Committee 21

Standards for Clinical Trials in Sexual Dysfunction in Women: Research Design and Outcomes Assessment

Chairperson:
ANITA H. CLAYTON

Members:
LORRAINE DENNERSTEIN
WILLIAM A. FISHER
SHERYL A. KINGSBERG
MICHAEL A. PERELMAN

Advisor:
ROBERT E. PYKE
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>INTRODUCTION</th>
<th>VI. OBTAINING A REPRESENTATIVE STUDY POPULATION – ISSUES IN GENERALIZABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. REGULATORY CONSIDERATIONS</td>
<td></td>
</tr>
<tr>
<td>1. REGULATORY GUIDELINES: FOOD &amp; DRUG ADMINISTRATION (FDA)</td>
<td>1. SAMPLING A REPRESENTATIVE STUDY POPULATION</td>
</tr>
<tr>
<td>2. EMEA</td>
<td>2. ADDITIONAL CONSIDERATIONS IN SAMPLE SELECTION</td>
</tr>
<tr>
<td>3. MEASURES</td>
<td>3. NATURE OF CONTROL GROUP</td>
</tr>
<tr>
<td>4. PROBLEMS WITH CURRENT GUIDELINES</td>
<td></td>
</tr>
<tr>
<td>5. STUDY DURATION</td>
<td></td>
</tr>
<tr>
<td>6. SAFETY OF LONG-TERM TREATMENT AND WITHDRAWAL</td>
<td></td>
</tr>
<tr>
<td>II. CHOICE OF STUDY MEASURES AND ENDPOINTS</td>
<td>VII. ISSUES IN STATISTICAL ANALYSIS</td>
</tr>
<tr>
<td>1. PROBLEMS WITH CURRENT ENDPOINTS: SEXUAL PSYCHOPHYSIOLOGY.</td>
<td>1. PLACEBO-CONTROLLED TRIALS</td>
</tr>
<tr>
<td>2. SEXUAL DISTRESS</td>
<td>2. DOSE-FINDING</td>
</tr>
<tr>
<td>III. OTHER ISSUES IN FSD ASSESSMENT</td>
<td>3. DURATION OF TREATMENT</td>
</tr>
<tr>
<td>IV. RECOMMENDATIONS FOR STUDY MEASURES AND ENDPOINTS</td>
<td>4. EFFICACY MEASURES AND THE CURRENT PROCESS AND PROBLEMS</td>
</tr>
<tr>
<td>V. DEFINITIONS OF THE SPECIFIC DISORDERS WITH APPLICABILITY TO THE RESEARCH QUESTIONS</td>
<td>5. SAFETY MEASURES AND ADVERSE EVENT MONITORING</td>
</tr>
<tr>
<td>VI. OBTAINING A REPRESENTATIVE STUDY POPULATION – ISSUES IN GENERALIZABILITY</td>
<td></td>
</tr>
<tr>
<td>VII. ISSUES IN STATISTICAL ANALYSIS</td>
<td></td>
</tr>
<tr>
<td>VIII. ADDITIONAL ISSUES IN FSD TRIALS</td>
<td></td>
</tr>
<tr>
<td>VIII. CONCLUSIONS</td>
<td></td>
</tr>
</tbody>
</table>

Financial Disclosures:
Anita H. Clayton, M.D.
Advisory Board/consulting: Boehringer-Ingelheim, Eli Lilly & Co., Concert Pharmaceuticals, Takeda, PGxHealth, TransTech Pharma, Inc., Wyeth
Grants: Biosante Pharmaceuticals, Boehringer-Ingelheim, Bristol Myers Squibb, Eli Lilly & Co., Novartis, Pfizer, Inc., Sanofi-aventis
Honorarium: Eli Lilly & Co., Sanofi-aventis, New England Research Institutes,
Royalty: Ballantine Books/Random House, Guilford Publications
Lorraine Dennerstein, M.D.
Advisory Board/consulting: Boehringer-Ingelheim
Honorarium: Boehringer-Ingelheim, Bayer Schering

William Fisher, Ph.D.
Advisory Board/consulting: Bayer, Boehringer-Ingelheim, Merck
Sheryl Kingsberg, Ph.D.
Consulting: Boehringer-Ingelheim, Tivamed, Wyeth
Honorarium: Boehringer-Ingelheim, Eli Lilly & Co., Johnson & Johnson
Michael Perelman, Ph.D.
Advisory Board: Boehringer-Ingelheim
Honorarium: Johnson & Johnson, Lilly USA, Palatin, Plethora, TransTech Pharma, Sempre
Robert Pyke, Ph.D.
Salary: Boehringer-Ingelheim
Only one pharmacological product has been approved for Female Sexual Dysfunctions (FSD) in the European Union and none in the United States since the last International Consultation (2004). This appears to be related more to the standards for clinical trials in women than to the efficacy of the drugs themselves. The focus of this chapter is on essentials of research design and measurement in female sexual dysfunctions. Accordingly, our recommendations rest upon classical procedures in experimental design and classical measurement theory. Therefore, consideration of rating the quality of level of evidence is not germane for this chapter.

As part of that process, we will review current recommendations, evaluate their appropriateness for regulatory approval, and discuss discrepancies and potential changes that may improve ability to demonstrate efficacy for a specific construct.

I. REGULATORY CONSIDERATIONS

1. REGULATORY GUIDELINES: FOOD & DRUG ADMINISTRATION (FDA)

Introduction of a new device or drug requires regulatory approval by a designated government entity. Unfortunately, as of 2009 only one regulatory body, the FDA, has issued any guidance on standards for clinical trials in women with sexual dysfunction. The FDA’s Draft Guidance of May 2000 (FDA-DG or DG, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071655.pdf) outlines Female Sexual Dysfunctions as “currently” consisting of the same four recognized components as in DSM-IV, decreased sexual desire, decreased sexual arousal, persistent difficulty in achieving or inability to achieve orgasm, and dyspareunia (no broader or more detailed definition for sexual pain syndromes). The FDA-DG states that “the definition of FSD continues to evolve” while not recognizing the Diagnostic and Statistical Manual (DSM-IV) criteria for these disorders. Similar to DSM-IV, however, it specifies that, for a diagnosis of FSD, “these components must be associated with personal distress, as determined by the affected woman.” Unlike DSM-IV, it does not recognize interpersonal difficulties as an alternative requirement, to personal distress.

The FDA-DG allows sponsors to target for treatment development any one of, or combination of, the four named components. It suggests some exclusion criteria “to increase the likelihood of demonstrating a treatment effect in clinical trials of FSD.”

- Relationship difficulties with a sexual partner
- Use of concomitant medications that could affect sexual function (e.g., certain antidepressant medications)
• The presence of medical conditions that could affect sexual function (e.g., depression, anxiety, the first few weeks of the postpartum period)
• Sexual dysfunction on the part of the woman’s partner

2. EMEA

The European Medical Authority (EMEA), while not issuing any guidance on FSD, has in the documentation of its approval of a testosterone transdermal system (TTS) (European Public Assessment Report (EPAR), 2006, effectively established precedents for standards of treatment development in hypoactive sexual desire disorder (HSDD). (www.emea.europa.eu/humandocs/PDFs/EPAR/intrinsa/063406en1.pdf) The EPAR details the discriminant validity of the sponsor’s proprietary patient-based measures of sexual desire (desire subscale of the Profile of Female Sexual Function© [PFSF©] and distress related to a deficiency of desire on the Personal Distress Scale© [PDS©]), in comparisons of post-menopausal women with HSDD to those without FSD, and then focuses on three measures of HSDD as leading to its acceptance of the TTS: satisfying sexual events, sexual desire as measured in the PFSF, and distress related to sexual desire as measured on the PDS. These measures remain proprietary and unavailable in sufficient detail for use except by the sponsor of the TTS, Procter and Gamble.

3. MEASURES

Satisfying Sexual Events (SSE): The DG recommends recording the number of sexual events/encounters daily “using diaries” for 4 to 8 weeks before treatment is initiated, and collecting such information weekly for the duration of the intervention. A count of satisfying sexual events is to be the primary endpoint. Satisfaction is to be determined by the patient herself. Partner diaries are allowed but not recommended as primary endpoints.

Sexual Distress: The Draft Guidance recommends measuring personal distress to ensure appropriate patient selection for trial participation but states that this dimension of FSD “should not serve as the primary endpoint for establishing effectiveness.”

Other Measures: Changes in vaginal/genital physiology are not accepted as main endpoints and must be linked to satisfying sexual events to be accepted. Neither is Health Related Quality of Life to be used as a primary endpoint.

Patient Reported Outcomes (PRO) Measures: The DG recommends that any new measures developed be thoroughly validated prior to use in definitive clinical trials. The USA Food and Drug Administration (FDA) issued its draft guidance in 2006 on PRO measures used as effectiveness endpoints in clinical trials (http://www.fda.gov/CDER/GUIDANCE/5460dff.pdf). [1] A PRO is a measurement of any aspect of a patient’s health status that comes directly from the patient. Thus data generated by a PRO measure can provide evidence of a treatment benefit from the patient perspective. PRO instruments need to demonstrate ability to measure the claimed treatment benefit and to be specific to the intended population and the characteristics of the condition or disease treated. The advent of this guidance had specific benefit to the field of female sexual dysfunction. Previously the FDA had relied on observable outcomes, such as daily calendar reports of intercourse frequency. It can be argued that intercourse frequency, relying on the relationship with a partner, may not directly reflect the domain of female sexual function that a clinical intervention seeks to change. The FDA guidance document notes that PRO-based evidence of improvement in symptoms may not be sufficient to substantiate a claim which may also need to demonstrate how any change in (for example sexual desire) translates into other specific endpoints (such as intercourse frequency).

The guidance recognized that for some treatment effects the patient is the only source of data. This, for example, applies to sexual desire where there are no observable physical measures of this concept. Self-completed questionnaires by patients directly capture the patient’s baseline and perceived response to treatment and are not affected by inter-observer variability. However, the PRO may be affected by inter-patient variability if the instrument is not easily understood and completed by patients. PROs may also be developed to define entry criteria to study populations, and evaluate adverse events.

The adequacy of the PRO instrument as a measure to support efficacy claims depends on the development of the PRO and demonstrated measurement properties. The FDA encourages sponsors to determine whether an adequate PRO exists to assess and measure concepts. If it doesn’t then a new PRO can be developed. A single item PRO can on occasion provide a reliable and valid measure (e.g. pain severity). Multiple domains are usually necessary for general concepts. Documentation is needed of the instrument development process, revealing the means by which domains were identified and named, how individual items are associated with each other, associated with each domain, and how domains associate with each other and the general concept of interest. In some measures domains can be aggregated into an overall score. The FDA will compare the patient population used in the instrument development process to the study populations enrolled in clinical trials to determine whether the instrument is appropriate to that population with regard to patient age, sex, ethnic identity and cognitive ability.
The guidance specifically states that the PRO instrument generation process is incomplete without patient involvement. Item generation generally incorporates the input of a wide range of patients with the condition of interest to represent appropriate variations in severity and patient characteristics such as age or sex. Adequate numbers of patients are needed to support the opinion that the specific items and instrument are adequate and appropriate to measure the concept. Detail is needed regarding item generation techniques used, including theoretical approaches used, population study, pools of items, selection and reduction of items, cognitive debriefing interviews, pilot testing, importance ratings and quantitative techniques for item evaluation such as factor analysis and item-response analysis. The method of data collection and all procedures and instructions and modes of administration must also be detailed. The rationale and appropriateness of the recall period must be established and measures taken to ensure that entries are made into patient diaries according to the study design. Response options need to be adequate, appropriate, associated with clear instructions and avoid potential ceiling or floor effects as well as not biasing the direction of response.

The FDA will review PRO instruments for reliability, validity, ability to detect change and interpretability. The socio-demographic and medical characteristics of any sample used to develop PRO instruments should be appropriate for future clinical study settings.

Measurement properties to be reviewed in clinical trials include:

- **reliability** (test - retest; internal consistency; inter-interviewer reproducibility for interviewer administered PROs only)
- **validity** (content - related; construct - related; predictive validity; discriminant validity)
- **ability to detect change** (calculation of effect size and standard error of measurement)
- **interpretability** (smallest difference that is considered clinically important such as minimum important difference; responder definition).

Modification of an existing instrument requires additional validation studies to confirm the adequacy of the modified instrument’s measurement properties. When an instrument developed in one language or culture is adapted or translated for use in another language and culture, then evidence is needed that the translation processes were adequate to ensure the validity of responses.

In summary, the FDA guidance on PROs represents an advance for the field of clinical trials for sexual dysfunction. The guidance clearly states that the PRO may be used as an endpoint, elevating the patient’s perception to that of an observable outcome. The FDA guidance has also spurred further development of instruments. The method outlined by the FDA does help ensure that instruments will adequately reflect the concepts being measured in any given population. However, the necessity for an integration of that guidance, which also incorporates important recommendations/corrections from sexual medicine professionals, is still awaited. [2,3]

4. PROBLEMS WITH CURRENT GUIDELINES

This Draft Guidance remains unaltered despite much criticism from many academic experts in the field of FSD. These objections are of four main types.

One criticism is to the primacy of SSE despite its absence from the criteria for the various forms of FSD recognized in practice and by DSM-IV. While a decrease in SSE is considerably “downstream” in consequences of the DSM-IV-recognized symptoms defining FSD, in fact, the rate of SSE is markedly decreased in the various types of FSD [4,5,6]. The mean decrease is about 70 to 80% in the various Boehringer-Ingelheim (BI) samples compared, with wide variation from patient to patient but with approximately 70-75% sensitivity and specificity vs women with no FSD. D. Shames, Director of the involved FDA division has stated that the primacy of SSE was established so that improvement could not be claimed based solely on subjective feelings of the subject/patient. It also serves to “level the playing field” between the various types of FSD in terms of the main standard set for proof of efficacy.

A second objection is to the absence of the DSM-IV recognized symptoms of FSD as primary outcome variables, e.g., improvement of sexual desire for women with HSDD. The FDA took the letter particularly into account in its presentation in response to the International Society for the Study of Women’s Sexual Health (ISSWSH) in 2006 and published the next year. [7]. Measuring sexual desire is explicitly recommended as a main endpoint for clinical trials of HSDD, but not to the exclusion of SSE as another primary measure.

Thus, it would not be unreasonable for sponsors to seek claims in terms of the main symptom that defines each of the other types of FSD, e.g., a measure of sexual arousal for women with sexual arousal disorder. However, such measures will apparently not be recognized by the FDA unless they meet the validation criteria recommended by the FDA in its Draft Guidance on Patient Reported Outcomes (PRO) of February 2006. Some measures meet these criteria, but not all are available for use in clinical trials. A path to validate currently available measures is also given which has been followed for at least one measure of a primary symptom of FSD, i.e., the desire domain of the Female Sexual Function Index [8,9,10,11].
A third objection is to the exclusion of distress as a main endpoint, but the FDA has not in practice excluded a well-validated measure of personal distress related to FSD as either a co-primary or a secondary endpoint.

A fourth objection is to the FDA emphasis on daily recording of symptoms of FSD, e.g., of sexual desire in women with HSDD. Recent evidence shows that women with HSDD do not recognize a 24-hour retrospective period as valid for assessing their desire, and that instead they find one to four weeks as a meaningful period for retrospective assessment. The finding was the same for pre- and post-menopausal samples. [12, 13]

Given these criticisms, there could be value in an iterative process involving coordination between the two regulatory agencies, FDA and EMEA, and including external input.

5. STUDY DURATION

The FDA’s draft guidance requires a six-month treatment period to establish efficacy; the two TTS studies that led to EMEA approval were of this duration, so a six-month treatment period is now the recognized transatlantic standard.

Objections can be raised to this, too. Surely it would be a rarely persistent patient who would comply with a new treatment, its inconvenience and side effects if she felt no improvement until six months had passed, leading to considerable likelihood of selective attrition of placebo-arm patients from clinical trials. Thus, a practical standard for onset of action should be recognized as closer to 1-2 months than 6 months. A practical standard for proof of efficacy would inform a similarly short duration of no longer than 12 weeks, as in typical clinical psychopharmacology indications (depression, anxiety, etc.). This would help not only to conform to practical considerations in routine clinical practice but also to separate the issues related to treatment response from those related to maintenance of efficacy and to long-term remission and recurrence. Three-month extension trials (following a 12-week acute efficacy trial), relapse prevention trials, long-term safety studies (when indicated, such as for hormonal therapies), and trials involving a comparator agent (when available) would add to drug information, but do not appear necessary for regulatory approval.

This recommendation is not for a relaxation of regulatory standards, but to establish standards for the separate goals of demonstration of induction of effect and for proof of maintenance of efficacy.

6. SAFETY OF LONG-TERM TREATMENT AND WITHDRAWAL

Data are accumulating that FSD are highly chronic. The mean duration of HSDD was over 5 years at entry into the fibanserin phase III trials of premenopausal women with HSDD. [4] Thus, determination of long-term efficacy and safety, including the safety of treatment withdrawal, should also be assumed to be regulatory concerns.

With the announcement of two one-year studies and a large two-year controlled study of 549 women, (some on placebo) of TTS for postmenopausal women with HSDD, http://clinicaltrials.gov/ct2/results?term=Intrinsa+female, it might be assumed that concerns for the safety of long-term hormonal therapy in postmenopausal women expressed by the FDA in 2004 (http://www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4082B1_02_A-FDA-Intrinsa-Overview.htm) have led to a recommendation for studies of this size and duration.

Similarly, with publications arising from the fibanserin ROSE study at the European Society for Sexual Medicine – International Society for Sexual Medicine (ESMS-SSM) meeting in Brussels in December 2008 and the International Society for the Study of Women’s Sexual Health in February 2009, it is reasonable to assume that one or more regulatory agencies have recommended that sponsors perform a clinical study sufficient to determine the effects of drug withdrawal after six months of treatment for FSD (at least in HSDD and based on CNS activity) http://www.em-consulte.com/article/175759, http://www.doguide.com/news/content.nsf/news/852571020057CCF68525751A005D6357.

Finally, additional guidance from other government regulatory agencies was not available at the time this chapter was written in 2009.

II. CHOICE OF STUDY MEASURES AND ENDPOINTS

1. PROBLEMS WITH CURRENT ENDPOINTS: SEXUAL PSYCHOPHYSIOLOGY.

Sexual psychophysiology is an emerging field of investigation with the number of publications increasing 10-fold over the last 30 years.[14] Sexual psychophysiology can be defined as the application of psychophysiological methods to the study of sexual arousal, with special emphasis on the interplay between subjective (cognitive and affective) and physiological determinants of sexual arousal.[15] A common finding in psychophysiological research is that correlations between subjective and genital responses are lower in women than in men.[16] The discordant pattern most frequently found in women is that genital responses occur while subjective sexual arousal is low or absent.

As well, there has been comparatively little research on the effects of the experimenter on results. A number of studies have explored differences
between volunteers and non-volunteers for sexuality studies and found differences in sexual experience, frequencies of sexual activity, sexual guilt, exposure to erotic materials, and sexual attitude. These findings suggest that psychophysiological studies are particularly susceptible to volunteer bias. Thus, psychophysiological measures cannot currently be recommended as suitable endpoints for clinical trials, consistent with the FDA guidance.

Rellini and Meston found that only the self-report questionnaire, the FSFI, significantly predicted which women with female sexual arousal disorder improved at post treatment. Event logs, vaginal photoplethysmography, and continuous subjective sexual arousal measured during exposure to erotic videos did not demonstrate or predict women's improvement at post-treatment. Nevertheless, these results provide strong evidence that validated self report questionnaires such as the FSFI are the most sensitive endpoints to detect treatment-induced changes in women's sexual dysfunction.

There is little correlation between other biologic measures such as sex steroids, specifically androgens, and self-report measures of sexual functioning. Low levels of androgens, specifically DHEA-S, have been linked to low sexual desire in only one study; however, treatment of distressing low desire with testosterone in women without demonstrated low androgen levels did increase desire over placebo. Elevated levels of sex hormone binding globulin (SHBG) appear to mitigate the increase in desire associated with testosterone treatment, and may be used as an exclusion criteria in studies. Peripheral measures of neurotransmitters have not been shown to accurately reflect central levels, nor to be correlated with sexual dysfunctions or response to non-hormonal, centrally-mediated treatments.

2. SEXUAL DISTRESS

Current definitions of sexual dysfunction include criteria that there is deficiency in a parameter of sexual function and that this causes personal distress or interpersonal difficulty (DSM-IV). Objective and subjective measures of domains of sexual function have been the focus of study for some decades. A number of instruments which measure the low sexual function component have been validated and have been translated into different languages. However these measures did not include items about personal distress. Consensus in definitions has led, in the last decade, to the development of measures of distress. When both components (low sexual function and sexual distress) are measured in epidemiological studies, prevalence of sexual dysfunction is greatly reduced. However, there is a differential effect of age on aspects of sexual function and sexual distress. For example with increasing age (and reproductive aging) the proportion of women who experience low desire increases, but the proportion of women distressed by low desire declines. Thus, if low desire and sexual distress are combined into the aggregate definition of hypoactive sexual desire disorder, no association with age is detected. Recent Australian epidemiological studies have investigated the relationship of psychological, socioeconomic, physiological and relationship factors to low sexual function and sexual distress components of FSD. In a cross-sectional population based study of 1002 women aged 20 to 70, utilizing the Sexual Function Questionnaire (SFQ) and the Female Sexual Distress Scale (FSDS), relationship factors were found to be more important to low desire than age or menopause, whereas physiological and psychological factors were more important to low genital arousal and low orgasmic function. Sexual distress was associated with both psychological and relationship factors. Sexual distress was positively associated with depression and inversely associated with better communication of sexual needs when results were adjusted for age, and other factors. In an 11 year, longitudinal study of mid-aged women using the Short Personal Experiences Questionnaire, a decline in all aspects of sexual function was reported over the duration of the study. Sexual distress (measured by FSDS only once, postmenopausally) was associated with higher depression scores and negative feelings for the partner. Sexual distress was predicted by prior negative feelings for the partner and a greater decline in total scores of sexual function. Sexually-related distress can be assessed by clinician interview and by patient self-report. Reliability and validity have been consistently demonstrated with the FSDS and PDS as a measure of the central nervous system construct of distress. Measurement of distress as an efficacy measure has been demonstrated.

III. OTHER ISSUES IN FSD ASSESSMENT

The primary focus by regulatory agencies on event measures rather than on whether the construct is comprehensively conceptually appropriate to the specific sexual disorder is based on a desire for quantitative behavioral outcome measures. For example, questions about maximum intensity of sexual desire and frequency of sexual desire collected on a daily basis may not be conceptually relevant to HSSD; neither the daily time frame nor the measurement of intensity are closely linked with the construct of HSSD. Similarly, definition of successful and satisfactory sexual events or encounters over time requires a subjective interpretation of sexual experiences by the woman potentially including sexual intercourse, oral sex, and partner- or self-stimulation, and other associated outcomes such as orgasm and...
emotional intimacy. Data clearly show other contextual factors influence reporting such as relationship status in sexual desire, and physiological and psychological factors in genital arousal and orgasmic function, that might need to be controlled-for in criteria for exclusion and inclusion in clinical trials, ultimately affecting generalizability. Such data seem to be even more difficult to interpret than numerical scores on self-report questionnaires, which have been established in control subjects and in target populations. Particularly in women with HSDD, the construct of low desire is only indirectly related to the number of sexual events, satisfying or not, as most if not all sexual events in women with HSDD are partner-initiated. Counting of satisfying sexual events reflects a behavioral outcome, rather than a biological or cognitive event (sexual desire), and may not correspond to the sexual problem about which the woman is distressed. [38] Attempts to quantify the minimum important difference (MID) for an SSE diary in 788 women with HSDD using anchor- and distribution-based estimates suggest a range of mean MID estimates between 0.04 and 0.46 SSEs per week. [39] Measurement of SSEs also fails to count other conceptually appropriate events such as receptivity to partner approach and subject initiation of sexual activity.

In addition, the FDA draft guidance for clinical development of drug products for treatment of FSD (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071655.pdf) requires recording of these events in daily diaries which do not have either face validity (daily measures in women of sexual desire or activity) nor have they been developed using the FDA's own guidance regarding PROs (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071975.pdf). Unfortunately, diaries are often inaccurate; this is particularly true of paper diaries that record behavioral events, which may be completed retrospectively without investigator knowledge, and thus, are subjected to recall bias which daily diaries are intended to avoid. [40] However, even when completed on time, for example, with electronic diaries, repeated daily measurement of the distressing sexual problem (e.g. low desire) may increase distress or may not accurately reflect the experience of the woman (e.g. she may not have previously experienced an awareness of sexual desire daily). Also, daily measures may themselves act as an intervention, potentially inflating treatment and placebo effects or magnifying distress.

Validated self-report questionnaires have been standardized in target populations, and permit privacy that may lead to more accurate reporting. (See report of Committee 6: Clinical evaluation and symptom scales: sexual dysfunction assessment) Structured interviews allow for follow-up questions that may better define the problem, but require training and monitoring to establish inter-rater reliability, and avoid clinician bias. Currently, standardized expert interviews have been developed by industry, but have not been psychometrically validated. Concomitant use of self-report measures and clinician evaluation may provide the most accurate assessment. Such an approach has been successful in demonstrating differences between study drug and placebo in other central nervous system-mediated disorders, such as major depression and pain disorders.

Partner assessments have been utilized in post-marketing erectile dysfunction (ED) and premature ejaculation (PE) studies, particularly analyzing the burden of disease, medication adherence, and response to treatment. A woman's sexual difficulties can be contingent on ED in her male partner, and PE in male partners has also demonstrated co-morbid sexual difficulties in the woman, and was associated with greater dissatisfaction and emotional distress. [41,42,43,44] No such studies have been published in women, but FSD would likely also be associated with dissatisfaction and distress in the partner. In addition, FSD trials usually exclude women whose partner has any reported sexual dysfunction. In planning studies to include partners of women with sexual disorders, care would need to be taken to avoid making desire discrepancy between partners into a sexual disorder, and to avoid using partner assessments as a study endpoint.

### IV. RECOMMENDATIONS FOR STUDY MEASURES AND ENDPOINTS

1. **Primary endpoints should directly measure the construct of interest sampled over an interval of time appropriate to the construct, through measures such as:**
   a. A **validated, self-report measure of sexual functioning** to include domains/subscales related to phases of the sexual response cycle (specifically sexual desire, arousal, orgasm and sexual pain – all processes mediated via the nervous system), and/or questionnaires or structured interviews assessing specific phases, and able to measure change over time with an intervention
   b. A **validated measure of sexual distress**

2. **Recall period should be 1 – 4 weeks; daily diaries are not recommended**

3. **Translations of self-report questionnaires must be linguistically/culturally validated with cross-cultural construct validation established.**

4. **Hormonal status should be considered to include separate evaluations/stratification of premenopausal vs. post-menopausal women, monitoring of menstrual cycle in premenopausal women, analysis of effect of hormonal therapies.**
5. Secondary endpoints may include physiologic measures, satisfying sexual events (this may be least important for HSDD), partner assessments, relationship satisfaction, etc.

V. DEFINITIONS OF THE SPECIFIC DISORDERS WITH APPLICABILITY TO THE RESEARCH QUESTIONS

All research designs in sexual dysfunction clinical trials must have participant inclusion and clinical trial outcome measures that reflect DSM IV-TR sexual dysfunctions. These have been outlined by Committee 2.

With regard to the DSM-IV-TR and its relevance to clinical trials, one of the major goals of the DSM-IV-TR is to allow for the differentiation between sexual complaints (short-lived, transient disruptions) from sexual dysfunctions (persistent problems that are distressing) and warrant treatment. The challenge remains to make more objective the essentially subjective line in the sand of what constitutes a complaint versus a dysfunction. Although Diagnostic and Statistical Manual for Diagnosis (DSM-IV-TR) [44] diagnoses and/or criteria are now often used uncritically by researchers and the legal profession, the DSM-IV-TR does not provide any indication of the huge variability in the level of empirical support for the reliability and validity of different diagnoses.

Similarly, the validity of the classification system is questionable given marked variations occurring across time (e.g., age, stage of life, etc.) and across cultures. Further, DSM-IV does not reflect differences in cultures when establishing norms of behavior and this remains a clinical judgment and may also be reflected in difficulties in validating clinical research across cultures.

There is a great deal of overlap or comorbidity of sexual disorders in women. The Female Sexual Function Index (FSFI), despite not initially meeting development criteria for a Patient Related Outcomes (PRO) measure [1,45], but often used as an outcome measure in clinical trials, shows high correlations between sexual constructs. For example, there was a surprisingly higher correlation between the FSFI arousal (psychological) and desire constructs than FSFI arousal and lubrication which are supposed to be measuring the same construct. [1,45] This also points to the notion that desire and arousal (psychological) may be the same construct in women or too interrelated to effectively tease out and measure independently. Clinical trials of sildenafil to treat Female Sexual Arousal Disorder (FSAD) are clear examples of the problem of overlap of disorders. These trials attempted to include postmenopausal women with only FSAD, but had a significant number of subjects who also met the criteria for HSDD or Female Orgasmic Disorder (FOD) (> 50% overlap) [46,47,48,49]. Some clinical trials of testosterone to treat HSDD in postmenopausal women also showed overlap between HSDD and FSAD or FOD with inclusion criteria that HSDD preceded the onset of other disorders. [20,21,22 23] Often in clinical and research settings, women are unable to differentiate the two constructs of desire and arousal, which begs the question—is it really the same construct we are measuring? More data are needed regarding the separation or overlap of sexual desire, sustained desire, receptive desire, and cognitive/psychological arousal.

VI. OBTAINING A REPRESENTATIVE STUDY POPULATION – ISSUES IN GENERALIZABILITY

1. SAMPLING A REPRESENTATIVE STUDY POPULATION

General Principles. Sample selection for clinical trials of candidate treatments for sexual dysfunction in women is driven by the specific objectives of the clinical trial in question, the degree of sample representativeness required to achieve these objectives, and the requirement for experimental integrity that permits unambiguous attribution of study effects to study treatments. Inclusion criteria for research participants must ensure (a) representation of the clinical phenomenon of interest at levels appropriate to the objectives of the study (e.g., study of safety, tolerability, or efficacy), and (b) representation of participant sample characteristics matched to the target population to which it is wished to generalize clinical trial results, which defines the external validity of the study. [50,51,52]

Exclusion criteria are determined by the requirement to ensure safety of study participants and by the need to avoid clinical trial sampling procedures from introducing confounding factors (e.g., recruiting women who are involved in concurrent treatment efforts that may be productive or interfering, or who are experiencing intractable relationship difficulties rendering their sexual dysfunction refractory to pharmacologic intervention) that would impair ability to attribute treatment effects or lack of effects to the treatment under study. The presence or absence of such confounding factors in the sample population, together with trial design and measurement parameters, help define the internal validity of the clinical trial, or the degree to which it is possible to unambiguously attribute clinical trial treatment effects to the treatment per se. [50,51,52] There is often a need to balance requirements for internal validity (achieving a pristine experimental design and sample recruitment strategy that permits unambiguous attribution of treatment effects to the intervention of interest) and external validity (ability to generalize from the
临床试验样本到目标女性人口中的性功能障碍

**Sample Selection in the Context of Clinical Trials.** Sample selection in the context of clinical trials is driven by the specific objectives of the Phase I, Phase II, or Phase III clinical trials. **Phase I trials** focus on assessing safety and tolerability of a study compound beyond data obtained in animal studies. While sampling for Phase I trials of candidate compounds in women may seem merely to require passive participation of physiologically normal women subjects, sampling for such trials may prove to be more complex than first apparent. Examination of safety and tolerability in clinical trials of treatments of sexual dysfunction in women will require healthy female participants, whose hormonal status (e.g., oral contraceptive use, menopausal status, hormone replacement therapy use), pregnancy status or pregnancy intentions, comorbid conditions and concurrent medication use permits generalization of Phase I safety and tolerability results to the intended treatment populations. Moreover, sample selection representativeness requires recruitment of participants who are likely to complete the study across the Phase I time interval and multiple potentially invasive procedures involved in single dose, multiple dose, and pharmacology studies, as systematic attrition of intolerant subjects or those who suffer adverse effects can introduce interval validity violations that will render safety and tolerability findings of limited or no value.

**Phase II clinical trials** focus on identification of an effective dose range for a candidate treatment of a female sexual dysfunction, and early in the Phase II trial a wide array of inclusion and exclusion decisions must be considered in the selection of an appropriate research sample. Following the general principle that sample selection is driven by the specific objectives of the research in question, it will prove critical to sample women who experience a moderate level of the clinical condition in question, as women with an insignificant or an overwhelming level of the condition will be unlikely to show any effective dose treatment response vs. placebo. Similarly, with respect to exclusion criteria, women experiencing lifelong sexual dysfunction, women whose partner relationships appear overwhelmingly to be a cause of their sexual dysfunction, and women for whom a specific medical or pharmacological issue appears to contribute significantly to sexual dysfunction, would not be appropriate candidates for a Phase II study of a candidate treatment. Selection of participants unlikely to attrit during the course of the trial remains a concern. Phase II sample selection should be guided by the need to create “greenhouse conditions,” in part via selection of a sample with appropriate levels of the dysfunction at focus and lacking characteristics that would render the dysfunction refractory to treatment, for identification of an efficacious dose of a pharmacologic treatment for a female sexual dysfunction. In seeking to do so, the tension of balancing internal validity with external validity is evident. In seeking this balance, many of the sampling issues of core concern in Phase III research, including sampling the diagnostic entity at focus at appropriate levels and exclusion of appropriate potential confounders without unduly restricting generalizability of results, come clearly into focus early in the Phase II trial process.

**Phase III clinical trials** aim to demonstrate the efficacy of a candidate compound for the treatment of a female sexual dysfunction within a representative sample of those experiencing such dysfunction and who might reasonably constitute a treatment target population. Accordingly, sample representativeness in relation to the population of women suffering from the dysfunction, balanced by the need for exclusion criteria sufficient to ensure experimental precision, are paramount concerns in Phase III research. Inclusion criteria for Phase III research are guided by the need to capture participants who represent the diagnostic entity under study, with representative severity of dysfunction (neither too little nor too much), duration of dysfunction, and distress attendant upon the dysfunction. At the same time, it is necessary to exclude those for whom the trial might prove unsafe and, to ensure internal validity, exclusion of those whose hormonal, comorbid illness or medication, and partner relationships might prove confounding factors that obscure ability to attribute research results to the study compound. Again, a careful balance between representativeness and precision is thus required. In Phase III trials, involving the active treatment of a woman’s sexual dysfunction that may be situated in and affect her partnered relationship, partner willingness and potentially partner consent must be considered. Enrollment of partners as clinical study participants may also be a wise choice to permit detection of positive or negative effects of treatment of the index patient on her partner [53].

2. ADDITIONAL CONSIDERATIONS IN SAMPLE SELECTION

**Contextual and Interpersonal Factors.** Within the general need to select samples that are appropriate for study of the research question at hand, and to balance sample representativeness with exclusion criteria sufficient to ensure internal validity, it is necessary to consider contextual and interpersonal factors that are relevant to both concerns. External validity may require representation of contextual and interpersonal characteristics in research sample participants (e.g., participants drawn from ethno-cultural groups that tend to inhibit female sexual expression; women with childbearing responsibilities; women with a burden of comorbid illness, or other stressors; women with sexual dysfunction whose partners...
also have sexual dysfunction), while the inclusion of samples with these complex characteristics may serve to obscure effects of a candidate treatment on endpoint indicators of female sexual dysfunction. A progressive approach to research sampling, moving from more highly select “greenhouse condition” samples in Phase II research, maximizing internal validity, to more representative and more heterogeneous samples in later stages of Phase III research, permitting broader generalization of findings with a foundation of highly internally valid studies, may serve as a means for balancing external and internal validity of research approaches.

**Hormonal Status.** In all research on female sexual dysfunctions, it is critical to systematically incorporate women’s hormonal status into sampling strategies on the basis of anticipated hormonal status moderation of effects of a candidate compound. At a minimum, it would be anticipated that pre, peri, and postmenopausal status women may differ on the basis of hormonal milieu and anatomic and physiologic aging in response to candidate treatments; as such, sampling and research designs must reflect this reality. In addition, it is likely that hormonal contraception (oral, transdermal or injection routes of delivery), and potentially ovarian cycle phase, may moderate treatment effects and sampling; research design, and assessment of covariate assessment and control must be integrated into research strategies in this area to address related external and internal validity concerns.

**Overlapping and Comorbid Conditions.** A number of overlapping and comorbid conditions may be both representative of women with a particular sexual dysfunction and confounding factors that could serve to obscure clinical trial results. It is likely that women with sexual dysfunction may experience medical illnesses, psychiatric conditions, and medication regimens that may contribute to their dysfunction, and sampling for research seeking to demonstrate efficacy of candidate compounds should exclude women with such cofactors. At the same time, women with sexual dysfunction in the treatment target population to which it is wished to generalize may well have these or other characteristics ranging from alcohol use to smoking to concurrent psychotherapy to relationship distress. Within a progressive approach to increasing external validity after establishing treatment efficacy within internally valid research protocols, participants with a representative range of concurrent relevant conditions may be sampled in research that seeks to identify boundary conditions of efficacy of a pharmacologic treatment of a female sexual dysfunction.

**Symptom Duration, Symptom Severity, and Symptom Distress.** For a variety of clinical, scientific, and regulatory reasons, it is clear that symptom duration, and as alluded to earlier, symptom severity, and symptom distress need to be considered in sampling strategies. Symptoms that are life long may be refractory to treatment, and symptoms that are highly intermittent are unlikely to show treatment effects within time-limited clinical trials. Similarly, symptoms that are exceedingly severe, or exceedingly minor, as measured via screening or interview metrics are unlikely to be sufficiently responsive to treatment in the context of a clinical trial. Finally, opinion generally holds that distress attendant to symptoms of sexual dysfunction is both necessary for diagnosis of such dysfunction and is a potential treatment-sensitive endpoint. Accordingly, avoidance of exceedingly high or low levels of distress may be advisable in sampling for clinical trials in sexual dysfunction.

**Additional Areas of Interest.** Evaluation of subject’s level of satisfaction with treatment may provide information about overall acceptability of the intervention, taking into account therapeutic efficacy, tolerability, and convenience. Quality of life measures may also be of value.

**The Partner as a Subject.** It is well recognized that sexual dysfunction is very often embedded in the context of a couple relationship. Not only may it be advisable for partner willingness and partner consent to be assessed, but it may also prove valuable in exploring treatment effects to examine treatment impact on the female participant as well as on her partner. Placebo-controlled clinical trials in sexual dysfunction have indicated that female partners of treated male patients may experience substantial positive benefits of partner treatment and exploration of the reverberating effects of treatment of women with sexual dysfunction on the index patient and her partner is certainly warranted [54,55]

### 3. NATURE OF CONTROL GROUP

Selection of a control group for clinical trials in female sexual dysfunction should be guided by the same imperatives as sample selection for those randomized to treatment conditions: representativeness of the clinical parameters under study as appropriate to a Phase I, Phase II, or Phase III study, balanced by the need for internal validity and experimental clarity. Research of this nature may involve crossover or repeated measures designs in which an individual woman participant serves as her own control, with consequently increased statistical power and appropriate consideration of drug washout periods; it may involve a placebo-control group in which a single sample of appropriate representativeness is randomized to treatment or control condition, or it may involve a randomized controlled trial involving randomization of appropriately representative participants to a placebo group, an investigative compound, and a comparator (when available) to determine the efficacy, equality, or superiority of a candidate treatment of a female sexual dysfunction.
Special sensitivity and effort to avoid systematic attrition of control arm participants who may not experience symptom relief and who may bear the burden of side effects is needed to avoid compromising the internal validity of the clinical trial.

VII. ISSUES IN STATISTICAL ANALYSIS

1. PLACEBO-CONTROLLED TRIALS

Large placebo-response in FSD trials makes it difficult to demonstrate significant drug-placebo differences. Large placebo responses (>50%) are likely to lead to failed trials.

Some concern may be raised for measures with placebo response over 40%, too; but a 42% placebo response on the primary endpoint for a trial of sildenafil for FSAD did not prevent significant statistical separation from placebo. A subgroup of patients with FSAD + HSDD showed poor drug-placebo separation but the placebo response was only 2% higher (44%). [47] Thus, subpopulation assessment may provide direction for future trials, particularly in terms of inclusion and exclusion criteria.

Low placebo response may obviate such problems; e.g., in the primary endpoint of a trial of a melanocortin receptor agonist in FSAD, the published main endpoint, a sexual satisfaction scale from 0 (extremely dissatisfied) to 5 (extremely satisfied), showed 82% of patients satisfied with bremelanotide vs 10% with placebo after twenty attempts at sexual intercourse. Earlier timepoints, after four and twelve attempts, were similar for placebo (12 and 15% satisfied, respectively) but showed increasing rates over time with bremelanotide (52 and 78%, respectively), p<0.01 for all of these comparisons. [56]

Placebo response in excess of 50%, however, can severely limit the assay sensitivity of a clinical trial. The ultimate example would be 100% placebo response, of course, which would allow no differentiation from active treatment. As an illustration of this problem, a daily diary measure of maximum intensity of sexual desire in premenopausal women with HSDD was used as a co-primary endpoint in the fibanserin phase III project. Although percent change from baseline was not a pre-specified endpoint and individual percent patient changes were not evaluated, the crude group placebo response on this measure exceeded 50% in all trials, at approximately 60-70% per trial, and the diary desire measure was the least robust endpoint in the series of trials. [57]

Relationship to measure, recall period, etc. A given measure tends to have a specific recall period, of one, seven, thirty days, etc. Controversy exists on what is an adequate period for a patient to make sense of her pattern of sexual symptoms, but women with HSDD have newly been debriefed to indicate unequivocally that they do not accept a one-day recall period, but do endorse a one to four week recall period related to sexual desire. [12]

In the same fibanserin studies as described just above, with a crude overall 60-70% response to placebo on the daily desire measure, the FSFI desire self-ratings, with a 4-week recall, showed 33% or less mean crude overall effect in the placebo groups. On other endpoints, e.g., changes in sexual distress (seven-day recall period) and proportion of patients rating themselves as improved on a global impression question (recall period: 4 to 24 weeks), placebo response was similarly low. The credibility of the e-Diary’s one-day recall results is thus questionable, as it is such an outlier in terms of placebo response. (Figure 1)

Relationship to Number of Treatment Arms. A supposition sometimes made is that the lower the proportion on placebo (the higher number of active arms; the higher the proportion on putatively “active” drug), the higher the placebo response. Too few large-scale trials in FSD have been done to establish this effect in FSD. The BI development program used 2-arm, 3-arm, and 4-arm trials (50, 33, and 25% on placebo, respectively). Of well-powered phase III trials, one out of two 4-arm trials had high placebo response; neither of two 3-arm trials had an excessive placebo response, and neither of the two-arm trials had a high placebo response. The two TTS Phase III and three Phase IIIb/IV trials in surgically or naturally menopausal women were all done with a two-arm design (50% on a single level of TTS, 50% on placebo) and were uniformly positive with moderate placebo response.

2. DOSE-FINDING

Duration. Duration of treatment in dose-finding studies must be consistent with onset expected and with practical clinical realities: if an agent shows no obvious onset of efficacy within one, or at most two, months of use, poor (or no) compliance thereafter, and thus poor efficacy should be expected. The maintenance dose should be achieved reasonably early, at least within the first month; consuming more time for titration of dose may be impractical.

Sample Size. In general, sample size should be large enough for adequate power of the trial to confirm whether further development of a treatment is worthwhile or not. Trials based on main endpoint separations and variance values consistent with less than 75% power cannot inform such decisions with much accuracy. Dose-finding done early in development with small numbers of subjects (n/treatment of 20-75) may help guide future development in dosage but can mislead by Type I and II errors as to whether the treatment actually works. An alternative
is to perform several trials each with moderate statistical power and look for patterns of confuence.

If done late in development, dose-comparisons multiply costs and patient burden in seeking the minimum effective dose because a large number of patients may be given inactive doses. This is likely to impact compliance and subject retention negatively.

Controls. Dose-finding not controlled with parallel double-blind placebo may mislead by overly positive results due to a high “placebo” effect. A positive control, if any were available, is not an adequate substitute for placebo because superiority over a positive standard is highly unlikely, and similarity to a positive standard may mislead because both agents may not have exceeded a placebo if one had been included. However, using a dose of the agent at what is demonstrably a no-effect level in Phase I studies may help surmount the difficulties of using a placebo. These include rejection by some ethics review boards, and poor enrollment due to low expectations of obtaining personal benefit.

Fixed vs Flexible Dosing. Fixed dose design can create a barrier to efficacy if side effects arise early and the onset of efficacy comes late, factors often leading to excessive subject dropout, especially in non-progressive conditions such as FSD. Yet, flex-dosing mixes in time effects, potentially obscuring dose effects. A forced up-titration to a fixed dose may be the best compromise to determine the minimum effective dose. Yet flex-dosing is the only way to optimize the dose for all patients in a treatment group, thus maximizing opportunity for drug-placebo separation and to best simulate effects to be expected in clinical practice.

Rescue down-titration may add value to salvage fixed-dose trials by helping lower dropout rates.

Maximum Dose. In early dosing studies of patients with FSD, it may be more germane to be guided in the maximum dose not only by toxicology results or side effects in an indication with high patient symptomatology (depression, etc), but more so by the side effect profile in normal female subjects, who tend to approximate closely the tolerability profile of women with FSD.

Timing of dosing. Sedative agents should be given at bedtime; activating agents should be given in the morning. Doses of agents which cause nausea, etc should be administered with meals if that does not impair absorption. In general, dosing instructions should seek to minimize side effects in a practical way.

The number of doses per day is another practical issue. Trials should keep the regimen practical for patient compliance, not exceeding twice daily dosing unless strongly informed by short half-life issues.

Intermittent Dosing. Trialists and sponsors should consider intermittent dosing if the indication amounts to an episodic disorder; e.g., in arousal, orgasmic,

Fig 1. Mean Change From Baseline with Placebo by Endpoints in Phase 3 HSDD Trials
sexual pain disorders; that is, if symptoms relate chronologically to sexual activity or if the agent works upon first dose. This schedule appears to have less potential for treatment of desire disorders.

Pharmacokinetics (PK). A short half-life may dictate dosing before sexual activity, or at the time of day just before sexual activity is most likely. If the half-life is much less than 8 hours, it may still be effective for FSAD if taken just before sex at bedtime but may require exclusion of, or special dosing alarm/reminder system for, women whose pattern is for sexual activity in the morning. The possibility of cumulative effects must also be considered, even for short half-life agents. For agents with long half-lives, cumulative effects (positive and negative) may overshadow first-dose effects, so trial designs must take this into account with slow increases in doses and emphasis on chronic rather than acute efficacy and safety.

Drug interactions must be considered for agents that are CYP-450 metabolized, especially for those with active metabolites.

Pharmacodynamics (PD). Pharmacodynamic studies early in drug development can obtain useful preliminary answers to several vital questions:

- Which come first, side effects or efficacy?
- How long must the patient be treated before a dose is found to be effective, or can be assured to be ineffective?
- Can an expected (early-onset) side effect be used to guide dosage for a late-onset drug?

Side effects and their severity may depend on the rapidity of upswing in plasma concentration, not just on the absolute Cmax; slowing absorption (even if only with food) may make a medication more tolerable and allow the patient to persist in compliance until she obtains efficacy.

Carefully chosen surrogate endpoints potentially can guide dose-finding in later studies, e.g., if an agent is sedating, prospective measures of sedation, attention, and/or cognition are likely to give stronger dose differentiation than simply a general inquiry about how the subject feels. Similarly, self-rating scales can be modified to a shorter recall period to fit the restrictions of early (phase I trials), e.g., a 4-week recall can be shortened to 1 week. [31]

3. DURATION OF TREATMENT

Three months may be adequate to test for induction of effects on feeling states (sexual desire, sexual distress) but may not be long enough to alter a couple’s dysfunctional pattern of sexual inactivity and/or dissatisfaction, whether it was a consequence of the FSD or not. Thus, this duration may be suitable for Phase II trials, but not for Phase III studies, in which sexual activity, and in particular the frequency of satisfying sexual events, is likely to be required as a primary endpoint. Several trials have shown concordance of improvement in feeling states and satisfying sexual activity over this period [20,21,22,23,58,59,60,61].

Female Sexual Disorders are rarely of short duration. Thus, controlled testing of maintenance for six months or longer may be useful to confirm maintenance of effect. A parallel trial with placebo for over a year is in process for the TTS (www.clinicaltrials.gov) but it may be more practical, particularly for retention of subjects on placebo, to use a randomized crossover from open-label “active” treatment to drug vs placebo after several months to obtain plateau maintenance effects. This has been used successfully in a 148-week design in premenopausal women with HSDD. [37] When considerations of withdrawal are relevant, this design may be particularly useful.

4. EFFICACY MEASURES AND THE CURRENT PROCESS AND PROBLEMS

All measures of a particular construct associated with the sexual disorder should have convergent validity. When well-validated measures fail to converge statistically with novel measures, concerns about the suitability of the novel measure arise.

Excessively high placebo response may be measure-specific, as explained above. Excessively low placebo response may also occur, even for measures that discriminate FSD patients well from normal women. Poor response on such a measure can suggest lack of efficacy if the active treatment does not separate from placebo on the measure. However, if it is an outlier among measures, the presumption instead should be that the measure is insensitive to change with treatment, as no history of ability to detect treatment responsiveness is currently available, because all the treatments are novel. This currently appears to be a particular problem for measures in FSD trials as data on treatment responsiveness of validated instruments is not published (e.g. proprietary tools used in TTS trials). This can occur because of irrelevance to patients or because of design flaws in the measure, if other measures in the same trial do show separation between active treatment and placebo, e.g., crude overall group placebo response was about 5% on a daily e-Diary question on intensity of sexual distress (not a thoroughly validated measure) in a set of Phase III trials of premenopausal women with HSDD, whereas crude overall group placebo response on the well-validated FSDD-S-R (measuring frequency but not intensity of sexual distress) was about 15-20% in the same women at the same time, and crude overall group desire improved 26-33% on the well-validated FSFI desire domain. Thus, the credibility of the e-Diary distress item results is questionable. (Figure 1)
This points to a major potential problem of daily measurement: that the measurement itself influences ratings. Whether the level of placebo response varied so greatly between measures (daily desire, 60% response or more; daily distress, 5% response, both in the same placebo group) because of the frequency of measurement, and/or because of the power of suggestion invoked by the measure, the large disparity between measures suggests a serious problem with daily measurements of feelings related to HSDD. Self-ratings with a 7 to 30 day recall period, the FSFI and FSDS-R, generally showed less than 40% crude overall group placebo response in the same program. This demonstrates the importance of less frequent measurement, which has been argued on theoretical and practical grounds. [62]

5. SAFETY MEASURES AND ADVERSE EVENT MONITORING

Spontaneous reporting of adverse events is the standard method, but it is known in general to give lower incidence rates than prospective, elicited reporting. The prime example in this field is that the first selective serotonin reuptake inhibitor (SSRI) antidepressant, fluoxetine, was originally described in the package insert as associated with a 1.9% rate of sexual dysfunction. Prospective, elicited reporting of sexual dysfunction with now-standard rating scales such as the Changes in Sexual Functioning Questionnaire (CSFQ) have shown rates of FSD more than 15-fold higher. [63]

The FDA’s new standards require prospective, elicited reporting of suicidality. The FDA has also mandated prospective, laboratory-based determination of whether an agent causes infertility. As appropriate, determination of relevant endogenous sex hormone levels should be used to inform long-term safety, e.g., carcinogenesis and cardiovascular disease. Thus, certain classes of compounds may require a longer treatment period to determine safety.

VIII. ADDITIONAL ISSUES IN FSD TRIALS

Comparator trials. No gold standard treatment for comparison exists, although tibolone, the Eros device, the testosterone transdermal system for postmenopausal HSDD, or behavioral treatments may serve as models to inform calculation of a sufficiently powered trial (not over-powered or under-powered).

Statistically significant change vs. clinically meaningful effect. Standards for statistical differences and clinically important differences must be determined a priori or anchored by a global measure of change: efficacy and effectiveness. Achieving a statistically significant difference between a new treatment and placebo is insufficient to define the value, or effectiveness, of the new treatment. The latter can be done only by finding a clinically meaningful difference between the two treatments. This can best be done by comparing the proportion of responders between the treatments.

Confluence of arbitrary responder values and clinical data to determine responders. Few normative population data are available to help establish responder criteria for adequate sexual function in pre- and post-menopausal women, but a value of 25, 30, or 50% of normal functioning might be chosen arbitrarily as a responder value for various endpoints. Data from comparisons of women volunteers with a given FSD diagnosis vs volunteers without sexual complaints suggest that such values would be clinically meaningful.

For the FSFI desire domain, the minimum score is 1.2 and the maximum is 6.0. A 20% improvement based on the maximum score would be 1.2. A 50% improvement based on the baseline score would be 0.9, but patients can improve by increments no smaller than 0.6, so 0.9 would amount to 1.2 anyway. The value of +1.2 to define a responder on this measure corresponds with the value determined by anchoring methods (see below).

For the FSDS-R, an improvement of 6 points has been recommended by the author of the scale, L. Derogatis, as a responder value, which corresponds for this 13-item, 52-point scale to approximately a one-half point change per item. It corresponded, in the ibanserin Phase III program, to a 20% improvement toward the scale maximum for functionality, i.e., zero. [4]

Confluence of anchoring methods to determine responders. Alternatively, anchoring a responder value in terms of a patient global impression question can be used.

For the TTS Phase III project in surgically menopausal women with HSDD, to define responders on the desire measure, the desire domain of the Profile of Female Sexual Function (PFSF), a value was chosen by anchoring for optimal sensitivity and specificity on the Patient Benefit Evaluation (PBE), i.e., 9.0. Comparing to the baseline mean on the PFSF, 21.3, a responder would be one with a 42% increase in desire.

In the TTS project, responders on the sexual distress measure, the Personal Distress Scale (PDS), were defined the same way, resulting in a responder value of -20 versus the baseline mean score of 65.05; this represents an improvement of 31%.

In the ibanserin Phase III project in premenopausal women with HSDD, responders on the FSFI desire domain or on the FSDS-R distress measure were defined similarly, but the anchoring used the mean difference (on the desire measure or distress
measure) between those self-rated on the PGI-Improvement as unchanged vs those minimally improved. The logic here is that a responder is one who achieves the minimum clinically important difference (improvement) between no treatment and treatment. The responder value was the same when calculated as the difference between those self-rated as minimally improved vs those much improved. This method of analysis seeks to determine the minimum clinically important difference between optimal and suboptimal improvement with treatment.

By either the PGI-anchored method of defining a responder, or by anchoring for optimal sensitivity and specificity on the PBE, the responder value was 1.2 for the FSFI-desire domain, and -6 (+1, depending on trial and method) for the FSDS-R. Use of independent data sets on women with HSDD not in a clinical trial, or in different clinical trials derived the same responder value.[11] Seeking and finding such a confluence of values helps assure the clinical meaningfulness of the definition of a responder.

On the 2-to-10 FSFI desire domain score (of two items, each rated 1-5), which in its standard analysis is multiplied by a factor of 0.6, an improvement of 1.2 represents a change of 2.0 in the raw score, i.e., exactly two scale points. The baseline mean in the fibanserin program was 1.8 for the FSFI desire domain. Thus, the PGI-anchored responder value represents a 67% improvement.

FSDS-R mean values of about 30 have been found for untreated premenopausal and postmenopausal women with HSDD or FSAD (vs scores of 3-4 in women with no sexual complaints) [4]. Thus, a decrease of 6.0 points would represent a 20% improvement for patients on the mean.

In any case, a reasonably large ratio of efficacy with active treatment over placebo is required not only for statistical significance but also for clinical relevance. This may alternatively be registered on effect size (mean difference divided by the common standard deviation at baseline) of at least 0.2 (small) to 0.5 (moderate), a statistically significant difference in the proportion of responders (usually >10%; e.g., anchored on the difference between no change and minimal improvement or between minimal improvement and much improvement on the patient’s global impression of change), or an (adjusted) odds ratio close to or exceeding 1.5. In the FSAD trial of Berman et al. [47] for instance, a subgroup with both FSAD and HSDD was included. Their adjusted odds ratio (OR) was 1.09 or 1.14 (primary endpoints question 2, and question 4), which demonstrated that inclusion of such dually–diagnosed patients formed the efficacy issue within the overall analysis, not a high placebo rate. The ORs for FSAD-only patients were impressive, with values of 7.98 and 10.95 on the two respective questions (p<0.001).

An example that involves comparing the proportion of responders is use of the testosterone transdermal system for HSDD in post-surgically menopausal patients. In the two trials of TTS in such patients, the % responders on PFSF desire, defined by anchoring (at a cut point of 8.9) to optimize sensitivity and specificity using a patient benefit evaluation question (yes or no meaningful benefit from the treatment) to define responders vs nonresponders, was reported in the SM1 and SM2 trials as favoring TTS over placebo significantly at 45.7/34.8% and 42.2/25.1%, respectively. For PDS distress, similarly defined, responders were also significant in both trials: 51.3/34.6% and 49.2/33.9% respectively.

The level of improvement required to define a responder on FSFS desire and PDS distress based on the TTS FDA advisory committee slide set of December 2004 was 42% for PFSF Desire (>8.99/mean baseline of 21.3). -31% for PDS distress (<-20/mean baseline of 65.05).

Remitters. A higher standard, but the one with the utmost clinical meaning, would be for a new treatment for FSD to produce a meaningfully higher rate of remission than placebo. Treatments for FSD cannot be expected to produce a high rate of remission, given that FSD may often, or perhaps even usually, be not only biologically mediated in part but also mediated and/or perpetuated by psychosocial factors. However, all new treatments should be evaluated not only for ability to induce response but also for ability to induce remission. Clinically meaningful superiority to placebo on the rate of remission should be an achievable goal.

On the FSFI desire domain, for instance, the cutoff between women with HSDD and normal women has been found to be >3.0 on the derived version of the domain. That is, achieving a score of 3.6 would be required for remission. The baseline means in the fibanserin program were 1.8, so an improvement of 1.8 would represent a 100% (mean) improvement. For the FSDS-R, the baseline value in the fibanserin program was about 30, and the established cutoff between HSDD and no FSD is 15. [4,30] This cutoff, representing a 50% group mean improvement, could be used to designate remitters.

Ethical Issues. Ultimately, somatic treatments will be prescribed mostly by clinicians not expert in FSD. Thus, an ethical issue arises as to how well they can screen/diagnose patients as candidates for such somatic treatments. To fill this need, acceptably simple (and brief) screening/diagnostic tools for non-expert clinicians must be well validated before approval of such agents. [64] A method for such validation is a one-way crossover study in which results of two diagnostic interviews are compared after a substantial sample of patients are screened using the new screening/diagnostic tool by a non-

...
trained, non-expert clinician, followed by diagnosis by an clinician expert/experienced in diagnosing FSD who is not permitted access to the non-expert’s diagnostic decision.

**FSD impairs quality of life.** However, it is, of course, non-life-threatening. Thus, the burden of proving a very high level of safety must be assumed by sponsors of treatments for FSD. No treatment for FSD should be recommended that has not stood the test of a large (multi-thousand-patient), long-term (at least 1-2 year) prospective evaluation of safety in the target population, using appropriate safety measures and appropriate controls, without causing life-threatening side effects or being associated with a rate in excess of placebo of laboratory abnormalities known or expected to be associated with long-term health risks, be they as specific to FSD as prolactin levels or as nonspecific to FSD as cholesterol levels. Thorough, perspicacious, scientifically valid risk management practices and studies must be assumed by sponsors after marketing, also, because of the significant multiple-effect to be expected for patient exposure compared to even large programs of clinical trials.

**Disclosure of results.** For the advancement of the field of treatment of FSD, it is incumbent on sponsors to disclose trial results. Proprietary concerns must be acknowledged and accepted. However, ordinarily in the past an ineffectiv treatment was seldom acknowledged in public. This can easily prevent or retard research funds from flowing from continuing fruitless mechanisms of action into useful directions.

Trial results (at least in the US) have become subject to disclosure in www.clinicaltrials.gov to be considered for publication in peer-review journals. This new standard has opened doors to disclosure. However, this remains to be seen, as the policy has been in effect only since 2007.

**Post-conduct issues.** Selective publication of positive results obscures difficulties to be expected in replication of trial results, and in effect, magnifies the efficacy of an agent. Thus, this is strongly discouraged in favor of full disclosure of all results in each phase of investigation of a new agent.

Post-hoc analysis of “evaluable patients” rather than intent-to-treat (ITT), and publication of the latter without the former, or failure to provide fair balance between conflicting analyses in publications, are also strongly discouraged. Like selective publication of positive results, these practices unfairly magnify efficacy expectations.

Issuing press releases before submitting results to peer-review publications, while a temptation for sponsors, serves only to negatively affect the view of the FSD scientific community and regulatory bodies alike about the utility of a new agent, and is to be discouraged.

**VIII. CONCLUSIONS**

Fine-tuning outcome measures to directly assess the specific FSD construct under evaluation should lead to improved ability to demonstrate efficacy when such an effect exists. Future clinical trials should include clear population definitions, direct and indirect measures of the specific FSD construct, and procedures to allow generalizability of diagnosis and treatment to the general/target population.

**REFERENCES**


22 Van Anders SM, Cherrick AB, Cherrick BA, Hampson E, Fisher RA. Preliminary clinical experience with androgen administration for pre- and postmenopausal women with hypoactive sexual desire. Journal of Sex & Marital Therapy 2005;31:173-185


28 Ter Kuile MM, Brauer M, Laan E. The Female Sexual Function Index (FSFI) and the Female Sexual Distress Scale (FSDS): psychometric properties within a Dutch population. J Sex Marital Therapy 2006; 32: 289-304

29 Okuz E, Malhan S. Reliability and validity of the Female Sexual Function Index in the Turkish population. Aile Hekimligi 2005;17:54-60


35 Hayes RD, Dennerlein J, Bennett CM, Fairley CK. What is the true prevalence of female sexual dysfunction and does the way we assess these conditions have an impact? J Sex Med 2008;5:777-787


52 Rothwell PM. External validity of randomized controlled trials: “To whom do the results of this trial apply?” Lancet 2005;365:82-93.