy for their children with autism. We therefore believe that these issues deserve attention and it is important that people are given a chance to assess the reliability of studies within this popular field when drawing conclusions about the true nature of the effects of IN-OT. Increased confidence in data from clinical trials involving IN-OT would entice funding agencies and pharmaceutical companies to support more research in the field.

Our calculations are focused on human studies, but we see no reason to expect that these results would differ for other species given OT intranasally, like non-human primates, as long as the dose is adjusted for body weight. Even though it is possible that the IN-OT literature so far contains a relatively high proportion of false positive findings we believe the behavioral effects of OT are indeed evolutionary conserved (6, 32) and that the evidence from rodent research is compelling. The problem here is not the underlying hypothesis that OT could be a modulator of human behavior, it is in the certainty in which we can have faith that any particular reported finding represents a true effect. Changes in research practices could increase the trustworthiness of the data.

We agree with Button et al. (14) who suggested that the way to handle these problems in neuroscience is for researchers to perform *a priori* power calculations, disclose methods and findings transparently (preferably making all data available to others within the field), and to work collaboratively to increase power and replicate findings. Researchers and funding agencies should acknowledge that better powered studies are needed, and funding be made available to make this possible. Our calculations show that sample sizes of hundreds of individuals are necessary to produce reliable data given these effect size estimates. If such studies are impractical for single research groups, multi-site collaborative studies are warranted. Large-scale collaboration efforts have been successful within fields like human genetics (33) and psychology (34) and replicated findings from these consortia can be considered more reliable. Further, gathering repeated measures from individuals on the same behavioral task can be an efficient way to increase statistical power, thus avoiding the need to recruit as many participants. Our simulations show that if the effect size is d=0.28 and the sample size is 49 individuals, 8 repeated trials (4 in each drug condition) can be enough to gain 80% power even when the correlation between trials is as low as 20%.

Considering the uncertainty of IN-OT findings, replications are needed. Although it is often argued that conceptual replications assess both the validity and generality of previous studies, and therefore should be considered more effective than direct replications, this is not necessarily true (21). In order to determine if IN-OT can affect human behavior in a specific manner, ideally studies need to be repeated using methods that are identical to those in the original study, or at least as close as possible. However, having several groups investigating the same behavioral phenotypes is not enough. The data these groups collect need to be presented transparently, for example deposited in publicly available databases, otherwise this will only lead to unreliable effects being presented due to reporting bias. Also, since the IN-OT studies are underpowered it does not make sense to try to replicate findings in samples of the same size as the original study. This will lead to many unsuccessful replication attempts due to insufficient power, undermining the original true finding. *A*

priori power calculations to determine the appropriate sample size needs to be performed before data collection starts. Finally editors of high impact journals should not reject replication studies on the basis of lack of novelty, as a replication may be more important than the initial finding.

While our assessment of the current state of the IN-OT field in psychology may appear pessimistic, we remain optimistic about the future of human OT research. There may be improvements in efficacy of manipulating the central OT receptor system, including more efficient intranasal delivery paradigms, or the development of small molecule agonists, or positive allosteric modulators (35). We believe that one of the reasons why human IN-OT studies are underpowered is because the current intranasal route of administration is not optimal for neuropeptides, leading to relatively small effect sizes. To test the veracity of this hypothesis we gathered data from animal studies investigating the behavioral effects of centrally manipulating the OT/vasopressin systems. Specifically, we focused on vole studies using the partner preference test to assess behavior, a literature well known by the authors. Effect sizes were estimated for 30 independent experiments comprising 668 individuals in total, weighted by the sample size of the individual experiments. This yielded a summarized effect size of $d = 0.76$. Although these vole studies are underpowered considering the small samples used (median $n = 22$, $1-\beta = 0.43$), these results indicate that larger effect sizes can be achieved by more efficient routes of administration. Central injections will for obvious reasons never be a common way to administer drugs in humans. However, there is an intriguing possibility that endogenous OT release could be stimulated pharmacologically and have a robust effect on OT-dependent behavior (35, 36). For example, melanocortin receptor agonists potentiate central OT release in the brain (37, 38), facilitate OT-dependent partner preferences in prairie voles (35, 37, 39, 40), and mimic the prosocial effects of OT in a mouse model of autism (41). These alternate means of manipulating the OT system, if used in humans, could potentially increase the effect size to levels similar to central injections of OT (36). In addition, research should be guided by information of the neurobiological mechanisms underlying the effects of OT on behavior gained from animal research. For example, OT receptors are concentrated in cholinergic brain regions regulating visual and auditory attention in nonhuman primates (6, 42, 43), consistent with the putative localization of OT receptors in human brain (44). Research in animals suggest that OT enhances the salience and reinforcing value of social stimuli, which is consistent with some evidence from human studies involving IN-OT (45–47) and genetics (48). Understanding the precise neural and cognitive effects of OT manipulation on the processing of social information can be used to increase the efficacy of OT-based therapies to improve social function in psychiatric disorders (36). Finally, in the meta-analysis by Bakermans-Kranenburg and Van IJzendoorn (11) IN-OT seems to have the largest effect size in individuals with autism. Only four studies of autism were included in the analysis and should therefore be interpreted cautiously. However, the larger effect size for autism (d=0.57) compared to all clinical studies (d=0.32) could indicate that focusing on disorders characterized by deficits in social-communicative skills could produce reliable results.

In summary, there are multiple ways for increasing the reliability of OT related research in humans so that we can have a true understanding of the function of OT in the human brain and maximize the therapeutic potential of this important neuropeptide.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Preparation of this manuscript was supported by NIH grants NIH grants R01MH096983 and 1P50MH100023. Additional funding was provided by NIH OD P510D11132 to YNPRC. HW thanks the Swedish Brain Foundation for financial support.

References

1. Bartz JA, Zaki J, Bolger N, Ochsner KN. Social effects of oxytocin in humans: context and person matter. Trends in cognitive sciences. 2011; 15:301–309. [PubMed: 21696997]
2. Churchland PS, Winkielman P. Modulating social behavior with oxytocin: how does it work? What does it mean? Hormones and behavior. 2012; 61:392–399. [PubMed: 22197271]
3. Evans SL, Monte OD, Noble P, Averbeck BB. Intranasal oxytocin effects on social cognition: A critique. Brain research. 2013
4. Guastella AJ, Hickie IB, McGuinness MM, Otis M, Woods EA, Disinger HM, et al. Recommendations for the standardisation of oxytocin nasal administration and guidelines for its reporting in human research. Psychoneuroendocrinology. 2013; 38:612–625. [PubMed: 23265311]
5. Ross HE, Young LJ. Oxytocin and the neural mechanisms regulating social cognition and affiliative behavior. Frontiers in neuroendocrinology. 2009; 30:534–547. [PubMed: 19481567]
6. Young LJ. Oxytocin, Social Cognition and Psychiatry. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology. 2015; 40:243–244. [PubMed: 25482173]
7. Young LJ, Flanagan-Cato LM. Editorial comment: oxytocin, vasopressin and social behavior. Hormones and behavior. 2012; 61:227–229. [PubMed: 22443808]
8. Pereira TV, Ioannidis JP. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. Journal of clinical epidemiology. 2011; 64:1060–1069. [PubMed: 21454050]
9. Van IJzendoorn MH, Bakermans-Kranenburg MJ. A sniff of trust: meta-analysis of the effects of intranasal oxytocin administration on face recognition, trust to in-group, and trust to out-group. Psychoneuroendocrinology. 2012; 37:438–443. [PubMed: 21802859]
10. Shahrestani S, Kemp AH, Guastella AJ. The impact of a single administration of intranasal oxytocin on the recognition of basic emotions in humans: a meta-analysis. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology. 2013; 38:1929–1936. [PubMed: 23575742]
11. Bakermans-Kranenburg MJ, van IJMH. Sniffing around oxytocin: review and meta-analyses of trials in healthy and clinical groups with implications for pharmacotherapy. Translational psychiatry. 2013; 3:e258. [PubMed: 23695233]
12. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007; 39:175–191. [PubMed: 17695343]
13. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
14. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews Neuroscience. 2013; 14:365–376.

15. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008; 19:640–648. [PubMed: 18633328]
16. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005; 2:e124. [PubMed: 16060722]
17. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Empirical evidence for low reproducibility indicates low pre-study odds. Nature reviews Neuroscience. 2013; 14:877.
18. Djulbegovic B, Kumar A, Glasziou P, Miladinovic B, Chalmers I. Medical research: Trial unpredictability yields predictable therapy gains. Nature. 2013; 500:395–396. [PubMed: 23969443]
19. Leng G, Ludwig M. Intranasal oxytocin: myths and delusions. Biological psychiatry. 2015 In Press.
20. Bakker M, van Dijk A, Wicherts JM. The Rules of the Game Called Psychological Science. Perspectives on Psychological Science. 2012; 7:543–554. [PubMed: 26168111]
21. Pashler H, Harris CR. Is the Replicability Crisis Overblown? Three Arguments Examined. Perspectives on Psychological Science. 2012; 7:531–536. [PubMed: 26168109]
22. Fanelli D. “Positive” results increase down the Hierarchy of the Sciences. PloS one. 2010; 5:e10068. [PubMed: 20383332]
23. Rosenthal R. The file drawer problem and tolerance for null results. Psychological bulletin. 1979; 86:638.
24. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews Drug discovery. 2011; 10:712.
25. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PloS one. 2008; 3:e3081. [PubMed: 18769481]
26. Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. CMAJ: Canadian Medical Association journal = journal de l’Association medicale canadienne. 2007; 176:1091–1096.
27. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science. 2011; 22:1359–1366. [PubMed: 22006061]
28. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science. 2012; 23:524–532. [PubMed: 22508865]
29. Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. Erroneous analyses of interactions in neuroscience: a problem of significance. Nature neuroscience. 2011; 14:1105–1107. [PubMed: 21878926]
30. Munafo MR, Stothart G, Flint J. Bias in genetic association studies and impact factor. Molecular psychiatry. 2009; 14:119–120. [PubMed: 19156153]
31. Bertamini M, Munafò MR. Bite-Size Science and Its Undesired Side Effects. Perspectives on Psychological Science. 2012; 7:67–71. [PubMed: 26168425]
32. Insel TR, Young LJ. Neuropeptides and the evolution of social behavior. Current opinion in neurobiology. 2000; 10:784–789. [PubMed: 11240290]
33. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. American journal of human genetics. 2012; 90:7–24. [PubMed: 22243964]
34. Open-Science-Collaboration . An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. Perspectives on Psychological Science. 2012; 7:657–660. [PubMed: 26168127]
35. Modi ME, Young LJ. The oxytocin system in drug discovery for autism: animal models and novel therapeutic strategies. Hormones and behavior. 2012; 61:340–350. [PubMed: 22206823]
36. Young LJ, Barrett CE. Can oxytocin treat autism? Science. 2015; 347:825–826. [PubMed: 25700501]
37. Modi ME, Inoue K, Barrett CE, Kittelberger KA, Smith DG, Landgraf R, et al. Melanocortin Receptor Agonists Facilitate Oxytocin-Dependent Partner Preference Formation in the Prairie

Vole. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology. 2015

38. Sabatier N, Caquineau C, Dayanithi G, Bull P, Douglas AJ, Guan XM, et al. Alpha-melanocyte-stimulating hormone stimulates oxytocin release from the dendrites of hypothalamic neurons while inhibiting oxytocin release from their terminals in the neurohypophysis. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2003; 23:10351–10358. [PubMed: 14614094]
39. Barrett CE, Arambula SE, Young LJ. The oxytocin system promotes resilience to the effects of neonatal isolation on adult social attachment in female prairie voles. Translational psychiatry. 2015 In Press.
40. Barrett CE, Modi ME, Zhang BC, Walum H, Inoue K, Young LJ. Neonatal melanocortin receptor agonist treatment reduces play fighting and promotes adult attachment in prairie voles in a sex-dependent manner. Neuropharmacology. 2014; 85:357–366. [PubMed: 24923239]
41. Penagarikano O, Lazaro MT, Lu XH, Gordon A, Dong H, Lam HA, et al. Exogenous and evoked oxytocin restores social behavior in the Cntnap2 mouse model of autism. Science translational medicine. 2015; 7:271ra278.
42. Freeman SM, Inoue K, Smith AL, Goodman MM, Young LJ. The neuroanatomical distribution of oxytocin receptor binding and mRNA in the male rhesus macaque (Macaca mulatta). Psychoneuroendocrinology. 2014; 45:128–141. [PubMed: 24845184]
43. Freeman SM, Walum H, Inoue K, Smith AL, Goodman MM, Bales KL, et al. Neuroanatomical distribution of oxytocin and vasopressin 1a receptors in the socially monogamous coppery titi monkey (Callicebus cupreus). Neuroscience. 2014; 273:12–23. [PubMed: 24814726]
44. Loup F, Tribollet E, Dubois-Dauphin M, Dreifuss JJ. Localization of high-affinity binding sites for oxytocin and vasopressin in the human brain. An autoradiographic study. Brain research. 1991; 555:220–232. [PubMed: 1657300]
45. Andari E, Duhamel JR, Zalla T, Herbrecht E, Leboyer M, Sirigu A. Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:4389–4394. [PubMed: 20160081]
46. Auyeung B, Lombardo MV, Heinrichs M, Chakrabarti B, Sule A, Deakin JB, et al. Oxytocin increases eye contact during a real-time, naturalistic social interaction in males with and without autism. Translational psychiatry. 2015; 5:e507. [PubMed: 25668435]
47. Guastella AJ, Mitchell PB, Dadds MR. Oxytocin increases gaze to the eye region of human faces. Biological psychiatry. 2008; 63:3–5. [PubMed: 17888410]
48. Skuse DH, Lori A, Cubells JF, Lee I, Conneely KN, Puura K, et al. Common polymorphism in the oxytocin receptor gene (OXTR) is associated with human social recognition skills. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111:1987–1992. [PubMed: 24367110]

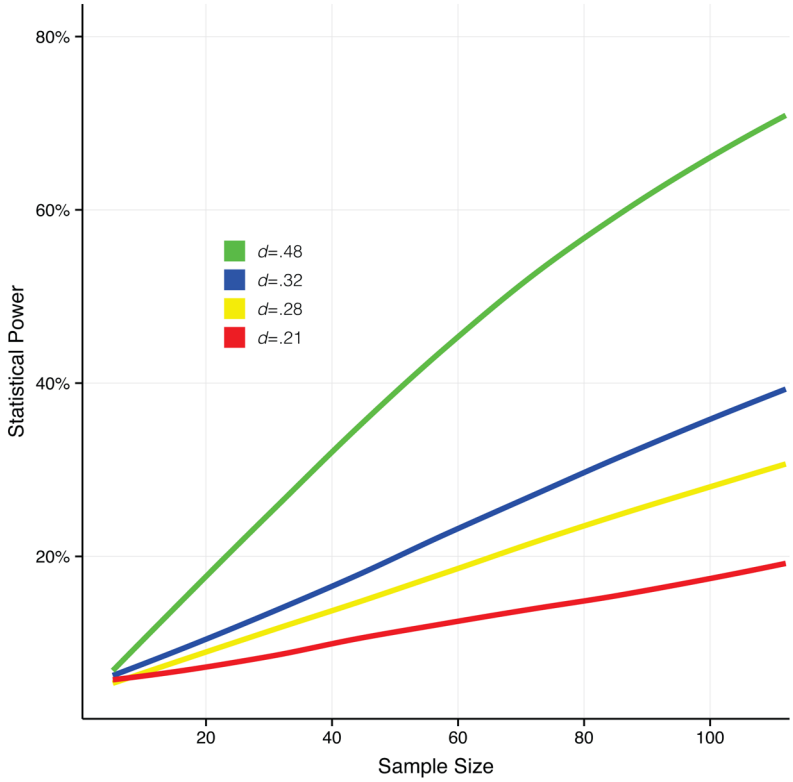

Figure 1. Statistical power as a function of effect size and sample size

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 ANOVA), with the alpha level set to 0.05. 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.

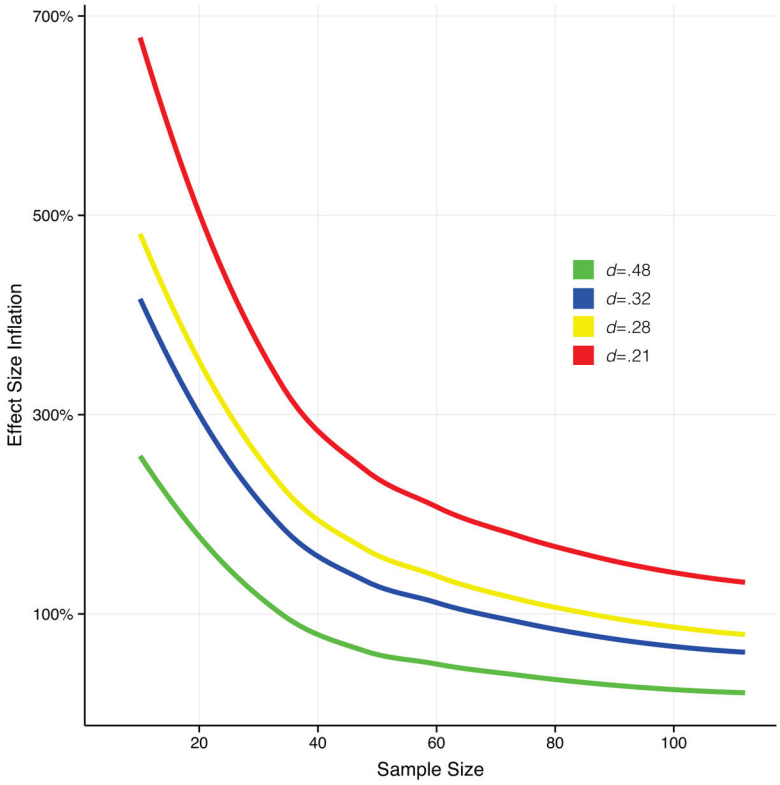

Figure 2. Effect size inflation as a function of statistical power

Effect size inflation is expected to occur when findings need to pass a certain threshold, in this case statistical significance, in order to be considered positive. The smaller the proportion of effects that pass this threshold, the larger the average effect size inflation will be. Since power is an estimate of the proportion of investigated effects that are statistically significant, low power is associated large effect size inflation. The figure shows simulations of effect size inflation. To generate this data we ran simulations in R (3.1.1) (13), using a similar approach as for the power simulations presented in Figure 1. Here, the amount of inflation was calculated by subtracting the true population effect size from the observed effect size of simulations reaching statistical significance ($p<0.05$ using one-way ANOVA), and dividing the difference by the true effect size. This estimate was then averaged over 1000 replicates. Due to the low power of IN-OT studies reported effect sizes within this field are generally inflated to a large degree.

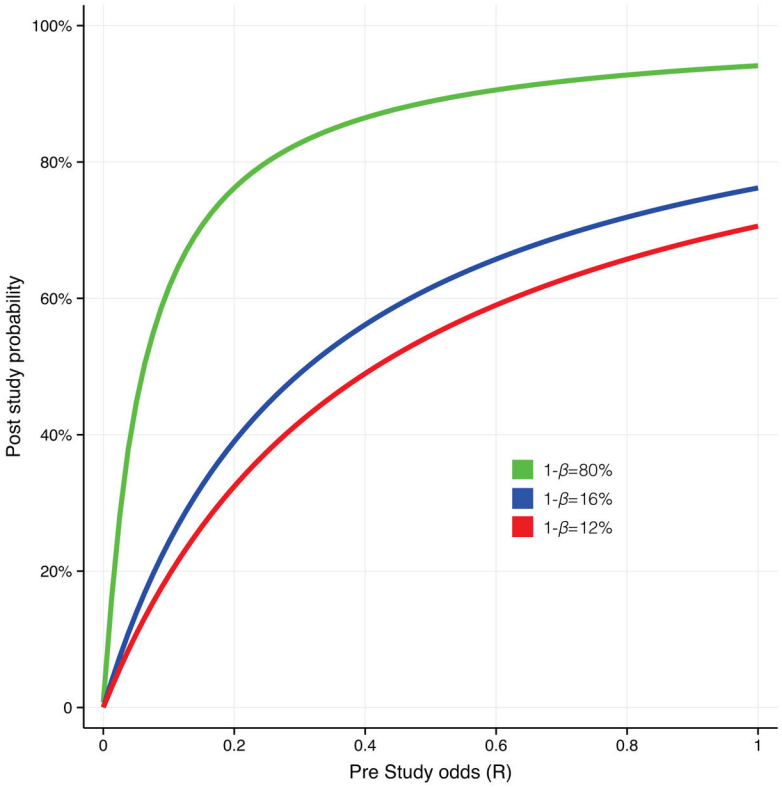

Figure 3. Positive predictive value as a function of pre study odds and statistical power

The probability of a research finding representing a true effect (the positive predictive value) is dependent on the pre study odds of an effect being true (R) and statistical power. The figure shows the relationship between pre study odds and positive predictive value, for the average statistical power in IN-OT studies in healthy subjects (16%) and clinical trials (12%), as well as the standard for minimal adequate statistical power (80%). Clearly, compared to adequately powered studies the probability of a research finding reflecting a true effect is strongly reduced for studies with 12% and 16% power, especially when the pre study odds are low.

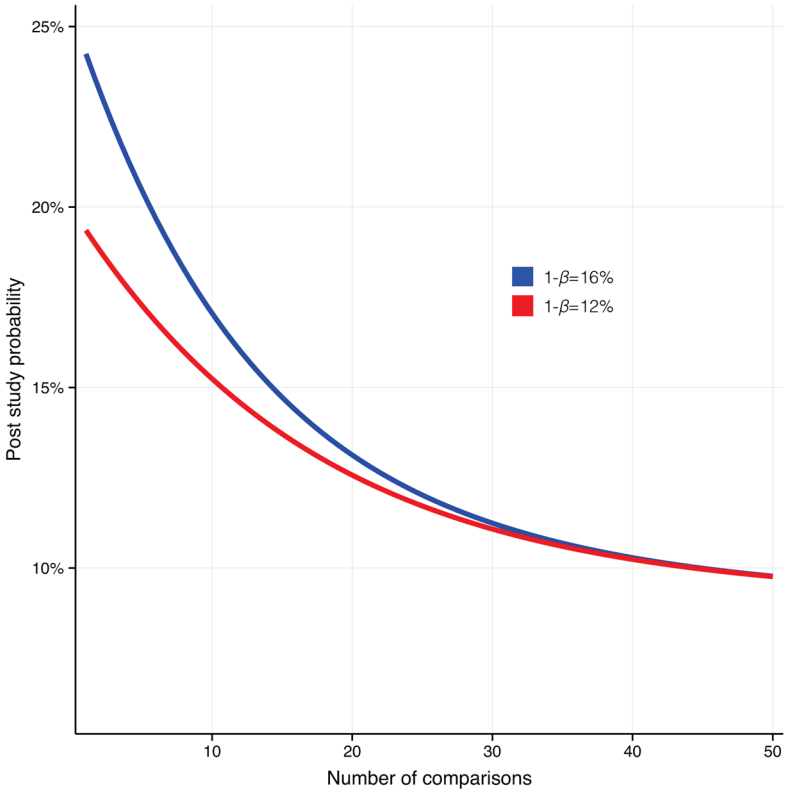

Figure 4. Positive predictive value as a function of multiple comparisons

The odds of a research finding representing a true effect (the positive predictive value) are reduced as the number of uncorrected tests within a study increases. The figure is showing how multiple comparisons without correction influence the positive predictive value for studies with 12% and 16% power.