Statistical and epistemological issues in the evaluation of treatment efficacy of pharmaceutical, psychological, and combination treatments for women's sexual desire difficulties

Meredith Chivers, Rosemary Basson, Lori Brotto, Cynthia Graham, Kyle Richard Stephenson

We were grateful to receive responses from Leonore Tiefer, Anita Clayton and Robert Pyke, and Richard Balon and Robert Segraves, to our commentary (Brotto et al., 2016) on Pyke and Clayton (2015). These commentaries raise a number of substantive statistical and epistemological issues relating to the evaluation of treatment efficacy in pharmaceutical, psychological, and combination treatments for sexual desire difficulties, and caution researchers to remain mindful of sources of bias as we do the science. In what follows, we discuss each of these issues in turn in hopes of encouraging our field to adopt the highest possible standards when carrying out and interpreting treatment outcome research.

Evaluation of treatment efficacy in pharmaceutical and psychological treatments

In their response and critique, Clayton and Pyke (2016) noted: "… the authors criticize the effectiveness of flibanserin based on absolute numerical change after subtracting placebo response rather than statistically significant change from baseline and difference from placebo, minimizing the effect of drug therapy, but rejecting this methodology for psychotherapy studies" (pp. XX). Clayton and Pyke's focus on the statistical significance of differences between drug and placebo conditions is unfortunate, given that effect size, that is, the standardized difference between two treatment means, is an equally (or more) important indicator of the real-world value of treatment when interpreted with guidance from patient-reported minimum benefit data. As we previously pointed out (Brotto et al, 2016), one meta-analysis of psychological treatment outcome studies for low desire in women found a large effect size, d = 0.91 (Frühauf, Gerger, Schmidt, Munder, & Barth, 2013) and, as we discuss later in this commentary, effect sizes are a more meaningful index of treatment effect than statistical significance resulting from null hypothesis statistical testing and its associated p value size. Focusing on statistical significance *on its own* is a problematic form of evaluating treatment effects. As noted in a helpful editorial on the perils of significance testing in evaluating treatment effects, "Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude –not just, does a treatment effect people, but how much does it affect them." (Sullivan & Feinn, 2012, p. 279).

In referring to improvements in sexually satisfying events (SSEs), Clayton and Pyke (2016) stated, "Flibanserin actually improved SSEs by a MEAN of about 1.6–2.5/month, placebo

by about 0.8-1.5/month (drug demonstrated double the effect of placebo (FDA Flibanserin Briefing Document, 2015, p. 30)." This approach to characterizing treatment effects, that is, describing them as "double the effect of placebo," is misleading because it does not accurately reflect the magnitude of a treatment effect. Using the above data reported in the FDA briefing document, we can calculate an effect size for flibanserin versus placebo treatment on SSEs. In Table 9, data from three studies are reported, with all studies reporting statistically significant effects ($p < .05$). Calculating Cohen's *d* for independent samples, the resulting effect sizes are: Study 147 $d = .22$; Study 71 $d = .22$; Study 75, $d = .18$. These effect sizes are considered small (see Sullivan & Feinn, 2012). Small effects can, however, have significant positive impact on patients' lives provided the net improvement outweighs costs associated with treatment.

Clayton and Pyke (2016) rightly emphasized the need to complement evaluation of treatment effects with determination of clinically meaningful response, defined as "the mean increase in SSEs associated with an improvement of one category on the Patient Global Impression (PGI) scale (about 1.5 SSEs/month)." Using this definition, they reported that 46% of flibanserin-treated women reached this goal versus 34% of placebo-treated women and commented, "Again, the important comparison is the statistically significant difference between the 2 groups, not the absolute difference. To do otherwise, eliminates the context of the response, an unacceptable approach in psychotherapy, as well as pharmacotherapy." We believe the important comparison is not statistical significance, but the standardized effect size interpreted in the context of the patient-derived minimal benefit of 1.5 SSEs/month. The appropriate statistic to evaluate the magnitude of treatment effect for these binary-outcome data comparing women who did and did not pass the bar of 1.5 SSEs, is an odds ratio (OR).

Consulting the FDA document, (FDA Flibanserin Briefing Document, 2015, p. 40), we calculated odds ratios for each of the reported effects: Study 147, *OR* = 1.68; Study 71, *OR* = 1.8; Study 75, *OR* = 1.5. Again, these effect sizes are considered small (Sullivan & Feinn, 2012). Readers wishing to learn more are encouraged to review recommendations by Sullivan and Fienn (2012) on using effect sizes versus statistical significance, and to also familiarize themselves with the recent debates over replication of research effects and the concerns about the widespread use of significance testing to assert the "trueness" of research and treatment effects (see Lindsay (2015) for an overview). The problems inherent in this approach, which are beyond the scope of this commentary, are highlighted in a series of tutorials on the pitfalls of relying too heavily on significance testing presented in the "Further Resources" section at the end of this commentary.

We want to emphasize again that the size of an effect on its own may also be uninformative if not interpreted within the context of the research question. Clayton and Pyke (2016) stated that the crux of the issue is that pharmaceutical treatments must meet a higher standard than psychotherapy interventions for HSDD. We agree with this statement and believe that the interpretation of effect sizes should be different for pharmaceutical versus psychotherapeutic interventions because of the fundamentally different nature of side effects associated with these treatments. Specifically, a majority of psychoactive medications (including flibanserin, sildenafil, etc.) are associated with common, detectable, and distressing side effects. Alternatively, most psychotherapeutic treatments (including both cognitive behavioral therapy and mindfulness treatments) are typically associated with positive long-term side effects, including improved physical health (e.g., Murphy, Mermelstein, Edwards, & Gidycz, 2012) and

overall quality of life (e.g., Hofmann, Wu, & Boettcher, 2014). As such, when considering the importance of effect sizes, researchers and practitioners should balance the helpful effects of medications with the cost of their side effects, whereas this process is generally not necessary for psychotherapies, where secondary effects are more likely to be beneficial e.g., cognitive therapies targeting sexual dysfunction also lessening the frequently co-morbid mood disorder. This difference is not a question of "higher" research standards per se (as suggested by Clayton and Pyke) but understanding that *different* standards are needed that acknowledge the reality of fundamental dissimilarities between treatments. An analogous situation might be the comparison of effect sizes of individual versus group psychotherapy. Even if effects are slightly smaller for group therapies (which might at first appear to be "weaker evidence" if identical standards are used), the real-world usefulness of group therapies may in fact be greater because of the lessened costs in terms of therapist time and effort. Such differences are important in informing science and practice and must be carefully considered before universal acceptance of identical standards for evaluating the quality of research evidence.

Are meta-analyses the answer?

In their commentary, Balon and Segraves (2016) asked, “Can anything be clearly concluded from these meta-analyses? Or are we witnessing what Alvan Feinstein (1995) called the statistical alchemy for the 21st century? … Feinstein (1995, 78) does not discard meta-analysis, but complains that “the meta-analysis of randomized trials concentrates on a part of the scientific domain that is already well lit, while ignoring the much larger domain that lies either in darkness or in deceptive glitters.” Fruhauf and colleagues’ (2013) meta-analysis of psychological treatment modalities seems to have some of the issues suggested by Feinstein (1995) e.g., using a

heterogeneous mixture of only four studies in the case of HSDD. This is a well-placed criticism; indeed meta-analysis is no panacea, and can be biased by a number of factors. Inclusion of only those studies that demonstrate significant effects is exactly what Balon and Segraves (2016) described as focusing on the "glittering," choosing only those studies under which a beam of light falls. Empirically sound meta-analyses typically seek to include data lost in the surrounding darkness i.e., those studies that weren't published (to avoid the "file-drawer effect"), usually because they didn't reach the goal of statistical significance that is typically required for academic publications (instead of considering effect size, among other criteria).

The meta-analyses by Gao, Yang, Yu, and Cui (2015) and Jaspers et al. (2016), examining efficacy and risks associated with flibanserin treatment, are excellent examples of variability in the strengths and weaknesses of meta-analysis, specifically "the file-drawer effect" to which Balon and Seagraves were referring. Gao et al. (2015) restricted their sample to four published RCTs and reported a standardized mean difference in SSEs of .59. Jaspers et al. (2016) replicated this effect using the same sample of published studies with an additional published study in which only women who showed improvement in an open-label phase were retained and randomized to treatment or placebo (Goldfischer et al., 2011). Jaspers et al. addressed publication bias concerns by including three unpublished studies; when these were added to the five published studies, overall the improvement in SSEs per month dropped from .58 to .49. The Jaspers et al. meta-analysis also addressed other concerns, including evidence quality (efficacy and safety), and use of SSEs as outcome variables, concerns also raised by Clayton and Pyke. In an exchange of commentaries about the Jaspers et al. meta-analysis, Laan, Jaspers, and Leusink (2016) succinctly stated what we believe should be the guiding principle of assessing treatment

efficacy and risk relating to any treatment: "We agree with Goldstein et al. that it is a clinician's task to diligently and routinely help patients to evaluate benefits, risks, and appropriateness of therapeutic options. When available, meta-analyses, not clinical opinion, should be the basis of such a risk and benefit analysis." (p. 1404).

Combining psychological and pharmaceutical approaches to treating sexual desire difficulties

Both Balon and Segraves (2016) and Clayton and Pyke (2016) suggested that future research could focus on direct comparisons between pharmaceutical and psychological treatments, or combination treatments for low desire in women. At this critical juncture of the clinical science on treatment of women's sexual difficulties, we welcome the opportunity to directly compare pharmaceutical and psychological treatments for low sexual desire. We are, however, cautious about assumptions that combination treatments may show additive benefits. Combination approaches (pharmacological + psychotherapy) can show greater efficacy in the short term; however, long-term follow up data have suggested that treatment benefits may not be retained. In some cases, (e.g., treatment of anxiety disorders) pharmacological treatment combined with psychotherapy can even be harmful, resulting in greater probability of relapse. In their JAMA publication, Barlow, Gorman, Shear, and Woods (2000) showed that relapse rates for panic disorder were higher for individuals receiving combined imipramine and CBT versus those receiving CBT and placebo, despite both combination treatments showing relatively similar efficacy in the short term. One interpretation is that, in the case of treating anxiety disorders, combination approaches impede learning and implementation of psychological

techniques that are necessary to preventing relapse because the pharmaceutical agent prevents full experience of the anxiety symptoms. Without the opportunity to fully experience these symptoms, patients are unable to disconfirm inaccurate threat perceptions regarding the consequences of their anxiety – the theorized core maintenance factor of anxiety disorders like panic disorder (e.g., Foa & Kozak, 1986). In other words, the medication is effective in treating the symptoms of the disorder, but may actually impede improvement in the core causes of the disorder. As a result, once medication use is terminated, the full experience of anxiety symptoms may lead to relapse.

If we extend these lessons to the treatment of sexual difficulties, we can forecast that women may become dependent on a medication with very modest benefits instead of learning new ways of cultivating sexual desire in their current context. Indeed, a goal of psychotherapy is that very few of our clients will return for long-term maintenance treatment because they, in turn, become their own therapists, able to identify challenges and implement effective solutions learned in treatment. Or, in the case of mindfulness-based approaches, learning acceptance leads to reduced distress, which ultimately paves the way for improved sexual desire when such inhibitions are removed. When treated with medication only, women do not learn acceptance and coping skills, beyond purchasing and consuming a medication. If that medication is combined with psychotherapy, women may not learn to cope with the symptoms or contexts that brought them in to treatment in the first place because the medication (partially) removes those symptoms via its direct action or via placebo effect. These are, of course, testable hypotheses that remain to be supported by data (or not). In summary, combination therapies for low desire in women may not offer the robust resolution of symptoms that Balon and Segraves suggested they might.

In their reflections on tailoring treatments, Balon and Segraves (2016) commented that, "One can also ponder if the extra effort spent on learning specific psychological interventions is necessary, given the power of non-specific interventions in supportive psychotherapy." Indeed, this is another empirical question. As later noted by Balon and Segraves (2016), and by ourselves in our initial commentary (Brotto et al, 2016), individualized treatment may be a meaningful way forward, recognizing the idiosyncratic factors that make treatments more or less beneficial for certain individuals. In medicine, individualized treatment considers factors like genetic complement, that is, whether the individual possesses certain genotypes associated with a disorder. In psychological treatments, we might also consider psychology traits, or phenotypes, that are associated with particular presentations of a disorder. A very straightforward example would be using a couples-based approach for women with sexual partners, versus other approaches for women without partners. To take this a step further, if one identified that an individual's sexual symptoms were associated with catastrophizing cognitions, incorporating elements of a CBT approach to recognizing and appraising catastrophic thinking and its impact on sexual response would be a likely approach. In the case of mindfulness, individuals who report disconnection with their physical sexual response, or difficulty remaining in the moment during sexual activities, might be well-served to learn techniques that enhance capacity to integrate physical sensation in to awareness and focus on the moment. We doubt that supportive therapy alone will be a one-size-fits-all solution, but this is an empirical question we look forward to addressing.

In their comments, Balon and Seagraves (2016) stated, "One would expect transient or situation specific problems to be more responsive to psychological interventions and more global

persistent problems to be more responsive to pharmacological interventions." The notion that transient sexual concerns are more amenable to psychological versus pharmacological treatment, and that longer term concerns warrant medical intervention, is a problematic shorthand for a more thorough case conceptualization. For example, long term, global deficits in sexual desire may have etiological roots in psychological phenomena, such as sexual trauma, absence of early sex education, anxiety in first sexual encounters, or other early influences that shape a person's sexuality. Long-term difficulties can also become compounded by other sexual dysfunctions in one or both partners. For example, sexual interest/arousal disorder and situational erectile dysfunction can arise subsequent to a couple grappling with chronic dyspareunia such that a given medication is unlikely to be sufficient. On the other hand, shorter term or situation-specific sexual complaints may also be related to biological, psychological, cultural, or any combination of those factors. Thus, assumptions such as those voiced by Balon and Segraves (2016) may result in ineffective or worse -- iatrogenic -- treatments. We discourage clinicians from assuming any simple shorthand for treatment and, instead, adopt a case formulation informed by a thorough patient history, including evaluation of present and past sexual experiences and sexual relationships, the developmental history, medical and psychiatric history and current status as well as aspects relevant according to culture, past and present stressors and personality factors. For example, the "three windows approach" proposed by Bancroft (2009) provides a useful framework for assessing situational/context factors, individual vulnerability factors, and medical/health factors. The diagnosis and its formulation (a combination of likely etiological factors), are then explained to the patient and preferably also to the partner who has been similarly evaluated. Then therapy begins, if it is needed over and beyond this assessment that, in

itself, can often be highly therapeutic. We must underscore, however, that this assessment is not intended to stand in for supportive therapy.

On the cultural context of sexual medicine and sex research and remaining mindful of bias

We deeply appreciate Tiefer's (2016) reminder that the conditions we are attempting to treat are constructs located within a specific cultural and historical context. Classification and quantification of deficits in desire brings to mind Wakefield's (1992) rubric for assessing disorder as "harmful dysfunction." Although this conceptualization seems, at first blush, to provide a necessary scaffolding for determining what is and isn't a disorder (harm judged within a social context; dysfunction as departure from a system's intended/evolved function), it doesn't take much scrutiny to see that we have little insight into the "true" design and function of the sexual response system beyond sexual pleasure and reproduction. Harm, or distress as it is currently phrased in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2013), is also multiply determined, and engages the many cultural factors that impinge upon the experience and expression of our sexualities.

Tiefer's (2016) incisive reminder that all discourse in sexuality, including treatment of sexual concerns, is systematically biased by the politics of the investigators and commentators is also well received. Although we may attempt to be unbiased in our critiques, the evidence we accrue, and the methodologies we employ, will ultimately always be influenced by the individual or group or culture and by their historical and social location. We would all be well served to check our biases at the door when making any pronouncements about what is correct, natural, biological, psychological, functional, and normal.

Tiefer (2016) pointed to qualitative approaches to understanding women's experiences as one possible means of accessing meaningful subjective experience and circumventing concerns about artificial constructs of sexual desire. Although qualitative methodologies offer an alternative approach to understanding individual experience, these data are ultimately being gathered and interpreted by people whose politics also shape decision-making, from the questions that are asked, to the themes that are extracted, to the inclusion of "lived experiences" that are typically highlighted in the first person voices included in qualitative papers. Well-trained and ethical researchers question these decisions and biases at every turn, querying the validity and reliability of their methods, their approach data to analysis, and their interpretation of effects. They also make their biases known and transparent.

Concluding remarks

If the small effects reported for flibanserin treatment are independently replicated, we will have more evidence for very modest and limited efficacy, and perhaps a better understanding of the factors associated with women's sexual desire concerns. Whether effectiveness will remain in the less controlled approach to treatment typically observed outside clinical trials, however, remains to be seen. Although the very modest demand for flibanserin may be entirely be due to the fact that Valeant is not yet permitted to engage in direct-to-consumer marketing, we predict that the small effect of an additional one-half SSE per month with flibanserin treatment will likely become diluted by a host of third variable concerns, including the total contraindication with alcohol use even when sexual activity is infrequent. As Balon and Segraves (2016) noted, this dilution will likely be attributable to powerful placebo effects. As clinical researchers, our job is to disentangle effects attributable to treatment from the

background noise of placebo effects that, in some cases, may also have very valuable lessons to impart regarding factors influencing women's sexual desire (see Bradford & Meston, 2009).

No research on any psychological or biological phenomenon is without bias in its interpretation, which is precisely why we have engaged in this dialogue. Shedding false dichotomies such as biological versus psychological causation, and its cousin, pharmacological versus psychological treatment, is among the first steps to appreciating the multifactorial determinants of the conditions we seek to ameliorate. Biology, psychology, and culture are so deeply intertwined in the etiology, symptom presentation, treatment seeking, and so on, that it would be a fools errand to try to disentangle and represent them as dichotomous.

Further Resources

http://www.psychologicalscience.org/index.php/members/new-statistics

https://youtu.be/iJ4kqk3V8jQ

http://rpsychologist.com/d3/NHST/

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th. ed.). Arlington, VA: Author.

Bancroft, J. (2009). Human sexuality and its problems. (3rd. ed.). Edinburgh: Churchill Livingstone.

Balon, R., & Segraves, R. T. (2016). Which emperor has new clothes? Biology versus psychology in the era of statistical magic. *Journal of Sex & Marital Therapy*.

Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2000). Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. *JAMA*, *283*, 2529-2536.

Bradford, A., & Meston, C. M. (2009). Placebo response in the treatment of women's sexual dysfunctions: A review and commentary. *Journal of Sex & Marital Therapy*, *35*, 164-181.

Brotto, L. A., Basson, R., Chivers, M. L., Graham, C. A., Pollock, P., & Stephenson, K. R. (2016). Challenges in designing psychological treatment studies for sexual dysfunction. *Journal of Sex & Marital Therapy*.

Clayton, A. H., & Pyke, P. E. (2016). RESPONSE TO COMMENTARY. NEED CITATION.

FDA Flibanserin Briefing Document, 2015: http://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/drugs/drugsafetyandriskmanagementadvisorycommittee/ucm449090.pdf

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin, 99,* 20-35. doi:10.1037/0033-2909.99.1.20

Frühauf, S., Gerger, H., Schmidt, H. M. , Munder, T., & Barth, J. (2013). Efficacy of psychological interventions for sexual dysfunction: A systematic review and meta-analysis. *Archives of Sexual Behavior, 42,* 915-933.

Gao, Z., Yang, D., Yu, L., & Cui, Y. (2015). Efficacy and safety of flibanserin in Women with hypoactive sexual desire disorder: A systematic review and meta-analysis. *The Journal of Sexual Medicine, 12,* 2095-2104. doi: 10.1111/jsm.13037

Goldfischer, E. R., Breaux, J., Katz, M., Kaufman, J., Smith, W. B., Kimura, T., ... & Pyke, R. (2011). Continued efficacy and safety of flibanserin in premenopausal women with hypoactive sexual desire disorder (HSDD): Results from a randomized withdrawal trial. *The Journal of Sexual Medicine*, *8*(11), 3160-3170. doi: 10.1111/j.1743-6109.2011.02458.x

Hofmann, S. G., Wu, J. Q., & Boettcher, H. (2014). Effect of cognitive-behavioral therapy for anxiety disorders on quality of life: A meta-analysis. *Journal of Consulting and Clinical Psychology, 82,* 375-391. doi:10.1037/a0035491

Jaspers, L., Feys, F., Bramer, W. M., Franco, O. H., Leusink, P., & Laan, E. T. (2016). Efficacy and safety of flibanserin for the treatment of hypoactive sexual desire disorder in women: A systematic review and meta-analysis. *JAMA Internal Medicine*, *176*, 453-462.

Laan, E. T., Jaspers, L., & Leusink, P. (2016). Appropriate perspective and context for newly approved medications, including flibanserin—Reply. *JAMA Internal*

Medicine, 176, 1404-1405.

Murphy, M. J., Mermelstein, L. C., Edwards, K. M., & Gidycz, C. A. (2012). The benefits of dispositional mindfulness in physical health: A longitudinal study of female college students. *Journal of American College Health, 60,* 341-348. doi:10.1080/07448481.2011.629260

Pyke, R. E., Clayton, A. H. (2015). Psychological treatment trials for hypoactive sexual desire disorder: A sexual medicine critique and perspective. *Journal of Sexual Medicine, 12,* 2451-2548.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26,*1827-1832.

Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279-282.

Tiefer, L. (2016). Apples and oranges: "Sexual Medicine" and the effort to deny that counting and classifying are political acts. *Journal of Sex & Marital Therapy*.

Wakefield, J. C. (1992). Disorder as harmful dysfunction: a conceptual critique of DSM-III-R's definition of mental disorder. *Psychological Review, 99,* 232.

