Committee 16

Research Methodology in Urinary Incontinence

Chairman

C. Payne (USA),

P. Van Kerrebroeck (The Netherlands),

Members

J. Blaivas (USA),

D. Chaikin (USA),

H. Herrera (USA),

U. Jonas (Germany),

L. Kusek (USA),

A. Mattiasson (Sweden),

L. Nyberg (USA),

T. Peters (U.K)

M-A Stothers (Canada),

A. Webers (USA)

Research Methodology in Urinary Incontinence

C. Payne, P. Van Kerrebroeck,

J. Blaivas, H. Herrera, D. Chaikin, U. Jonas, L. Kusek, A. Mattiasson, L. Nyberg, T. Peters, M-A Stothers, A. Webers

I. INTRODUCTION

Lower urinary tract dysfunction comprises a group of common diseases, and we need far more knowledge of their origin, diagnosis, treatment, and ultimately prevention than we have today. Clinical research is a precondition for any progress in these areas. The task for the present committee is to provide recommendations for good research practice, including principles of trial design and correct statistical methodology. In addition, the aim is to give recommendations, when possible, on current concepts and outcome measurements, as well as to provide specific recommendations for various methods and their application in different groups of patients. Other ICI committees report on the etiology, epidemiology, pathophysiology, prevention, and economics of lower urinary tract dysfunction. This committee only covers these areas briefly, when appropriate.

Methodology and terminology in incontinence research should comply with standards established by the International Continence Society (ICS) [1 - 15]. The ICS Standardization Committee describes the "now-state" (how to do things right using present knowledge) with regard to interventions in patients by a number of "Outcome Groups" [12 13 14 15] while the future "desired state" (to do the right things in order to develop the area) is processed by a number of work groups with the collective name "Clinical Research Assessment". The recommendations from the ICS Standardization Committee on "General Outcomes" [12] and subcommittee recommendations [13,14,15] lower urinary tract dysfunction in women, men, frail older people, children, and neurogenic disorders are integrated into this chapter. Recommendations from national working groups have also been included when appropriate, for example, the Urodynamic Society's recommendations for outcome research [16,17,18].

The aim of clinical research is clear – namely, to offer freedom or relief from symptoms, and eventually to prevent the origin of disease. The need for high quality research is similarly evident. The prevalence and impact of genitourinary disease has been underestimated in the past, impairing research efforts. Many factors, including embarrassment and ignorance keep patients from seeking care and contribute toward making incontinence a silent disease. A gradually increasing degree of openness in society means that patients feel less guilty about urinary incontinence and other lower urinary tract symptoms. When these factors are coupled with the marked demographic trend toward an older society it produces an explosive increase in demand for incontinence therapy. At the same time, our knowledge of the etiology of incontinence, optimal treatment strategies, and prevention is greatly inadequate. In order to meet this need effectively, we need to intensify research. Proper choice and use of methods decides whether or not we will be successful in our efforts. The quality of research is not only important immediately (to improve treatment) but in a larger sense that research funding is severely limited in relation to need. There is great competition for government funding and other resources. While there is ample evidence that urinary incontinence is a greatly troublesome disease creating a major impact on patient's quality of life, priority for funding research may be compared against heart disease, cancer, infections, as well as in relation to other non-life-threatening diseases. We must therefore constantly strive to produce the highest quality work so as to make the best use of each research dollar and to encourage future competitive funding. We must also acknowledge the wide spectrum of interest in incontinence research. While clinicians are primarily interested in specific disease outcomes, government bodies may require more generalized assessment of global impact on quality and quantity of life.

There are many goals of research—foremost to improve care of patients, but also to promote understanding of the disease process. We need a broad spectrum of information if we are to not only understand which treatments work but also how and why they work (or don't). The ultimate goal is to produce credible research. When research is inherently credible due to strong study design the impact is maximized. The clinical application of the research will be hastened and other investigators will be energized to use the information in their own quest for knowledge. The recommendations given in this document have less the character of definitive standards than of guidelines and options, although on occasion the current state of evidence is such that standards are available. We have tried to keep a discursive style, but also give firm and practical advice when possible and appropriate.

II. GENERAL RECOMMENDATIONS FOR CLINICAL RESEARCH IN INCONTINENCE

1. The planning phase of a clinical study on Incontinence

Meticulous planning is essential in all clinical research. The work done before initiating a research protocol does not in itself guarantee success but it is the obligatory first step in that direction. At the same time, inadequate planning can doom even carefully conducted protocols. Both prospective and retrospective studies require the same deliberate approach in the planning stage. The background and rationale for the study as well as the study objectives and/or hypotheses must be clearly elucidated and documented. There are many mistakes that should be avoided in the planning and conduct of a clinical study. A detailed list of common pitfalls in preparing and writing protocols has been taken from Spilker [19] to illustrate many avoidable errors (Table 1).

The first step of the planning phase involves reviewing previous and, if possible, ongoing work in the field. A study may be well performed but clinically irrelevant. A thorough knowledge of related clinical work is the cornerstone of protocol development. Fortunately, in urinary incontinence the work of the Cochrane Collaboration (www.cochrane.org) and its Cochrane Incontinence Group (www.otago.ac.nz/cure/) provide a tremendous asset to the potential investigator. The Cochrane reviewers have registered over 1300 trials in the field and produced 16 reviews (with others in progress) that cover most issues in conservative therapy as well as many other topics. This concentrated collection of data allows researchers to hone in on key questions

Table 1: Common pitfalls in preparing and writing protocols (from Spilker 1984)

A. Study objectives

1. Expressed too generally to allow a specific study design to be constituted
2. Ambiguous or vague
3. Not achievable with the current study design. The study may be too complex or there may be inadequate resources to conduct the study

B. Study design

1. Insufficient statistical planning—the design will not adequately address study objectives
2. The design chosen is beyond current state of the art
3. Inadequate validation of outcome measures
4. Inadequate statistical power. The chosen sample size is too small to detect clinically meaningful differences
5. Inappropriate use of active or inactive controls
6. Lack of placebo or double blind when one or both should be incorporated
7. Dose regimen too restrictive (e.g., range of allowed doses, alterations of dosing for adverse reactions)
8. Failure to consult with statistician regarding randomization process

C. Inclusion/exclusion criteria

1. Too stringent to allow adequate numbers of subjects to be enrolled. Overly stringent criteria also reduce the generalizability and thus the impact of research
2. Too broad to create homogenous groups.

D. Screen/baseline/treatment

1. Time periods for data collection are either too long or too short for optimal conduct of the study
2. Too few or too many measurements are requested
3. Subjects may be inappropriately entered into the study before complete screening
4. Excessive blood volume removed for testing or an excessive period of fasting is required. This is especially common in pharmacokinetic studies

E. Drug packaging/dispensing

1. Drug packaging that does not permit all options allowed by protocol to be followed

F. Study blind

1. Study blind easily broken because of "obvious" characteristics (e.g., adverse reactions, changes in laboratory parameters, drug odor) that are difficult or impossible to adequately mask
2. Study blind easily broken by observation of drug interactions or other situations by the investigator (e.g. marked improvement in study group or changes in blood levels of concomitant drugs)
3. Study blind inappropriate

G. Data collection and analysis

1. Poorly designed data collection forms
2. Incorrect statistical methods used to analyze data, including baseline comparisons
3. Failure to make the primary research question the main focus of the analysis
4. Reliance on within group rather than between group comparisons in parallel group trials
5. Overreliance on p-values without presenting confidence intervals

H. Overall

1. Ambiguous language that allows different interpretations
2. Too many comparisons requested. Five of every 100 independent comparisons will be statistically significant by chance alone, when alpha is 5% and there are no true differences between the comparison groups.
3. Lack of internal consistency in the protocol
4. Discretionary judgments allowed by the investigator. This may seriously affect the quality and quantity of data obtained
5. Presentation/reporting fails to accord with CONSORT guidelines

and focus their research. Recommendations of the group help to assure that future studies will be interpretable in the context of past work. This is an appropriate starting point for any literature search.

A rule of thumb for all research is that one should seek the least complex approach to adequately answer or address a given problem, hypothesis or question. The project must provide a convincing answer to the question in an efficient manner. At the same time, the committee has struggled with the desire to gain more from clinical research than an answer to treatment efficacy. We still need to understand how our treatments work. It may be helpful to think of the dilemma as a balance between breadth and depth [19]. While it is important to remember that only a limited number of questions should be posed in one study protocol it is still relevant to record as many observations as possible without jeopardizing subject recruitment or retention with onerous demands.

2. Study design: types of clinical trials

Discussion of the various types of studies and other aspects of study design can be described as the framework by which the study objectives will be met. Different considerations might be made for etiological, epidemiological and pathophysiological studies on the one hand, and for clinical trials on the other. Ideally, clinical research should be prospective, controlled and randomized. However, some studies can only be performed retrospectively and good clinical research methodology does not always mean controlled or blinded studies. Sometimes open and uncontrolled studies are accepted, as in a phase I study when a new pharmacological agent is tested for the first time in humans, and in pilot studies where a new surgical technique is being developed.

The following definitions of Phase I-IV studies have been constructed from Senn [20].

- ***Phase I studies:*** The first studies with the actual drug in humans. Often, but not exclusively carried out in healthy volunteers. Pharmacokinetics and tolerance information are obtained in these studies.
- ***Phase II studies:*** The first attempts to prove efficacy of a treatment. These are often the first studies in patients. Dose finding is a common objective of such studies.
- ***Phase III studies:*** Large-scale "definitive" studies carried out once probable effective and tolerated doses of the drug have been established with the object of proving that the drug is suitable for registration.
- ***Phase IV studies:*** Studies undertaken either during or after registration with the purpose of discovering more about the drug safety and efficacy in different populations. Such studies are often larger and simpler than regulatory studies and may lack a control group.

Randomized Controlled Trials (RCTs) are the most important method for demonstrating the effectiveness of treatments. Phase III studies are usually RCT in format, using the regimen indicated from phase II, in comparison with placebo and/or a comparator, such as equivalent drug or other therapy. Parallel groups are recommended as the first choice, while the crossover design is best suited for intra-individual comparisons (see below).

a) Types of trials

- ***Parallel Designs:*** *Parallel clinical trial* designs offer subjects only one treatment during the study. In a placebo-controlled trial, the patient is assigned to receive placebo or the active drug according to the study protocol, with the predetermined outcome measure obtained at the time of the study follow-up. During a drug trial, the dosage of drug may be held steady or, in a variable dose trial, the dosage may be increased to maximize clinical benefit or decreased if side effects occur.
- ***Crossover trials***: An alternative to the parallel groups design is the *crossover trial*, in which patients experience both arms of the study [21, 22, 23]. Like parallel designs, randomization is important in the limitation of bias; in crossover trials, patients are randomly allocated to receive the treatments in one order or the other. Crossover trials allow for within subject comparisons, which may provide a more precise measure of treatment effectiveness. The effect of variance between subjects is removed, unlike in parallel group designs. Crossover studies are particularly well suited for small study groups with chronic stable disease states in which the primary objective is to measure a short-term response in symptoms. The treatment itself should not have a long lasting result once it is stopped. The duration of treatment is important in this design. If the duration of treatment is too short, the treatment may not show its effect or make too small an effect to be adequately measured. If the duration of treatment is too long, compliance may be poor or the disease may not remain stable. *Crossover effects* may occur, where the results of the first treatment linger and affect the second treatment. To avoid crossover effects, a *washout period* is planned where subjects receive either no treatment or placebo. To ensure adequate disease stability before the start of the study, a *run-in period* of monitoring relevant signs or symptoms can be undertaken. Those with transitory or labile disease can then be excluded.

- ***Equivalence trials:*** The primary objective of an equivalence trial is to demonstrate that two treatments are similar in outcome or that there is no difference between treatment and controls. This can be of relevance when one treatment is significantly more cost-effective, offers a better quality of life, or is less toxic or time consuming for the patient when similar clinical outcome can be achieved. In this scenario, it is the primary objective to demonstrate that a more conservative therapy is no different than the standard by a degree of measure that the investigator feels would make the two treatments equivalent. It should be emphasized that this type of trial is not the same as failing to find a difference between two groups. In contrast, this is a powerful design when appropriately employed. The goal is to demonstrate that the observed difference between two treatments is small with a narrow confidence interval. Specific statistical methods are needed to ensure adequate power to find a difference between the groups if one truly exists.

The magnitude of clinically unimportant differences may be quite small, which is one reason why equivalence trials often need large sample sizes [20,24]. Equivalence trials need to be designed and conducted with particular care. Unlike the other approaches, equivalence trials aim to demonstrate that the trial arms are equivalent or, at least, are not particularly different. For a conclusion of this kind to be valid, researchers should be especially vigilant that failure to demonstrate a difference is not merely a consequence of poor study design and procedures.

3. Study conduct and statistical considerations

Research must be planned early, and planned often. All issues should be addressed at the start of the planning process, and many will need to be re-addressed at suitable times throughout the project. Many of these issues are statistical; indeed, the major statistical input to a study should be at the design stage, including planning the data analysis in advance to follow the design of the study. Leaving this until the end of a study will almost always lead to difficulties that cannot be resolved, resulting in a study which is at best inefficient, and at worst inconclusive.

The issues covered here relate to: study design; sampling strategies; randomization and stratification; primary and secondary outcomes; inclusion and exclusion criteria; blinding and effects on validity; control of bias; sample size considerations; pragmatic and explanatory trials; data analysis; and reporting of randomized controlled trials (RCTs). Only the principal features of study design and analysis will be covered here; extensive coverage is available elsewhere [24]. Regarding presentation, the Consolidated Standards of Reporting Trials (CONSORT) statement provides guidelines for reporting the design, detailed methods, and results of RCTs [25]. The original statement [26,27] has recently been revised with the aim of improving clarity and, where appropriate, increasing flexibility [28–31]. Many of the points discussed here relate to those guidelines.

a) Study design

The most fundamental planning issue is whether the study is observational or experimental. Observational studies include a variety of designs, from cross-sectional descriptive studies (where the primary purpose is estimation of the prevalence of incontinence in a defined population) to case-control designs and long-term prospective or retrospective cohort studies. Observational studies may be purely descriptive, or they may be analytic when designed with a control or comparison group. The limitation that all comparisons based on observational data have in common, however, is that it is not possible to ensure that one is comparing like with like. In particular, the bias that results from differential selection effects (both patient and clinician induced) cannot be eliminated, even by the use of advanced statistical methods.

Properly planned and executed, the RCT is the optimal approach to limiting selection bias [32] RCTs compare outcomes in groups of subjects with the allocation of treatment determined by chance. In this way, the treatment groups will not differ in any systematic fashion, and comparisons between them will be unbiased [21]. Subject assignment must be concealed during enrollment (e.g., by separating allocation from the process of recruiting subjects, and by using sealed envelopes or, preferably, telephone randomization). In addition, treatment allocation must be concealed during the trial (e.g., through blinding with or without use of placebo). In some studies, blinding of subjects and health care providers may be impossible or undesirable. This can occur with some surgical trials, or with studies of health care delivery. In all cases, however, those personnel collecting outcome data should be blinded to the subjects' treatment allocation.

RCTs provide the optimum level of evidence about the clinical effectiveness of different interventions [28]. Observational studies can contribute useful information on many aspects of health care [33], and may be necessary pre-cursors to a randomized trial, but the central position of the RCT in terms of influencing patient care should and will continue. The classical (two) parallel groups RCT is not the only option within the experimental paradigm; complex randomized designs such as factorial and cross-over designs may overcome some of the limitations of and objections to the standard approach [23].

The type of study design for a clinical trial is dependent upon the *primary research question*. In particular, the primary question is critical in determining the *sample size* needed for the study. Obtaining adequate numbers of patients to address the primary research question is crucial to avoid the problem of an underpowered study and insufficiently precise estimates of the comparisons between the treatments (for details see below). A sufficient *power* (probability of detecting a difference if it exists) is required to minimize the risk of a Type II or *beta error* – that is, failing to find a difference between the treatment groups even when one exists.

Secondary factors that influence the choice of a study design may be related to the natural history of the disease, the treatment itself or patient endpoints. Patient-related endpoints may be short term such as changes in signs or symptoms or may be more long term such as increased survival. Once the sample size required for the primary research question has been calculated it is usually obvious if the study can be performed in a *single institution* or if a *multicenter* study is required. Single institution studies have the benefit of being less complicated since all personnel are on site and study coordination is less difficult. However, if a large sample size is required a single institution may take years to accrue the required number of patients. While multicenter trials are more complex from an administrative point of view and are generally more expensive, they provide larger numbers of patients in a shorter period of time, and have benefits in terms of the generalizability of the research findings.

b) Strategies

Whether a study is descriptive or analytical, the first practical issue to resolve is the selection of participants. Some studies require a sample that is representative of the community overall. In principle, this is achieved by taking a simple random sample from a known population. In practice, a list of all eligible and consenting individuals is obtained and then a sample is drawn by a method in which each member of the population has an equal probability of selection ('epsem'). Even in ideal circumstances, however, some sophistication on this basic method is usually desirable or necessary. For example, in *stratified sampling*, subjects are arranged into subgroups and the sampling is performed within each subgroup separately. This ensures that the sample is representative of the population in terms of these subgroup characteristics. In *multi-stage random sampling*, the population is first divided into 'primary sampling units' (such as hospital, health center, or surgeon), and a sample of primary units is selected. The 'secondary sampling units' (usually individual subjects) are then selected just within the primary sampling units that have been selected. A special case of multi-stage random sampling is *cluster sampling* where all individuals within each primary unit are included. Standard procedures for sampling should be followed [34].

It is important to note that, while the technicalities of random selection of subjects for a study are closely related to the random allocation of patients in a clinical trial (and indeed there are similar issues in trials relating to stratification and clustering [35]), there is an important distinction in the objectives of the two procedures. First, the (ideally random) *selection* from the population of eligible subjects concerns the *external validity* or generalizability of the study findings (RCT or otherwise). Independent of this, the random *allocation* of subjects in an RCT is concerned with the *internal validity* or comparability of the trial groups.

In principle, sampling should involve random selection. In practice, however, this ideal is rarely met outside of large-scale epidemiological studies. Rather, a consecutive series of patients at a particular health care setting over a specified period of time often forms the basis of recruitment to a study. Where this is the case, it is crucial to provide descriptive information about the study sample, so that broad representativeness can be judged. This is as important for trials as for observational studies. Guidelines for reporting of RCTs include requirements to state the study population, give details of inclusion and exclusion criteria, and present clearly the numbers of eligible subjects who were not randomized and the reasons [28,29,36]. Nevertheless, "the basic logic of clinical trials is comparative and not representative" [20]. In other words, the principal benefit of conducting a randomized trial is to provide groups that allow valid comparisons to be made.

c) Randomization and stratification

Randomization is the process of allocating subjects to groups by chance [21,32]. Neither the subject nor the clinical staff responsible for recruitment to the trial should be able to predict to which group the subject will be assigned. Randomization removes treatment selection from the hands of the clinician thereby removing selection bias.

In order to minimize bias, the randomization process must be concealed from those recruiting subjects to the trial [28,36]. This can be achieved most effectively for multicenter trials by the use of central telephone randomization. Sealed opaque envelopes can be used for studies performed at a single institution. In drug studies, a pharmacy can maintain identical treatment drug and placebo already randomly allocated into individual subject portions. These are distributed consecutively as subjects are enrolled in the study.

• ***Simple randomization*** can use computer-generated random numbers, either prepared specifically for the

trial or using existing tables of random numbers where the digits of 0-9 appear with equal likelihood in each entry. Treatments are assigned to odd or even numbers. As the total number of subjects in the trial increases, the balance of numbers and characteristics of subjects between the groups improves. In small trials, however, balance is not assured by simple randomization. Appreciable imbalances in subjects per group may be particularly important in a multicenter study where imbalances in assignment can occur within individual institutions.

• ***Block randomization*** is one method used to prevent imbalances in subject numbers assigned to each group, particularly when the number of subjects in the trial is small. With block randomization, the total sample size is divided into blocks of a given size. Within each block, the group is assigned so that there are equal numbers allocated to each group. To prevent investigators from learning the block size and being able to guess order of assignment, the block size can be varied, usually at random from a small number of alternatives.

Most disease states have factors known to influence the outcome of treatment. A form of randomization that accounts for such factors is called *stratified randomization* [21,32]. Stratified randomization ensures equal distribution of subjects with a particular characteristic in each group. Stratification is usually restricted to a small number of factors, in particular those most likely to influence outcome. Despite its complexity, stratified randomization is usually helpful in a multicenter trial, so that both the numbers of subjects in each group and the important factors influencing the outcome can be balanced within each site. An alternative method exists to cater for more factors at once, known as *minimization,* where the characteristics of individuals already randomized alter in a systematic manner the chances of a given subject being allocated to the different trial groups, so as to maximize the resulting balance of these factors [21,32].

d) Primary and secondary outcomes

Specific discussions of the most appropriate outcome measures for particular studies of incontinence will be dealt with elsewhere in this book; the purpose here is to define the general concepts of primary and secondary outcomes in the context of RCTs, which are relevant to both sample size determination and data analysis. The distinction between these two sets of outcomes depends on the context of the trial, and should be decided at the planning stage of the study. Primary and secondary outcomes should not be confused with the distinction between primary and secondary analyses of trial data, which will be discussed later. Primary outcomes are those viewed by the researchers to be of central interest. Trial results that lead to major changes in patient care will be based on primary outcomes.

The number of primary outcomes in a particular trial will depend on the nature of the interventions and the number of independent domains. The number of primary outcomes is usually limited to three, and rarely will there be reasonable justification for more than six. Sample size calculation is based on the primary outcomes and is unlikely to be based on more than two outcome measures. The number and nature of outcome domains in a particular study will vary depending on the study's perspective (e.g., those of patients, clinicians, regulatory bodies, and health care purchasers). In almost all situations, the outcome set should include a dimension representing the viewpoint of the patient (such as a questionnaire relating to symptoms and impact on quality of life) as well as an appropriate clinical outcome measure.

Secondary outcomes are the remaining outcome measures and could be relatively large in number. They are not the focus of the main study objectives and are rarely used directly in sample size estimation. Secondary outcomes are often subject to the dangers of multiple hypothesis testing, for which suitable corrections should be considered as described below. Analyses of secondary outcomes are often best viewed as exploratory, i.e., as hypothesis-generating exercises for which independent confirmation is essential.

e) Inclusion and exclusion criteria

Inclusion and exclusion criteria should provide a relevant population to address the study question, and together define the heterogeneity or homogeneity of the study population. Broadening the inclusion criteria can make a study more generalizable and facilitate recruitment. Making the entry criteria too broad, however, may dilute the effect being sought in the most suitable patients. If the study population is defined too narrowly with many exclusion criteria, applicability of the results may be limited and subject recruitment may be difficult.

Inclusion criteria govern what patient characteristics are required for eligibility to enter the study. Some exclusion criteria such as age, weight and gender are determined implicitly by corresponding inclusion parameters. Issues of patient safety determine other exclusion parameters (e.g., avoiding nephrotoxic drugs in patients with renal insufficiency). All parameters should be precise enough to allow the study to be reproduced by other groups of researchers.

The most important inclusion criterion is how the disease in question is defined. Eligibility criteria are critical to both the interpretation of the study and its reproducibility. If possible, established international criteria for the presence and severity of disease should be used. Inclusion criteria should screen for patients who are known 'non-responders' to the treatment being studied.

Including these patients can result in false negative clinical trial results since they do not have a reasonable expectation of improvement.

f) Informed consent

Peer review of protocols by a multidisciplinary team may include members of the scientific community, clinicians, pharmacists, the public, the legal profession and religious representatives. Each member of this team reviews the protocol from their particular type of expertise and in doing so aids in safeguarding patient health and well-being.

Informed patient consent is required for participation. The length and depth of detail in consent forms vary widely between institutions. In the most extreme, they involve exhaustive pages of information, which explain every alternative treatment and its pros and cons in detail. A general list of requirements for a consent form includes: name of the investigators and contact numbers, a detailed description of the new treatment and its known side effects, rationale for why the new therapy may be better compared to standard therapy. A summary table of the results of previous studies using the drug can be helpful in some circumstances. A statement that the patient may decline to be in the study with no subsequent consequence to their ongoing medical care is generally provided and whether or not remuneration is expected. An understanding that the patient will be randomly assigned to treatment should be included.

A review committee should be established prior to initiation of the trial. In addition to reviewing results of the study for safety monitoring they may conduct an interim analysis to ensure that a treatment is not producing unacceptable levels of side effects. Rules for stopping the study in this case are agreed upon usually prior to the start of the trial. Emergency procedures for unblinding a patient are put in place in the case of a severe side effect or concomitant serious illness.

g) Bias, blinding, and effects on validity

All human players in a clinical trial can introduce bias (systematic error), which can result in erroneous conclusions regarding treatment effects. Bias can occur in every aspect of a clinical trial from the process of randomization to observation of the outcome variables and the statistical analysis itself. Bias occurs because of previously conceived ideas held by those involved, which unconsciously affect their actions and observations. In addition to *observer bias*, an amount of *observer error* is inherent in outcome measures that require clinical interpretation. To avoid or limit bias, blinding should be employed whenever possible, with concealment of allocation and blinding of outcome assessors being the most important. *Blinding* is the process by which key elements of knowledge are withheld that can otherwise lead to bias. Blinding should not be confused with *concealment of allocation*, referring to withholding knowledge of assignment in advance, which is a prerequisite for the validity of any trial (Moher, Schulz et al, 2001; Altman et al, 2001) [28,36].

Unblinded trials are conducted in an open manner where both subjects and clinicians are aware of which treatment has been assigned. While certain types of therapy may require investigation in this manner (e.g., some surgical trials), there remains considerable opportunity for bias. Both subjects and clinicians may have preconceived ideas regarding the benefits of a particular treatment that can influence the reporting of symptoms and/or their outcome.

In a ***single blind*** trial, the subject is blinded to group assignment. It may be advantageous for the clinical staff to be aware of the assignment to allow them to monitor the health and safety of individuals, since the potential effects of the treatment (side effects) will often be known in advance. Single blinding ameliorates biased reporting of symptoms and/or side effects by subjects. However, clinical staff can influence data collection and change other aspects of subjects' care when they know which study treatment subjects are receiving. Moreover, particularly when a placebo is used in a trial, clinicians can systematically introduce co-interventions (or even the treatment under study itself) to the placebo group, thereby potentially diluting any differences between the trial arms.

In ***double blind*** trials, both parties who could influence outcome are unaware of group assignment. Often this is just the subjects and the clinical team responsible for their care. More generally, the term *double blind* relates to subjects and research personnel responsible for the measurement and assessment of outcome [21,28]. While this reduces potential sources of bias considerably compared with unblinded or single blind trials, it does introduce other levels of complexity. For example, safety monitoring must be performed by a third party.

Triple blind trials include blinding of subjects, outcome assessors, and those involved in data analysis. If the same persons carry out data analysis and safety monitoring, it can be difficult to ensure proper monitoring of complications and outcome. It might be argued that subject safety may not be properly ensured unless the monitoring committee knows which arm of the study is the treatment and which is the control (placebo). For the same reasons, clinical staff may not feel comfortable participating in such a study.

If investigators are aware of the results of interim analyses, this may cause bias by influencing how vigorously any given patient is recruited into the study. Another opportunity for bias occurs if an appreciable

number of subjects drop out or withdraw during a study, and fail to provide outcome data. This can be particularly problematic if withdrawal is related to group assignment and if it unequally affects one arm of a parallel group design. In this scenario, both the monitoring team and the trial data analyst must carefully consider the reasons for subject withdrawals.

h) Sample size considerations

Sample size should be calculated in the planning stage of all studies. There are many formal equations to assist in this process, details of which will not be given here [21,37,38,39]. Rather, the emphasis for this discussion is on the concepts involved and the information required for the calculations to proceed. Determination of sample size is not an exact science. Many decisions about design and analysis are interrelated with specifications for sample size, and the process does not have a single solution. This is no reason to abandon the exercise, but reinforces the need to include someone with appropriate statistical expertise in the research team.

There are three fundamental approaches to sample size calculation. One is based on the required precision of an estimate. The second requires that the study have adequate probability (power) of detecting a given (target) magnitude of effect. The third aims to demonstrate equivalence between treatment groups. In all cases, appropriate adjustment for attrition (loss to follow-up) should be performed.

The first of these approaches is relevant to both descriptive and analytical investigations. The basic issue is one of precision (measured by the standard error, SE) or margin of error (which depends on the SE but is more specifically defined as half the width of the 95% confidence interval [CI] around the estimate). The higher the level of precision specified in advance (i.e., the smaller the SE and the narrower the CI), the larger the sample size will need to be. However, the margin of error depends on the nature of the primary outcome variable, i.e., whether it is a continuous variable (such as maximum urinary flow rate) or a binary variable (such as the presence or absence of self-reported urge incontinence). For a continuous variable, the variability (standard deviation) of the measure must be estimated for relevant patients; this may be derived from some combination of clinical experience, the literature, or a pilot study. The larger the variability, the larger the sample size required. For a binary variable, its prevalence must be estimated in the population to be studied, since the SE for such variables depends on their prevalence.

The second approach, based on power, is the most commonly used. It requires similar prior information, including estimates of the variability for continuous measures and the magnitude of proportions for binary variables. In addition, it requires specification of three other quantities: the *significance level*, the *power*, and the *target difference*. The *significance level*, termed alpha, is conventionally, though not necessarily, set at 5%. *Power* is defined as the probability that the study will detect (as statistically significant at the alpha level specified) a given target difference between the groups, if such a difference exists. Power is commonly specified in the range of 80% to 90%, which implies a risk of not detecting the target difference of between 20% and 10%, respectively. For a trial involving anything other than minor risks and expenditure, a power closer to 90% than 80% would seem preferable [24], which leads to a larger sample size (as does a stricter alpha level of, say, 1%). This is most pertinent when a lack of statistical significance is obtained in a small trial, particularly when the sample size was not planned using a power calculation [21]. This is the basis for the adage that "the absence of evidence is not evidence of absence" [20]. A planned unequal allocation to the trial groups also requires an inflation of the sample size [21], as does interim analyses. By multiplying the number of significance tests performed, studies with interim analyses generally require stricter significance levels at each analytical point [20,37].

The *target difference* is the last, and arguably the most important, quantity that must be specified for the power-based approach to sample size calculation. The target difference is defined as the minimum difference needed for clinical significance. Clinical significance is an entirely different concept from statistical significance. Investigators must estimate the clinical significance as the magnitude of difference (in means or proportions) that would lead to a change in clinical management for the target group of patients. The smaller the difference, the larger the required sample size. Statistical significance means that the observed difference, whatever its magnitude, cannot reasonably be considered as being due to chance. Statistical significance (denoted by the p-value) represents the strength of evidence against the null hypothesis [40]. The degree of clinical significance can be inferred only with the additional information of a confidence interval for the comparison between groups.

The third general approach aims to demonstrate equivalence between trial groups [39]. The same specifications are made as in the power-based approach, except that instead of specifying a particular target difference to be detected, the calculation is centered on the magnitude of difference beyond which the researchers would no longer accept that the treatments are 'equivalent'. The study is designed to have adequate power to produce a confidence interval for the difference between the groups, which does *not* include values greater than this limit.

There is no single answer for sample size determination; often the calculation proceeds around a 'circle of specifications' (involving, say, power, target difference and sample size) many times, starting and stopping at different points. For instance, it is not uncommon to commence with the 'textbook' approach of specifying power and target difference (along with alpha and the standard deviation) and calculating the sample size, then to reverse the argument by starting with how many subjects could be recruited and determining what differences could be detected with various probabilities! Furthermore, the ideal of the target being the *minimum* for clinical significance cannot always be met; rather, the aim in practice is to produce a convincing argument (among the researchers themselves, and also to funding bodies and regulatory agencies) that the sample size has an adequate chance of detecting differences that are (a) feasible, and (b) worthwhile detecting in clinical terms. A common failing is selecting a target difference that is too large, often derived from differences that have been observed or published previously rather than based on considered clinical judgment. Preliminary investigations (often termed 'elicitation exercises') into the levels of treatment effects that patients themselves consider worthwhile should be carried out much more commonly than is the case at present. Likewise, more evidence is required concerning the relationships between the responsiveness (sensitivity to change following treatment) of clinical and patient based outcome measures.

i) Pragmatic and explanatory trials

There is an important distinction between *pragmatic* and *explanatory* trials [41, 42] and correspondingly, between *intention-to-treat* and *per-protocol* approaches to data analysis [20, 37]. In pragmatic trials, data are analyzed by intention-to-treat, according to the group to which subjects were randomized, regardless of the extent of compliance with the intended treatment. In explanatory trials, data are analyzed accounting for compliance. This per-protocol approach may exclude serious non-compliers, analyze data according to treatment actually received, or allow for degree of compliance in a statistical model. At first sight, the explanatory approach appears more attractive. However, there are considerable limitations to the explanatory approach, particularly when the intention is to draw inferences from the trial to wider clinical practice.

The purpose of randomization is to produce groups that are, on average, comparable. A per-protocol analysis retains this property only in the unlikely situation when non-compliance is unrelated both to the patient's underlying state of health and the treatment received [20]. The intention-to-treat approach in pragmatic trials retains the full benefits of randomization and has the advantage that the comparison will more closely reflect the relative effectiveness of the treatments when applied in real clinical practice, where non-compliance obviously occurs [43]. In pragmatic trials, the interventions are designed to be as close as possible to treatment options in clinical practice (including 'cascades' of patient management choices) and entry criteria are usually relatively liberal. Pragmatic trials may involve a wide variety of outcome domains, including patient-completed questionnaires, and an economic evaluation of outcomes. Because of intention-to-treat data analysis, pragmatic trials will tend to yield lower estimates of treatment differences than explanatory trials. It may be of interest to gauge the effect of treatment given full compliance, so full data analysis may incorporate elements of intention-to-treat and per-protocol approaches [20].

The follow-up time for a trial should be at a fixed point (for logistical reasons, this is in practice often a short time window) relative to randomization rather than when treatment was actually received, since again this is the only way of ensuring a valid comparison. The planned timing of follow-up should allow for any likely delays in receiving treatment, e.g., due to surgical waiting lists.

In summary, it is established practice that unless there are strong reasons to the contrary the primary analyses (for both primary and secondary outcomes) of an RCT should be on an intention-to-treat basis [28, 36]. Secondary analyses incorporating non-compliance and/or which treatment was actually received may be justified in addition to the primary analyses. Appreciable loss to follow-up in a trial (which is not the same as non-compliance with intended treatment, lack of efficacy, or the observation of adverse events) may present serious problems both in terms of generalizability of the findings to the wider population and, in the case of differential loss to follow-up across treatment groups, to the validity of the comparisons. Indeed, strictly speaking any missing outcome data means that not all of those allocated to the various randomization groups can be included in the analysis [28, 44] and this might lead to the conclusion that the term 'intention-to-treat' should only be used if follow-up is complete. In practice complete follow-up occurs only rarely. Under current guidelines, intention-to-treat relates more to the broad strategy adopted by the researchers for data analysis [45]. Results should always be accompanied by a full and clear statement of how deviations from intended treatment and missing outcome measures have been handled in the analysis. The discussion should include how missing outcome data may have affected the conclusions [44]. Sensitivity analyses can be used to test the exclusion of, or assumptions about, missing values; practical examples of such analyses are becoming more common [46].

j) Data analysis

This section will not contain any technical details of statistical methods, which are available in standard texts [21, 47] but rather will summarize concepts of data analysis. The emphasis here will be on RCTs, although many of the complex methods mentioned (e.g., multiple logistic regression analysis) are used in similar ways to analyze observational data. Appropriate techniques of data analysis will depend on the nature of the outcome variable. In practically all situations, hypothesis tests should be two-sided, rather than one-sided. One-sided tests are only appropriate if a difference in one direction is not just unlikely, but would not be of interest.

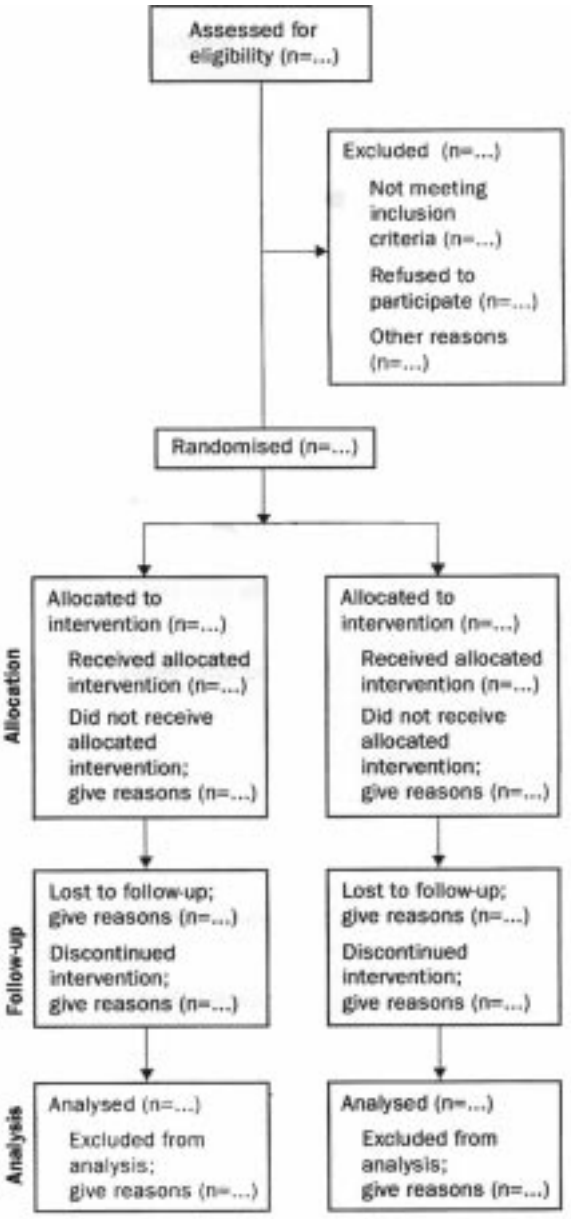
Regardless of the type and complexity of statistical techniques used in analysis, the general underlying principles behind hypothesis testing and estimation apply. In particular, the statistical significance of a hypothesis test should be interpreted critically. The actual p-value should be considered, rather than just whether or not it is below an arbitrary threshold such as 5% [28]; indeed, the p-value is better considered a measure of the strength of evidence against the null hypothesis, on a continuum or 'shades-of-gray' [40]. The direction and magnitude of the trial comparison should be presented with an appropriate confidence interval to indicate the possible clinical significance and precision of the comparison [28].

Data analysis for numerical outcome variables may use parametric or non-parametric methods. Simple parametric methods require that the data follow a normal distribution, while non-parametric methods do not have this requirement. Parametric methods of testing mean values include t-tests, confidence intervals for differences between group means, and analysis of variance. Regression techniques address more advanced issues such as stratification in randomization and allowance for baseline measures. Non-parametric methods include the Mann-Whitney test to compare two independent samples as in a parallel groups trial and the Wilcoxon matched-pairs signed-ranks test for paired data such as from a crossover trial [23]. Binary outcome variables can be analyzed using chi-square tests and confidence intervals for comparing proportions, and multiple logistic regression [48]. For time-to-event data (such as survival data), methods of data analysis include life tables, Kaplan-Meier survival curves, log rank tests, and Cox's proportional hazards regression [49].

How, then, should the analysis of data from an RCT proceed? An outline of the various stages of data analysis can be gleaned from the CONSORT statement [28, 36]. The following discussion will concentrate on the underlying concepts of data analysis at a particular follow-up time relative to randomization, and considers initially the simplest case of just two trial groups. Multiple treatment groups will be covered briefly, but repeated measurements on outcomes and interim analyses involve considerably more complex methods of planning and analysis, for which expert help is essential [20, 23, 37, 50].

The first stage of data analysis is to address the representativeness of randomized subjects compared to the target population of eligible patients. The number of eligible patients who were and were not randomized should be provided, along with reasons for the latter. The presentation of this information is facilitated by use of the ***CONSORT flow diagram*** [28, 36]—indeed, its use is associated with improved quality of reporting of trials generally [51]. Descriptive statistics should be given of important characteristics of health care professionals approached for involvement in recruiting subjects to the trial, both for those taking part and those declining.

The second stage of data analysis is to compare the two groups at randomization (baseline) including demographic, prognostic, and outcome variables. A common error at this point is to rely on statistical testing for these comparisons [20, 21, 37]. If the randomization procedure has been performed correctly, then any statistically significant differences in baseline characteristics must be due to chance. Statistical testing of this kind is *not* a test of the comparability of trial groups; rather, it is a test of the allocation procedure [20, 21,37] It may be seriously misleading, particularly if lack of a statistically significant difference for a given characteristic is taken to imply comparability. Trials are not designed to detect potentially important differences in baseline characteristics that might be large enough to influence the comparison of the outcomes between the trial groups. The magnitude of this potentially influential difference for a baseline measure depends on the strength of its relationship with the outcome, and not on a p-value at randomization. Therefore, baseline comparability is best assessed by simply obtaining descriptive statistics for the groups and making a judgment as to whether any observed differences are likely to be influential or not. If differences are likely to be influential, they should be considered in the analyses. Notable exceptions to this are baseline measures of the outcome variables, which should be considered in the analysis regardless of the situation at baseline, since removing variance in the outcome measure that is purely attributable to differences between individuals at baseline has potentially marked benefits in terms of precision and power [20]. Investigators should consider stratifying the randomization on any strongly prognostic variable (for reasons of efficiency rather than bias). Since there are practical limitations to how many variables a trial can stratify for, a technique known as minimization



Flow diagram of the progress through the phases of a randomised trial

may also be considered [20, 32]. Any variables stratified or minimized at randomization should be allowed for in the analysis [20].

The next stage of data analysis is to perform the primary analyses for the outcome variables. Primary outcomes should initially be analyzed by intention-to-treat comparisons of the groups as randomized, both using hypothesis tests for statistical significance and confidence intervals (CIs) for comparisons between the groups to assess clinical and statistical significance, usually adjusting for baseline measurements of the outcome variable. With a small number of primary outcomes, multiple testing is not a concern. However, when a large number of statistical tests are performed for secondary outcomes, corrections to the observed p-values should at least be considered. Similar issues of multiple testing for different outcomes are involved when there are more than two groups.

The most commonly used procedure for multiple testing is the Bonferroni correction [20, 21, 52]. The Bonferroni correction is fairly conservative in reducing the risk of a statistically significant effect occurring purely by chance, at the cost of reduced power for individual outcomes. This is particularly pertinent when, as is usually the case, the outcomes are positively associated with one another. While there are alternative procedures that improve this deficiency, none of them are entirely satisfactory [20]. It is emphasized that whatever strategy is adopted to deal with multiple testing, the major errors are to rely solely on p-values rather than present CIs as well, to over-simplify the presentation of p-values to just "NS" or "$p<0.05$" rather than to quote the actual p-values, and above all to report selectively the results of significance tests.

Another example of a "multiplicity" is where there are more than two treatment groups, e.g., when different doses of a drug are being investigated or when more than one 'active' procedure is being compared with placebo [20]. Similar issues to multiple testing of different outcomes are involved here, but there are a greater variety of commonly used procedures available to deal with the central concern of finding a difference purely by chance. Standard methods for dealing with this multiple comparisons problem include the procedures attributed to Tukey, Newman-Keuls and Dunnett [20, 23, 53].

More complex primary analyses adjust for baseline measurements and potentially important prognostic variables (including but not exclusively those that were unbalanced at randomization). They may also involve adjustments for center effects and the investigation of differential treatment effects across centers in multicenter trials [24]. The correct approach for continuous outcome variables is to use the (regression-based) technique known as the analysis of covariance [20, 37]; the equivalent approach for binary outcomes is to use logistic regression. A commonly employed alternative for continuous outcome variables is to analyze simple change scores from baseline to follow-up (either in absolute or percentage terms), but for reasons of both bias and precision this is inferior to regression methods [20, 37]. It is good practice to present both the (unadjusted) simple intention-to-treat results alongside those from the regression methods. In any case, the results from alternative analyses such as these should be compared in a sensitivity analysis of the conclusions [24].

Secondary analyses of trial data include per-protocol analyses with adjustments using regression methods for pertinent process measures such as degree of compliance with the allocated treatments. Secondary analyses

also include planned subgroup analyses, such as different intervention effects across different age, ethnic, or disease severity groups. Subgroups should be analyzed by using appropriate interaction terms in regression models [28, 37] Using interaction terms rather than performing repeated, separate, subgroup-specific analyses considerably reduces the risk of false positive findings. [54]. Subgroup analyses should be carried out sparingly, specified in advance (preferably with a clinical rationale), and above all should not be reported selectively [28, 54].

k) Reporting of randomized controlled trials

The CONSORT statement is specifically designed to provide standards for reporting RCTS [28, 36]. Adherence to these guidelines and the use of flow diagrams in particular is associated with improved quality in reporting of RCTs [51, 55]. Errors in presentation of statistical information are extensively covered in many textbooks [21, 47]. This section will emphasize the most important points on reporting of RCTs, to ensure an objective and comprehensive presentation of the trial itself, and also to facilitate any subsequent synthesis of research evidence including formal meta-analyses of RCTs. Meta-analyses are themselves the subject of separate reporting guidelines, the QUORUM statement [56]. Guidelines for reporting studies on diagnostic tests (the START document) will be published in 2001. However, such guidelines are not a panacea [31] deficiencies in reporting are still common [55].

The CONSORT statement recommends clear statements about the objectives of the trial, intended study population, and planned comparisons. Subgroup or covariate analyses should be clearly specified and justified. The method of randomization should be stated, as should the unit of randomization; in most cases, this will be the individual subject but occasionally an aggregate group of subjects will be allocated jointly in a cluster randomized design [35]. Cluster randomized designs are also now the subject of separate reporting guidelines [57]. For all trials, specifications for the sample size calculation (primary outcomes, target differences, etc.) should be stated and justified. In addition, the precision actually obtained in a study must be presented. This requires confidence intervals as well as the observed p-values, at least for primary outcomes but preferably for all outcome variables. The principal confidence intervals should be for comparisons *between* the groups, rather than for differences in the outcomes *within* the trial groups [20, 21]. Results should include a trial flow diagram, with numbers and reasons for the exclusion of eligible patients, randomization, and subsequent losses to follow-up [51]. Protocol deviations should be described and explained [37]. Finally, the discussion should include a brief summary of the trial's findings, possible explanations for the results, interpretation of the findings in light of the literature, limitations of the trial including internal and external validity, and the clinical and research implications of the study [28].

l) Conclusions

In conclusion, it is crucial that those intending to embark on research into incontinence plan the details of the study in advance. Many of the decisions to be made involve statistical issues; therefore it is vital that someone with relevant expertise is involved in the planning from the start. Statistics in general has been described as a combination of mathematics, logic and judgment [20], and this applies particularly to clinical trial design, conduct, analysis and reporting. Naturally, formally qualified biostatisticians are not the only professional group with the necessary expertise to address these issues, particularly since in the planning of studies the above three characteristics are probably stated in increasing order of importance. However, individuals with relevant statistical expertise are in a good position to contribute to research projects in these ways, if they are consulted sufficiently early in the process including at the piloting stage.

Furthermore, the benefits of such expertise will only fully be derived if the individuals are involved on an ongoing basis in the conduct of the trial. This is equally true of all the disciplines relevant to studies of health care technology and organization, including social scientists and health economists as well as statisticians and clinicians. Increasingly, the major funding bodies and international journals expect a sufficiently multidisciplinary team to carry out and report on health services research. If for no other reason than because of their central position in influencing the purchasing and provision of health care, this is especially important for randomized controlled trials.

4. Outcome research in patients with LUTS, including incontinence

Introduction

No single measure can fully express the outcome of an intervention. While every clinical trial must have a primary endpoint, complete collection and reporting of data is essential to progress in understanding and treating disease. It is good to know that a drug or procedure appears to be "safe and effective". It is better to know that treatment A is superior to treatment B. It is ideal to understand why one treatment is better than another—to understand why a treatment works for a particular patient and not for another. Understanding at this level requires correlation of outcomes with anatomic and physiologic variables. This degree of detail is often not obtained and is only rarely reported. Reports tend to

concentrate on success or failure in achieving the primary endpoint (e.g., cure of stress incontinence); however, to understand outcomes, we need detailed data on improvement and deterioration in anatomy, symptoms, lower urinary tract function, complications of the intervention, and the effect on quality of life. Both subjective and objective measurements should be recorded and reported. Functional changes can occur without obvious symptoms and symptom improvement can occur without urodynamically demonstrable changes; therefore, it is necessary to correlate the subjective response with physiologic response if we are ever to fully understand therapies. Perceptions of the patient, doctor or therapist are frequently at variance and this must be reported. Patients' expectations may also influence the outcome of a study [58].

To obtain maximum information, it is important to chose and define the correct endpoints at the beginning of the study. Outcome variables should be chosen so that they will be relevant and may be incorporated into practice at the end of the study. We agree with recommendations from the ICS Standardization Committee [12]. For clarity, we have structured the recommendations as follows:

a) Baseline data:

b) Observations:

1. Patient's observation/Subjective measures

2. Clinician's observation/Objective measures

c) Tests

1. Quantification of symptoms—void diary and pad tests

2. Urodynamics

d) Follow-up

e) Quality of life measures

f) Socioeconomics

g) Toward a simple, inclusive outcome measure

a) Baseline data/demographics

Data collection in clinical research begins with complete demographic description of the subjects including age, race, sex, duration of symptoms, prior treatments, comorbidities, medications, etc. It is prudent to inquire about the use of naturopathic and alternative medicines since these can impact metabolism and clearance rates of certain conventional pharmaceutical agents. Obstetric and gynecologic history is important in women. Recommendations for minimum data collection are made in the proceedings of the NIH Terminology Workshop for Researchers in Female Pelvic Floor Disorders [59]. While few trials will be large enough to analyze the effect of these demographic factors on outcome, the potential use of meta-analysis makes a complete database valuable.

b) Observations

1. Subject's observation and subjective measures:

Validated patient completed symptom questionnaires and other validated instruments are recommended in trials for LUTS and incontinence. In addition to specific symptoms, the respondent's overall opinion of the condition should be included. Different methods to obtain this measure include: a question with a forced choice, a graded response, a statement with a Likert scale agree-disagree response, and a statement with a visual analog graded scale response. These global response instruments should have a symmetrical design with equivalent opportunity to express a negative as a positive outcome. Questionnaires should always be administered in private and by a third party. An ideal instrument would record all symptoms related to the lower urinary tract and relevant associated organ systems. At a minimum, this would comprise:

- Incontinence, stress induced
- Incontinence, urge induced
- Incontinence, other
- Frequency and nocturia
- Urgency
- Voiding/emptying symptoms
- Protection (e.g., pad use)
- Coping measures
- Pain
- Sexual function
- Bowel function

An ideal instrument would record both the objective severity of the symptom (e. g., how many times nocturia) and the impact or bother produced by the presence of the symptom (e. g., much greater for the individual who is unable to fall back to sleep easily). There is no general symptom measure with established methodological reliability. Therefore, researchers should clearly describe their instrument and procedure and provide reliability data or indicate their absence. As there is no one universally accepted, "ideal" instrument, trials are often conducted using multiple instruments to assess different domains. In the future, consideration may be given to use of the International Consultation on Incontinence Questionnaire (ICI-Q long form). Developed in response to the first Consultation, this is one instrument that meets these specifications. Unfortunately, clinical experience with this new instrument is lacking. Committee 6 provides a detailed discussion of available validated symptom scores.

RECOMMENDATIONS:

- **One or more validated symptom instruments should be chosen at the outset of a clinical trial to accurately define baseline symptoms and other areas in which the treatment may produce an effect.**
- **The same instruments should be administered after intervention throughout follow-up.**

2. CLINICIAN'S OBSERVATION AND OBJECTIVE MEASURES:

We have traditionally included functional, primarily urodynamic, data in the evaluation of lower urinary tract disorders. It is equally important to investigate the possible presence of anatomic changes in the lower urinary tract and its supporting structures. For example, in evaluating the results of stress incontinence surgery, there are few papers that report both anatomic and functional results adequately. Therefore, while one may get some idea about the effectiveness of a particular operation, it is impossible to determine if failure occurs because of technical factors (recurrent hypermobility) or due to an inherent limitation of the procedure (intrinsic sphincter dysfunction). Similarly, reports of biofeedback training for incontinence provide data about continence after intervention little information about muscular function is provided. Do patients fail because the intervention itself was unsuccessful (pelvic muscles remain weak) or because of an inherent limitation of the technique (incontinence persists despite successful muscular reeducation)? We can only make major progress in treating lower urinary tract dysfunction by merging a full understanding of the patient's symptoms with a detailed assessment of function and a complete description of anatomy. In parallel, a new concept for classification of objective observations in lower urinary tract disorders that recently has been suggested [60], visualizing all types of lower urinary tract dysfunction as being neurogenic—either primarily neurogenic because of disease/damage in the nervous system or secondarily neurogenic because of disease/damage in the lower urinary tract and/or their supporting structures. Only complete evaluation of both structure and function will allow us to ultimately devise an optimal classification of LUT disorders.

RECOMMENDATIONS:

- **Clinicians' observations of anatomy should be recorded using standardized, reproducible measurements.**
- **Pelvic muscle and voluntary sphincter function should be reported using a quantifiable scale.**
- **These measures should be repeated after intervention and correlated with primary clinical outcome measures.**

c) Tests

1. QUANTIFICATION OF SYMPTOMS—BLADDER DIARY AND PAD TESTS:

The diary (voiding diary, bladder diary, or frequency-volume chart) is a self-monitored record of selected lower urinary function that is kept for specific time periods. Variables include fluid intake, episodes of incontinence, pad use, voiding frequency (diurnal and nocturnal), total voided volume, mean voided volume, and the largest single void. Accuracy depends on the subject's ability to follow instructions. Reproducibility depends on the parameters used and improves with the number of days that self-recording is obtained. Diaries are reliable for assessing the number of incontinent episodes. In most instances, a single 24-hour diary is sufficient. Longer diaries (48-72 hours) are more reliable but have decreased subject compliance. The circumstances under which a diary is kept should approximate everyday life, and should be similar before and after intervention to allow for meaningful comparison. Reliability and validity data for specific diaries should be provided if available, or their absence indicated [61], [62] [63], [64]. The period of time the diary was used should be noted [65].

Urinary diaries are important in the evaluation of LUTS because they document functional bladder capacity, diagnose diurnal and nocturnal polyuria, and diagnose fluid restriction that may affect continence or other LUTS. Incontinence studies often use the number of incontinence episodes on the diary as the primary endpoint. While this may provide a clear endpoint, it does not provide the information necessary to interpret the data completely. Voided volumes are critical in this regard. Might urge incontinent patients fail to improve with anticholinergic medications because bladder capacity was normal at the outset? Is improvement in continence correlated with improvement in bladder capacity? If we are to understand our interventions completely, we need complete data.

Pad tests can be divided into short-term tests, generally performed under standardized conditions as office tests, and long-term tests, generally performed at home over 24 to 48 hours. Pad-weighing quantifies the amount of incontinence. 24-hour pad tests are reliable instruments for assessing the amount of urinary loss. Increasing the test duration to 48 or 72 hours increases reliability but decreases subject compliance. For short-term tests, the experimental conditions should be described. Standardized bladder volumes are recommended. The pad test quantifies incontinence in a way no other measure can replicate; therefore it can provide the key link in understanding outcome. A patient who experiences a decrease in the number of incontinence episodes from four to two per day may not be satisfied if the volume of urine loss is high. Similarly, cure of incontinence may not have a great impact on a patient with trivial volume of urine loss at baseline.

RECOMMENDATIONS:

- **Clinical trials of incontinence and LUTS should include bladder diaries as an essential baseline and outcome measure.**
- **The diary should include measured voided volume (for at least one day if a multi-day diary is employed).**
- **24-hour diaries are adequate for most studies.**
- **Clinical trials of incontinence and LUTS should include a pad test as an essential baseline and outcome measure.**

2. URODYNAMICS:

Detailed recommendations on the indications and conduct of urodynamic investigation are found in the report from committee 7. This discussion is limited to the role of urodynamics in clinical research. Urodynamic studies take on two major roles in research—describing subjects at entry and defining outcome. Most clinical trials do not enroll subjects based on specific urodynamic diagnoses but rather based on reported symptoms. This is appropriate because:

- Urodynamic tests add significant cost to clinical trials
- Urodynamic tests are not universally available
- No urodynamic test has 100% sensitivity or specificity

Subjects should not be stratified by urodynamic diagnosis. With the possible exception of a high detrusor leak point pressure in children with spina bifida, there are no studies that clearly define a predictive role for urodynamic testing in the management of LUTS and incontinence. One of our primary research goals should be to collect data to determine the predictive value of urodynamic testing prior to intervention.

We recommend the use of urodynamic studies to accurately characterize baseline lower urinary tract function and dysfunction. This information greatly facilitates understanding of the underlying disease and the actual effect of treatment, and even provides insight that can help improve urodynamic tests. How will we advance our understanding of the pathophysiology and treatment of urge incontinence if we do not perform cystometry during clinical trials?

How can we improve our selection of patients for surgical treatment of stress incontinence if we don't carefully study which patients succeed and which fail? Urodynamic tests are among the best tools currently available to understand the basic physiology and mechanisms of disease; these tests must be a fundamental part of our research efforts. Nevertheless, urodynamic testing is far from perfect in representing lower urinary tract function and dysfunction, and our research efforts should also be directed toward the development of new and better tools.

Interpretation of urodynamic signals remains an art, the art of detecting artifact. Direct interpretation of urodynamic data without careful and critical investigation of the accuracy and reproducibility of the measurements is inappropriate. Accurate urodynamic interpretation requires continuous observation and quality control of all signals with plausibility control. In addition to anatomical and physiological knowledge at least some basic knowledge of biomechanics is needed (e.g., muscle mechanics, fluid dynamics). The elementary physical-biomechanical properties of parameters and measurement should be understood. Procedures for performing urodynamic studies must be carefully standardized to ensure that consistent techniques are used for different subjects; this is particularly critical for different centers in a multicenter study.

The exact same technique must be used at baseline and follow-up. Studies with urodynamic endpoints require an evaluation of whether or not the study reproduces the symptom under investigation. Another source of potential error is investigator bias. In multicenter studies, this can be avoided by using a central reader for urodynamic tracings, after detailed annotation by the primary observer. In all studies the reader should be blinded.

RECOMMENDATIONS:

- **At this time, clinical studies should enroll subjects by carefully defined symptoms, not urodynamic findings.**
- **To determine the predictive value of urodynamic tests, urodynamics must be performed at baseline but subjects enrolled without prejudice of urodynamic test results.**
- **In the ideal clinical study, urodynamic tests are performed at baseline and exit to correlate symptom changes with physiologic changes.**
- **When these ideal conditions cannot be met, urodynamic tests should be performed on a subset of the larger group.**
- **In all trials, standardized urodynamic protocols (based on ICS recommendations) are defined at the outset. In multicenter trials, urodynamic tests are interpreted by a central reader to minimize bias.**

d) Follow-up

Minimal standards for evaluation of treatment outcomes in urinary incontinence have been presented by Blaivas et al. [16] in a report approved by the American Urological Association and the Urodynamics Society. Their recommendations are in agreement with the ICS, although they are more detailed and specific for certain patient groups and disorders. In addition to standard pre- and post intervention evaluation, they recommend evaluation of surgical, prosthetic, and implant therapies no less often than 1 to 3 months and 12 months after treatment, and thereafter at yearly intervals for as long as possible.

The method by which data were collected should be specified, e.g., prospective questionnaires or retrospective chart review. Individuals collecting data should be identified, e.g., independent research nurse, clinician. The interval between the time of evaluation and the last treatment should be specified. The exact type of data collected at each time point in follow-up will vary by individual studies and should be defined at the study's outset. Some general data are mandatory to collect at each post-treatment interval: the total number of patients treated, the number of subjects actually evaluated in the study, and the total number of subjects lost to follow up and the reasons why they were lost. Indications for retreatment and the time interval since the last treatment should be specified. Efficacy assessment should be done at a specific time interval after the last treatment. The protocol should further specify the criteria by which treatment success or failure is determined.

e) Quality of life measures

Health related quality of life (HRQOL) is a multidimensional construct that refers to an individual's perceptions of the effect of a health condition and its treatment on quality of life. Primary domains of HRQOL include physical, psychological and social functioning; overall life satisfaction and well-being; and perceptions of health status. Secondary domains include somatic sensations (symptoms), sleep disturbance, intimacy and sexual functioning, and personal productivity (e.g., household, occupational, volunteer, or community activities). It is important to know not only how successfully treatments eliminate incontinence, but also how a treatment affects a patient globally. Nonetheless, HRQOL can never be the sole outcome of clinical research. Our focus must always be on how successfully we have treated the target condition or symptom. If a treatment is effective but does not improve HRQOL due to some adverse effect, the treatment can be improved. The combination of HRQOL data and more traditional objective endpoints will allow us to understand the reasons behind our success and failures.

Three measurement approaches are commonly used to assess HRQOL: generic, condition-specific and dimension-specific. (These instruments are explained in the report from committee 6. Here only a few aspects of relevance to research are discussed.) Generic HRQOL instruments are designed to be used across groups by having established age and gender norms. Condition-specific instruments are designed to measure the impact of a particular condition. These instruments tend to be more responsive than generic instruments in detecting treatment effects. Symptom scales are considered condition-specific; generally, these scales should include measurement of the presence of a symptom as well as the "bothersome" or "troublesome" nature of it. The majority of generic and condition-specific instruments are multidimensional, i.e., they measure more than one aspect of HRQOL. Dimension-specific instruments, in contrast, are designed to assess a single component of HRQOL, such as emotional distress. The trend in assessing HRQOL outcomes has been toward the use of a multidimensional generic and/or condition-specific instrument, supplemented with dimension-specific instruments as needed.

The selection of an HRQOL instrument should be based on the purpose of the study. Descriptive epidemiological studies should consider both generic and condition-specific instruments. Intervention studies should include a condition-specific instrument. Dimension-specific instruments should be used when more detail about a specific subdomain of HRQOL is desired. Researchers should define HRQOL for their study, clearly describe their instrument(s) and data collection, and provide reliability data if available. Selected instruments should be reliable and sensitive. In adopting HRQOL instruments, results obtained in the study population should be compared with published norms. If a new instrument will be used in a study, adequate pretesting should be done to establish its clinimetric characteristics (e.g., reliability and sensitivity) and an established instrument should also be used to provide a comparison.

RECOMMENDATIONS:

- **Research in incontinence and LUTS should include both generic and condition-specific HRQOL instruments.**
- **Changes in HRQOL after therapy should be correlated with changes in individual symptoms, and with physiologic and anatomic outcome measures to learn how the particular therapy is working.**

f) Socioeconomic data as outcome measures [66]

A full discussion of the economic impact of urinary incontinence is detailed in the report of committee 14. We recommend that cost analyses be planned with clinical studies whenever possible. Costs are to some degree artificial, in that they are established by economic and political factors that are subject to change at any time. However, when basic units of work, time, and resources are carefully defined, cost analyses remain useful even if market forces change monetary costs in an unforeseen manner.

Economic measurements are divided into two broad categories: descriptive and comparative data. Descriptive data include the socioeconomic cost caused by the disease and its current treatment, and comparative data provide an economic evaluation of different treatment strategies and interventions.

Descriptive data: Cost of illness studies that are prevalence-based or incidence-based provide a baseline against which the economic consequences of a new intervention can be measured. They provide useful basic information for policy makers, as well as for researchers developing new treatments. Generally, such studies take a societal perspective and include direct costs (i.e., costs to the health care system or to patients) and indirect costs (e.g., loss of productivity due to disease or treatment, premature mortality).

Comparative data: Economic evaluations allow comparison of different courses of action in terms of their costs (inputs) and their consequences (outcomes). There are four types of evaluations:

- Cost Minimization Analysis (CMA) is appropriate when two interventions have an identical outcome and only costs need to be compared.
- Cost Effectiveness Analysis (CEA) is appropriate when two interventions for the same disease have similar outcomes, but to different degrees. Outcomes are measured by variables such as cure, function restored, symptom-free days, events avoided, or life-years saved. Costs and outcomes are compared and the additional cost to achieve an incremental unit of effectiveness is calculated.
- Cost Utility Analysis (CUA) is similar to cost effectiveness analysis, but outcome is expressed as a single measure incorporating survival and QoL, usually quality-adjusted life years (QALY). Cost utility analysis allows comparisons of treatments in different diseases.
- Cost Benefit Analysis (CBA) expresses the value of the outcome directly in monetary terms and allows comparison of interventions both inside and outside healthcare.

Costs: Costs of an intervention are a function of resource utilization (quantities) and cost (price). Data on utilization of relevant resources is usually collected directly within a trial, while costs are calculated outside the trial. Costs should be fully allocated including overhead and depreciation.

Economic evaluation: Socioeconomic decisions depend on knowing both the cost and outcome of therapies. It is not easy to define a single outcome measure that is acceptable and meaningful to patients, physicians, and health care purchasers. Ideally, a composite construct incorporating all the dimensions of disease would be most useful for economic evaluation.

RECOMMENDATIONS:

1. The type of economic evaluation should be chosen before starting a trial, based on the specific objectives to be addressed. Analysis is based on intention-to-treat, and dropouts must be handled in the same way as for the primary outcomes analysis of the trial.

2. Typical resource use to be collected in a study is:

a. Direct

- personnel (physician, nurse, technician) time;

- diagnostic tests, laboratory analyses;

- treatments (drugs, physiotherapy, etc.);

- treatment of side-effects;

- surgical interventions;

- days of hospitalization;

- miscellaneous (pads, laundry, etc);

b. Indirect

- days of absence from work.

3. Very few economic evaluations have been done in the field of urinary incontinence and more experience is needed to make firm recommendations. Researchers should consider both a condition-specific outcome measure for economic evaluation, and a quality of life instrument with utility properties to allow for comparison to other diseases.

g) Putting it Together: toward a simple, inclusive outcome measure

One group of researchers has proposed a simple scoring system to assess outcome of incontinence therapy combining the important non-invasive outcome measures—patient's perception, voiding diary, pad test [67]. Each item is scored "0" for cure, "1" for improvement, and "2" for failure. The total score (with range of 0 to 6) represents cured, improved (good, fair and poor) and failure (same or worse). This system has been used to assess the results of sling surgery and injection therapy

[68] [69] This system has two important advantages, in that it is applicable to all types of incontinence and therapies, and it is inexpensive with no special equipment required. Consideration should be given to further investigation of this system in clinical trials.

Another instrument that has been developed and used in clinical trials is the SEAPI-QMM system [70]. The acronym SEAPI includes subjective and objective assessment of stress incontinence, emptying ability, anatomy, protection, and inhibition (urgency). The QMM includes a validated quality of life questionnaire, a mobility assessment and a mental status assessment. This is more cumbersome and expensive to use in its entirety but is more detailed; it may also be used without completing all domains.

III. CONSIDERATIONS FOR SPECIFIC PATIENT GROUPS

1. Men with LUTS, including Incontinence

We concur with recommendations in the ICS report on "Outcome Measures for Research in Adult Males with Symptoms of Lower Urinary Tract Dysfunction" [14]. The unique factors influencing research on lower urinary tract symptoms in adult men are the presence of the prostate and the possible presence of benign prostatic obstruction (BPO).

a) Prostate Size: If treatment could potentially change prostate volume, measurements of volume should be made before and after treatment. The method used to measure volume and its reliability and validity should be provided if available or their absence indicated. Timing of post-treatment testing depends on the treatment's mechanism of action. Correlation of outcome with change in prostate size should be reported. Consideration should be given to stratifying patients by prostate volume, as it is clear that response to medical therapy, at least, may be volume dependant.

b) Bladder outlet obstruction: As discussed in more detail by committee 7, routine urodynamic studies cannot be recommended prior to clinical trials on LUTS and incontinence. Urodynamic studies have not been demonstrated to predict response to treatment. However, detrusor pressure-uroflow studies (pQS) are highly desirable and should be included to document the presence and degree of change in bladder outlet obstruction whenever feasible. Results should be presented as stated in the ICS 1997 "Standardization Report on Pressure Flow Studies of Voiding, Urethral Resistance and Urethral Obstruction." Change in flow rates in response to treatment is sensitive, but the degree of change is meaningless unless pre-treatment detrusor voiding pressure is known. A slight decrease in outlet resistance might produce a pronounced increase in maximum urinary flow rate if outlet resistance is low before treatment. Conversely, a large decrease in outlet resistance might result in only a small increase in maximum urinary flow rate if outlet resistance was high before treatment and an element of obstruction persists. Reduction of residual urine volume after treatment indicates improvement of outlet conditions; such a reduction is likely to be more important in assessing treatment response than in establishing a diagnosis. Methods used for the assessment of bladder outlet obstruction should be stated and reliability and validity data should be provided if available, or their absence indicated. At this time there is not conclusive data demonstrating a differential response to treatment by degree of outlet obstruction and we do not recommend stratifying patients. This is an important area for future research.

Recommendations:

- **If treatment could change prostate volume, measurements of volume should be performed before and after treatment.**
- **Consider stratifying patients by prostate volume.**
- **Whenever feasible, detrusor pressure-uroflow studies should be performed before and after treatment to document the presence and degree of change in bladder outlet obstruction.**

2. Women with LUTS and incontinence

We concur with recommendations for outcome research in women by Blaivas et al. in 1997 [16] [17]. We also refer to the ICS recommendations for outcome measures in women with lower urinary tract dysfunction [13] and the Proceedings of the NIH Terminology Workshop for Researchers in Female Pelvic Floor Disorders [59]. Unique factors influencing research on lower urinary tract symptoms in adult women include (1) ***hormonal effects*** on the lower urinary tract; (2) ***obstetric history*** and the influence of vaginal childbirth on the development of pelvic floor disorders; (3) assessment of ***pelvic organ prolapse*** and other measures on physical examination; (4) ***definitions of outcomes*** after treatment of lower urinary tract symptoms; and (5) ***sexual functioning.***

a) Hormonal effects

Our knowledge of hormonal influences on lower urinary tract function and symptoms is inadequate. Although estrogen has been advocated as a treatment for lower

urinary tract symptoms, conclusive evidence of its benefit is lacking. A recent prospective study suggests that estrogen may actually be a risk factor for incontinence [71]. Evaluation of women in research studies should include assessment of menstrual and hormonal status. Information collected by history or questionnaire should include menopausal status (premenopausal; postmenopausal without hormone therapy; postmenopausal with hormone therapy); and hormone therapy if used (type, dose, and route of administration for estrogen and progestin).

b) Obstetric History

The unique influence of vaginal childbirth on the structure and function of the female pelvis remains incompletely understood. The need for basic clinical data on the study population and the specific aims of each study will determine the level of detail obtained for obstetric history. At a minimum in all studies, the number of vaginal deliveries should be ascertained. Other variables of potential interest include infant birthweight, length of second stage of labor, operative versus spontaneous vaginal delivery, use of midline or mediolateral episiotomy, and obstetric analgesia.

c) Pelvic Organ Prolapse

Studies of surgical treatment of incontinence (and other study types as appropriate) should include assessment for pelvic organ prolapse using the staging system approved by the ICS, the Pelvic Organ Prolapse Quantification (POP-Q) system [9] as described in the report of committee 8c. The POP-Q system includes measurement in centimeters of six vaginal sites relative to the hymen, plus three other measurements. The hymen marks the zero point of reference; positive numbers refer to prolapse beyond or distal to the hymen, and negative numbers refer to locations above or proximal to the hymen. Ordinal stages are defined by the most advanced site of prolapse affecting any of the six vaginal sites, as follows:

Stage 0: no prolapse.

Stage I: One or more of the vaginal sites or cervix is located at – 2 cm (2 cm above the hymen).

Stage II: One or more of the vaginal sites or cervix is located at – 1 cm, 0 cm, or + 1 cm (1 cm above or below the hymen, or at the hymen).

Stage III: One or more of the vaginal sites or cervix is located more than 1 cm beyond the hymen, but not to the maximal extent of protrusion.

Stage IV: Maximal extent of protrusion at one or more vaginal site or cervix.

Other measurements have not been standardized, such as assessment of urethral mobility (e.g., estimation on physical exam, cotton swab testing, perineal ultrasound, lateral cystogram), identification of paravaginal defects and perineal descent, pelvic muscle assessment, and pelvic imaging (e.g., defecating proctography, static or dynamic pelvic MRI). Detailed descriptions of their measurement should be included if they are used. Data should be presented as a continuum, not as a dichotomy of "normal" versus "abnormal" until those terms are clearly defined by evidence of clinical relevance.

Following recommendations made by the NIH Terminology Workshop for Researchers in Female Pelvic Floor Disorders, in general, prolapse is defined as descent of Stage I or greater at any site. An optimal anatomic outcome (cure) after intervention is defined as Stage 0, or no prolapse. A satisfactory anatomic outcome (improvement) after intervention is defined as Stage I. An unsatisfactory anatomic outcome (persistence or recurrence, failed treatment) after intervention is defined as Stage II or greater, or no change or worsening from pre-treatment stage.

d) Definitions of Outcomes for Lower Urinary Tract Symptoms in Women

The recommendations for outcomes from the NIH Terminology Workshop for Researchers in Female Pelvic Floor Disorders are detailed below. The recommendations emphasize that outcome after treatment for urinary incontinence should be defined in terms of stress incontinence symptoms, signs, and testing, but also in terms of associated symptoms and unwanted (side) effects resulting from an intervention, after return to baseline activities and medications. The suggested outcome definitions are detailed below. If these definitions are not adopted it is still imperative that researchers specify the outcome measures that will be used to define cure, failure, and improvement for each individual study protocol.

1. Stress Urinary Incontinence:

***Cure* of stress urinary incontinence is defined as:**

1. resolution of stress urinary incontinence symptoms;
2. resolution of the sign (negative full bladder cough stress test, performed under the same conditions as pre-treatment). In studies using urodynamics after intervention, absence of genuine stress incontinence should be documented.
3. no new symptoms or side effects. New symptoms or side effects should be specifically described and could include:

- new urinary symptoms such as urinary urgency, frequency, urge incontinence, with or without urodynamic changes of detrusor instability;

- change in sexual function;
- development or worsening of pelvic organ prolapse;
- adverse effect on bowel function;
- onset of urinary tract infections;
- surgical complications, such as foreign body reaction to grafts, or development of fistula or diverticula;
- osteitis or osteomyelitis;
- neuropathy; and
- others.

• Failure of treatment of stress urinary incontinence is defined as any one of:

1 persistent stress symptoms with the number of incontinent episodes unchanged, or worse, by voiding diary;

2 positive full bladder cough stress test (performed under the same conditions as pre-treatment) or genuine stress incontinence confirmed by urodynamic studies; and

3 presence or absence of new symptoms or side effects, as listed above.

• Improvement of stress incontinence includes:

1 persistent stress symptoms but with the number of incontinent episodes decreased by voiding diary;

2 positive full bladder cough stress test (performed under the same conditions as pre-treatment) or genuine stress incontinence confirmed by urodynamic studies; and

3 presence or absence of new symptoms or side effects, as listed above.

Since improvement has no standard definition, if improvement is used as an outcome, it must be specifically defined. In addition, when more than one characteristic is used to define an outcome (i.e., symptoms and sign), the characteristics will not be concordant in some situations. Possible categories to describe these situations include: (1) patient-observed treatment effect, with absence of stress symptoms and no side effects, but positive full bladder cough stress test; and (2) provider-observed treatment effect, with persistence of stress symptoms, no side effects, and negative full bladder cough stress test.

2. Urgency, Urge Incontinence, and Other Urinary Symptoms:

For outcomes related to urgency, urge incontinence and other urinary symptoms, *cure* is defined as the patient's statement (by history or questionnaire) that the symptom(s) is no longer present. *Failed* treatment is defined as the patient's statement that the symptom(s) is no better or worse, with objective data from a urinary diary. *Improvement* could include the patient's statement that the symptom(s) is less frequent or less bothersome, with evidence from a urinary diary.

Outcomes for detrusor overactivity should be defined separately for symptoms, as described above, and for urodynamic findings. Cure of detrusor overactivity is defined as the absence of involuntary phasic detrusor contractions on filling cystometry. Failure is defined as unimproved or worsened detrusor overactivity on urodynamics. Improvement has not been standardized and should be precisely defined for each study.

Although these recommendations advance the concept of global pelvic floor evaluation and emphasize the interrelatedness of pelvic organ function, there are limitations in compressing such broad outcome measures into only three categories. It is still critical to know whether a treatment corrects the intended problem. For example, if an operation reliably cures stress incontinence but causes dyspareunia, it may be more useful to report that there is a high cure rate plus a high complication rate. While appropriately emphasizing the significance of complications and adverse events, this system does not provide a means to fully express such complex outcomes. It also leaves a rather broad range of "improved" patients that must be further defined; when complete cures are relatively uncommon, this may diminish the impact of the outcome.

Sexual Function. Assessment of sexual function is an important part of measuring the impact of lower urinary tract symptoms on quality of life in women. Assessing change in sexual function should be a routine part of all studies of treatment for urinary symptoms. Some condition-specific quality of life instruments include sexual function, such as the Incontinence Impact Questionnaire[72]. A validated condition-specific instrument for assessing sexual function in women with urinary incontinence or pelvic organ prolapse has been recently published [73], and this or another valid instrument should be used in all surgical studies and whenever a higher level of detail regarding sexual function is appropriate.

Recommendations:

- **Data on hormonal status should be collected on women in all studies of incontinence and LUTS.**
- **At a minimum, data on vaginal parity should be collected on women in all studies. Additional obstetric history should be obtained as appropriate for individual studies.**
- **Studies of surgical treatment of incontinence (and other study types as appropriate) should include assessment for pelvic organ prolapse using the ICS staging system, the Pelvic Organ Prolapse Quantification (POP-Q) system.**

- **Outcomes (cure, failure, improvement) must be clearly symptoms and signs defined at the outset of all studies, based on changes in Complications and side effects may be included in the definition of outcomes but should also be reported separately.**
- **Assessment of sexual function should be included in all studies.**

3. Frail older and disabled people

We agree with recommendations for outcome research in frail older people as reported in the ICS Subcommittee on "Outcome Measures for Research of Lower Urinary Tract Dysfunction in Frail Older People" [15]. In addition, please refer to the full report of Committee 10c regarding conservative treatment in the elderly.

Frailty is defined as "a state of reduced physiological reserve associated with increased susceptibility to disability [74]" There remains a wide variation in functional capacity within this definition ranging from those requiring some assistance with activities of daily living to those suffering from dementia and severe physical handicaps. For this population there is little validated research showing long-term efficacy of treatment for urinary incontinence. Research in this population is difficult because of:

- heterogeneity of the population resulting in difficulty designing studies that account for comorbidity, drug use, intercurrent illness, and shorter life expectancy;
- lack of standardized terminology to define and measure cure and improvement;
- lack of validated research tools to measure baseline and outcome variables;
- lack of long-term follow up to gauge impact, durability, outcomes, and applicability of interventions;
- lack of information on the natural history of incontinence.

a) Considerations in study design

1. Baseline clinical data: Descriptive data regarding the patients' current care setting should be fully described (e.g., type of setting of the study such as home or nursing home; patient-staff ratio; usual continence care; direct and indirect costs of current care; patient, family and/or staff expectations; description of caregivers and their training; and system incentives or disincentives that may influence management options). Associated factors influencing incontinence or the potential response to treatment must also be accounted for (e.g., environmental factors contributing to incontinence such as toilet access, and associated comorbid conditions that influence incontinence or the effectiveness of intervention). Bowel status and concurrent medications are important in this population. Mobility is often impaired in these patients; impaired mobility impacts urinary control, therefore mobility should be assessed using validated instruments. Finally, the functional level and cognitive state of the patients should be characterized using standardized scales (Bartel Orcats ADL scales [75] [76] and Mini-mental status Scale Examination [77], respectively).

There are age specific influences on lower urinary tract function but normative data are generally lacking in this frail population. In addition, the test-retest reliability and sensitivity to change of the more invasive measures of lower urinary tract function are poorly documented in the frail elderly. It is probably not appropriate to repeat invasive measures at follow-up in this frail population unless these measures are fundamental to the outcome of the intervention being studied.

The following information should be addressed and reported at the time of follow-up whenever possible:

- number and reason for dropouts and deaths (i.e. were they trial related)
- compliance issues (by patients, staff or caregivers), such as compliance to exercise programs, toileting protocols, or drug use)
- type of bladder training or toileting programs (if any)
- other intercurrent treatment including medication not directly related to bladder function that may influence outcome
- socioeconomic data including impact of the intervention on the patient
- changes in caregiver or staff status or numbers
- cost of the treatment
- cost-benefit data
- patient and/or caregiver satisfaction with the intervention
- risk benefit data

Because comorbidity and drug use contribute to the presence and severity of incontinence in this population, they should be stabilized before enrollment.

b) History and symptoms

Research in this group cannot be based solely on patients' subjective reporting of symptoms. In some cases, the patient's perspective of the problem may be less relevant than that of family members and caregivers. Patient-derived symptom response as an outcome measure should be supplemented by objective data from diaries, etc., and data derived from caregivers.

c) Outcome and other measures

It must be acknowledged that almost all measures used in the study of incontinence in the community dwelling population require separate validation for use with the frail elderly. In addition, establishing clear "clinically significant" outcomes and understanding the full socioeconomic costs of therapy are of particular importance in this population as the patients are often unable to participate in decision making.

d) Conclusion

Research methodology for studying incontinence in the frail and housebound elderly is fraught with pitfalls. This has compromised the usefulness of past research. There is a great need for basic research to validate practical and useful outcome measures that will allow meaningful results to be obtained. In addition, an understanding is required of the importance of defining clinical rather than statistical significance.

RECOMMENDATIONS:

- **This is a heterogeneous population requiring a detailed study design and careful description of baseline clinical data if results are to be interpretable**
- **There is a need for validation of all instruments and procedures used in incontinence research for the population of frail elderly patients**
- **"Clinically significant" outcome measures and relationships of outcome to socioeconomic costs are critically important to establishing the utility of treating urinary incontinence in this population.**

4. INCONTINENCE IN CHILDREN

In general, conducting clinical research in children is more difficult than in adults for a variety of reasons. However, the need for quality clinical research in children has been emphasized in an official report from the United States National Institutes of Health (NIH) from March 1998, published in response to statements from the 1996 U.S. Congress Appropriations committees calling for increased and improved funding of pediatric medical research. The document [78] sets forth the policy and guidelines on the inclusion of children in research involving human subjects that is supported or conducted by the NIH. The goal of this policy is to increase the participation of children in research so that adequate data will be developed to support the treatment modalities for disorders and conditions that affect adults and may also affect children. The document points out that, "The policy was developed because medical treatments applied to children are often based upon testing done only in adults, and scientifically evaluated treatments are less available to children due to barriers to their inclusion in research studies". The American Academy of Pediatrics has reported that only a small fraction of all drugs and biological products marketed in the U.S. have had clinical trials performed in pediatric patients and a majority of marketed drugs are not labeled for use in pediatric patients. Many drugs used in the treatment of both common childhood illnesses and more serious conditions carry little information in the labels about use in pediatric patients. It is the stated policy of NIH that children (i.e., individuals under the age of 21) must be included in all human subjects research, conducted or supported by the NIH, unless there are scientific and ethical reasons not to include them. Appropriate exceptions are listed in the document. The specific responsibilities of all involved parties—principle investigators, institutional review boards, involved institutions, peer review groups, and the NIH—are detailed. Finally, and perhaps most importantly, the document describes levels of risk and the corresponding nature of assent required for participation in research studies. All clinical investigators that work with children should be familiar with the contents of this NIH document.

Four overriding issues separate pediatric research from the general recommendations. First, physiology varies widely within the group referred to as "children", differs from adults, and changes with time. Because children are growing, any treatment, especially pharmacological and surgical therapy, may affect them profoundly in the long term. This is particularly true of the immature brain, nervous system and other incompletely developed systems. Second, compliance with therapy is more complicated as children may depend on caregivers to administer treatment in many studies. Third, reporting of symptoms and outcomes may be difficult. The child may be unable or unwilling to respond. Symptoms reported by a caregiver may not be interpreted in the same way as the child. Finally, the issue of informed consent becomes even more complex with children.

The pediatric population is not a homogenous group; neonates, infants, pre-pubescent children, and adolescents clearly differ physiologically and psychologically. The effect of illness and the treatment of that illness must be carefully studied in each age group. Studies should be robust enough to allow for evaluation of varying age groups when relevant. Urinary incontinence in children falls into four main categories: neuropathic (myelomeningocele and other less common neurogenic etiologies), pure nocturnal enuresis, detrusor overactivity, and dysfunctional voiding without neurologic disease. This issue of age groups is most crucial in children with myelomeningocele. These children may be

on drug therapy from a very young age onward; the long-term safety of medications in children must be established in all age groups. Therapy for other causes of incontinence in children tends to start at a later age, by which time size is the main difference between children. We recommend that clinical studies have long-term (five years or more), open label extension arms to monitor safety, particularly focusing on normal growth and development and the effects on treatment of liver and central nervous system function. Most importantly for incontinence studies, normal maturation may significantly enhance or obscure response to an intervention.

Assessment of compliance with therapy is always difficult, and even more so with children. Compliance with voiding diaries, a significant issue in the adult population, may be even more problematic with children. Children may "act out" and refuse medications or other treatments. Children may be willing to comply with instructions from one parent or caregiver but not another. Personal problems of the caregiver may dramatically affect the child's compliance with a treatment protocol. We can only recommend that this potential problem be recognized and given even more attention than in trials with adults. Adequate support to the family member consenting to the trial may aid in compliance with treatment. Specific compliance issues should be identified whenever possible. If a treatment is not accepted by either the adult or the child (e.g., tablet size too large, taste of the medicine not acceptable, behavioral treatment schemes too rigid), then it cannot be effective in practice, no matter how theoretically beneficial it may be.

The NIH document details appropriate levels of consent required based on the risks inherent to a particular study. Depending on the age of the child, consent may be given by the parent in a purely surrogate role or the child may participate to some degree in the process. However, true informed consent of the subject is not possible in the vast majority of cases when children are involved. We recommend that an effort be made to include the child in the discussion of the trial with age specific language and illustrations when appropriate. It is important to include the primary care giver, when the consenting adult will not be administering the treatment. Such complex relationships exist where childcare is shared amongst more than one adult, or where an employee for the purposes of childcare exists, either inside or outside the home. As always, we recommend that that study designs ensure that children are always offered the standard of care when such exists. In fact, because so few treatments have ever been studied properly in children, there are many areas in which no treatment can properly be called "safe and effective".

Outcome measures are not as well developed in children as in adults. Validated, age-specific symptom and disease-specific quality of life instruments must be developed for the pediatric population. Early efforts in this area have been reported for dysfunctional voiding [79] and daytime incontinence [80] much more work remains to be done. Invasive urodynamics can rarely be used (except in the neurogenic population), as parents will not allow repeated instrumentation of the child. The reproducibility of urodynamic investigations in children is still under investigation.

RECOMMENDATIONS:

- **We support the NIH statement calling for increased clinical research in children. All investigators that work with children should be aware of the details of the document and particularly the issues surrounding informed consent.**
- **Long-term follow-up is of critical importance in the pediatric population in order to ascertain the effect of a treatment on normal growth and development**
- **Research is needed to develop standardized outcome measures including validated, age-specific symptom and disease-specific quality of life outcome measures.**

5. NEUROPATHIC LOWER URINARY TRACT DYSFUNCTION

Modern neurourologic care is generally successful in preventing late complications in neurogenic patients, maintaining renal function, and promoting independence in self-care. Lifelong urological follow-up is mandatory and there are many areas for further research to improve the lives of these patients. These recommendations add to those described before and focus on the specific characteristics of the neurogenic patient. Specific discussion of treatment in the neurogenic population is contained in reports from committees 10d and 11c. Reports from committees 2, 4, 7, and 15 are also relevant to this population. Statistical methods and research outcome are identical as described in the general recommendations. Emphasis is given to:

- classification of the neurogenic patient
- the specifics of history and evaluation, necessary for research studies
- the urodynamic evaluation, which is the key investigational tool in the evaluation of this specific, complex and difficult patient population

a) Classification

Classification of neurogenic voiding dysfunction has

three primary aims—to aid in discriminating or identifying an unknown underlying neurological disease process, to characterize the nature of the dysfunction so as to develop a treatment plan, and to assess the risk of secondary effects (e.g. on the upper tract) which may influence the necessity and aggressiveness of treatment. The latter two are clearly relevant to research in neurogenic incontinence and must be reflected in study design and patient description.

Despite this, it is difficult to find a good classification system of neurogenic voiding dysfunction as a base for research. The published systems are reviewed in detail by Wein [81]. Both the disease process and the site of the neurologic lesion(s) are relevant in the study of neurogenic voiding dysfunction, yet even this information is inadequate to predict the functional characteristics for an individual patient. There is no one that meets the broad needs of classification in this group. Typical or classic cases are often well described but it is especially difficult to handle patients with mixed and incomplete lesions. Thus, the classification systems necessarily oversimplify or become extremely cumbersome. Finally, it must be acknowledged that the complexity of neurologic diseases and variations in individual behavior almost always call for a customized approach to therapy, further complicating research in the neurogenic patient. All of these factors complicate study design as it becomes difficult to create workable inclusion and exclusion criteria.

b) History and evaluation

Study planning is best undertaken with the cooperation of urologist, neurologist, and other clinicians, who have specific interest and special training in the neuropathic patient. Baseline data collected by history in subjects with neuropathic lower urinary tract disorders should include:

- bladder volumes by diary or examination (functional, total capacity, post voiding residual urine);
- mechanism of bladder evacuation: normal or volitional, reflex evacuation, spontaneous involuntarily, Credé, sterile intermittent catheter (SIC), clean intermittent catheter (CIC), intermittent catheter by second person, suprapubic or urethral catheter;
- use of external appliances (e.g., diaper or pad use, condom catheter, urethral catheter, suprapubic tube);
- the typical time span of continence following last bladder evacuation.

Objective assessment of sacral nerve function should be determined. This includes:

- Perineal sensation (S 3-5)
- Bulbocavernous reflex (S 2-3)
- Bulboanal reflex (S 3-4)
- Cutaneous anal reflex (S 4-5)

Issues such a skin breakdown and fecal impaction frequently become relevant in this population compared to neurologically intact individuals.

c) Urodynamics

In contrast to the general recommendations, baseline urodynamics are required for research studies of the neurogenic incontinence. Because the nature of the lower urinary tract dysfunction cannot be accurately predicted based on the history and physical findings, urodynamic classification is mandatory. Neurogenic disorders commonly cause complex and generalized lower tract dysfunction, combining bladder and urethral sphincter abnormalities. In addition, data should be collected on symptoms and the underlying neurologic disease. While urodynamic classification alone is suboptimal, it is clearly preferable to classification by symptoms or disease alone (e.g., a study involving patients with hyperreflexic neurogenic bladder and coordinated sphincters will be more generalizable than one of urge incontinence in neurogenic patients or all multiple sclerosis patients).

Urodynamic studies in neurogenic disorders are qualitatively different compared to non-neurogenic disorders. For each subject, bladder function, sphincter function, and the coordination between the two must be fully described. In addition to data on stable or unstable filling, compliance is also of major importance. Elevated detrusor leak point pressure predicts upper urinary tract deterioration in children with myelomeningocele [82] and is important in all patients with non-compliant filling. Detailed analysis of voiding dynamics becomes more important (e.g., simultaneous Pves/Pabd during voiding, voiding time, shape of the Pd and Q curves) because of the possibilities of functional obstruction and impaired contractility, which are uncommon outside of the neuropathic population. Because the bladder and sphincter may be dyssynergic, assessment of sphincteric activity is essential. This may be accomplished by surface EMG of the pelvic floor, needle electrodes, fluoroscopy, ultrasound, or direct measurement of urethral pressure.

RECOMMENDATIONS:

- **Detailed urodynamic studies are required for classification of neurogenic lower urinary tract disorders in research studies because the nature of the lower tract dysfunction cannot be accurately predicted from clinical data. Videourodynamic studies are preferred but not mandatory.**
- **Change in detrusor leak point pressure should be reported as an outcome as appropriate, and can be considered a primary outcome in addition to symptom response.**

- **An area of high priority for research is the development of a classification system to define neurogenic disturbances. Relevant features would include the underlying diagnosis, the symptoms, and the nature of the urodynamic abnormality.**
- **It may sometimes be appropriate to group patients withurodynamically similar neurogenic bladder disorders of different etiologies in a clinical trial. However, great caution must be used if patients with progressive disease (e.g., multiple sclerosis) are grouped with patients having a stable deficit (e.g., traumatic spinal cord injury).**

IV. CONSIDERATIONS FOR SPECIFIC TYPES OF INCONTINENCE RESEARCH

1. Behavioral and physiotherapy trials

Non-pharmacologic, non-surgical treatments for incontinence comprise a wide variety of tools often grouped under the name of behavioral treatment. Because these treatments are generally very safe and applicable to most incontinent patients, there may be a tendency to use less stringent protocols. This must be discouraged.

The type of therapy must be defined with sufficient detail that other investigators can reproduce the study. The type of behavioral therapy should be clearly stated, including the duration of the total treatment period, duration of each treatment session, and number of treatment sessions. The time between qualification for study entry and start of therapy must be specified. Any devices used must be properly described. The background and training of the therapist should be defined. All instructions, training, and educational materials given the subjects should be reproduced or referenced. A complete description of all differences in the experience of the treatment and control groups should be provided.

As in other studies, the study population should be identifiable. When urodynamics are not used to describe the pathophysiology, other valid measures are employed. The usual clinical outcome measures suffice. In order to progress in our understanding of these treatments is important to correlate clinical outcome with physiologic changes. If the intervention is intended to increase the strength of pelvic floor muscle contraction, this should be measured and correlated with continence. Outcome measures in related organ systems (e.g., gastrointestinal and sexual functioning) should also be considered, as well as possible adverse outcomes.

It is important to distinguish between *specific* and *non-specific effects*, such as improvement related to the extra attention of the therapist, motivation, confidence gained, etc. The goal is to isolate what a particular therapy achieves on its own. However, in behavioral therapy, the non-specific effect is widely considered to be an essential, desirable and important part of the effect of the therapy. It therefore needs to be evaluated along with the specific effect. Carefully designed randomized controlled studies should allow separation of specific and non-specific effects. This is particularly important with techniques such as electrical stimulation and biofeedback where particular instrumentation or equipment may be credited with results that could be due to the efforts of the therapist.

It is often difficult to perform double-blind studies of behavioral technique. Clinicians and subjects often cannot be blinded. In quality assessment of studies, double blinding is often one of the criteria of methodological quality. It may not be reasonable to demand double-blinding in all behavioral studies, or, if double blinding is not accomplished, to consider such research less valuable. It is more realistic that we demand the 'most optimal and possible level of blinding'. This means that a relevant control group is established, that every effort is made to blind as many persons as possible, and that appropriate measures surrounding this issue are discussed in the manuscripts.

Recommendations:

- **Treatment protocols must be detailed to the degree that the work can easily be reproduced**
- **A structured examination of pelvic floor function should be included before and after treatments that are aimed at pelvic muscle training**
- **More work is needed to separate the specific and non-specific effects of treatment**

2. Device Trials

The United States Food and Drug Administration (FDA) had established detailed guidelines for studies on intra-urethral and vaginal devices and urethral bulking agents in the treatment of urinary incontinence [83].Although devices and bulking agents differ considerably in risks to research subjects, they are grouped together for the purpose of FDA regulation. Requirements for the protection of human subjects are appropriate for the study of bulking agents, but are probably excessive for research on devices. Any researcher considering this area of investigation should be familiar with this FDA document, which outlines the entire conduct of studies from design through outcome mea-

sures. For the most part, these guidelines follow the general recommendations. Some specific issues invite comment.

1. Inclusion is limited to patients with "urinary incontinence due to ISD (intrinsic sphincter deficiency), as evidenced from urodynamic studies or radiographic assessment". While the concept of ISD is well understood, there is no consensus on its definition for clinical care or research.
2. Female subjects "must demonstrate an abdominal leak point pressure less than 65 cm H2O". There is no evidence to support this particular cutoff, and the clinical significance of this value is questionable given the wide variation in techniques for leak point pressure measurement.
3. The potential study population is markedly limited by exclusion of mixed incontinence, failure of a previous injection procedure for stress incontinence, neurogenic bladder, previous implantation of an artificial urinary sphincter, and patients taking medications affecting the bladder. These patients could potentially benefit from therapy, but cannot be included in research by this guideline.
4. The initial evaluation calls for urodynamic testing and a pad test but not a voiding diary. We recommend that voiding diaries be included in all incontinence studies.
5. Along with routine data collection, all studies must include urodynamic testing, cystoscopy, and pulmonary and liver function results at 12 month visits. Although this is because of issues specific to bulking agents, the requirements include all devices.
6. The Stamey grading scale (0-3) for stress urinary incontinence is recommended as the primary outcome measure. There is little evidence that this measure is as valid or reliable as other measures such as voiding diaries, pad tests, and leak point pressure measurements. While the Stamey grading scale is required by the FDA, researchers should use a variety of outcome measures as described in the general recommendations and in the specific recommendations for women.

One other important area of concern in device studies is patient recruitment procedures. We strongly support reporting according to the CONSORT guideline, including the flow diagram (Figure 1) for subject enrollment and follow-up. Subjects should be enrolled in a manner that minimizes selection bias. The protocol should detail the procedure by which consecutive patients meeting the inclusion criteria are selected. All situations in which a patient meets the inclusion/exclusion criteria but is not offered enrollment by the investigator should be documented. The number of patients who decline enrollment should be stated, along with the reasons. There should be a complete accounting of all patients in the study including the reasons for subject withdrawal; recommended loss to follow-up should not exceed 20% over the course of the study per the FDA.

RECOMMENDATIONS:

- **Researchers should be familiar with the FDA guidelines for research in devices. However, vaginal support devices, urethral stents, and urethral bulking agents are not intrinsically similar and these guidelines should be refined such that recommendations are appropriate to the risk involved in the treatment.**
- **Full reporting of studies following the CONSORT flow-chart (even for observational studies) will help to define the degree of selection bias inherent in this type of research**

3. PHARMOCOTHERAPY TRIALS

Drug trials are necessary so that new drugs can be clinically and scientifically evaluated for quality, efficacy and safety [47] [84] [85] [86] [87] [88]. Since the 1960's administrative bodies such as the Food and Drug Administration have required that new pharmaceuticals undergo controlled investigations to establish efficacy. In order to comply with laws governing the release of new drugs to the general public, various phases of drug trials are undertaken. The specific stages and of study design have been discussed in detail in section IIB. Pharmacotherapy trials in incontinence have come closer to the ideals presented in the general recommendations than have other treatment modalities. Incontinence research has been greatly advanced in recent years with the introduction of new medications that have been carefully studied in several large RCTs. While the financial backing of the pharmaceutical industry has been largely responsible for this superior research, new conflicts and problems have arisen due to the changing economics of research. As stated in a joint editorial endorsed by members of the International Committee of Medical Journal Editors, ". . . published evidence of efficacy and safety rests on the assumption that clinical trials data have been gathered and are presented in an objective and dispassionate manner. . . We are concerned that the current intellectual environment in which some clinical research is conceived, study subjects are recruited, and the data analyzed and reported (or nor reported) may threaten this precious objectivity" [89]. Several of these issues are discussed below.

a) Payment for drug studies

Especially in the US, proceeds from clinical trials have become an increasingly important supplement to clinician income. Clinical research, previously limited to a few academic institutions, is now spread through all segments of the medical community. While this may improve the variety of patient representation in studies,

it also makes safeguarding the rights of research subjects more difficult. Competition for revenue from research, aggressive advertising for research subjects, and dependence of clinicians on income from pharmaceutical companies are trends that bear close attention. Most quality peer-review scientific journals require a declaration of conflict of interest. It is preferable that researchers do not receive money directly from industry sponsors. An acceptable alternative is to have research funds paid into an appropriate research account and dispensed by an independent third party.

b) Clinical direction

In clinician-initiated, government-funded research, there has always been a lead investigator who is ultimately responsible for all aspects of the work. This paradigm may not be applicable to pharmaceutical research. The structure of the trial is determined by the company (perhaps with input from a group of consultants); there are typically a large number of sites, each of which enrolls relatively few subjects; and data analysis is performed centrally, often under the direction of the sponsoring company. Clinicians at each site cannot be intimately familiar with the entire process of the study. When results are reported, the paper may be written by an outside agency, and then passed to authors for editing and comments. This presents a real problem with favoritism and inevitably dilutes the force, impact, and responsibility of authorship. Standards of authorship defined by many journals should be followed and rely on the honor system for compliance. Academic leaders should work to establish standards for interactions between clinicians and industry.

A final issue of special relevance in trials of pharmaceutical agents (although germane to other treatment modalities) is the controversy regarding placebos in clinical trials. Regardless of whether a drug is effective or not, simply giving a drug to a patient may produce a beneficial response. To assess if a drug has an effect over and above the placebo response, it is usually tested against an inactive substance (placebo). In incontinence, the placebo effect may be quite large, anywhere from 30-50% in recent published studies. To account for this, investigators and regulators have generally demanded a placebo arm in most clinical trials of medication. On the other hand, the Helsinki Agreement (1989) states that "far from being useful, a placebo is unethical: in any medical study every patient including those in the control group, if any, should be assured of the best proven diagnostic and therapeutic method". Clinicians need to know how a new drug compares with established treatment. The FDA does not require placebo-controlled trials of drugs for approval. However, the sponsor will generally prefer to compare the drug with a placebo and not with a competitor, since it is usually easier to detect a difference between treatment and no treatment, compared to two active treatments. Researchers must carefully consider these issues in designing a relevant, ethical study. The report of committee 9 also addresses issues related to placebos.

RECOMMENDATIONS:

- **While considerable progress has been made in pharmaceutical research on urinary incontinence, few reports have followed the CONSORT document; this decreases the clarity and impact of the work. All randomized clinical trials should follow these reporting guidelines.**
- **Continuity in clinical direction from design through authorship is highly desirable. All authors should be able to accept responsibility for the published work and all potential conflicts of interest should be fully disclosed.**
- **Investigators must be sensitive to the conflicts regarding the use of placebos in clinical trials. While placebos are often desirable from a scientific standpoint, every consideration should be given to making sure that the interests of the subjects are kept at the forefront in designing safe, ethical research.**

4. SURGICAL STUDIES

Standards for surgical trials are detailed in recommendations from the ICS, the Society for Urodynamics and Female Urology (SUFU), and the American Urological Association (AUA) [12,13,14,15,16,17]. We support the adoption of these standards by clinical and basic science researchers, the peer review process, specialty and sub-specialty organizations, the health care industry, regulatory agencies and ultimately by clinicians. While discussion of surgical therapy for incontinence mainly applies to females with stress incontinence, most of these points are equally applicable to males undergoing surgery for post-prostatectomy incontinence and related problems.

Entry: The choice of surgical treatment involves significant selection bias on the part of both the patient and surgeon. If the study is not a randomized controlled trial, this bias should be acknowledged and the number of patients treated by other methods at the same institution during the same period should be reported. Of particular importance are those patients who would be eligible for surgery but who were not offered surgery or did not select an operation. Patients undergoing a different operation for incontinence other than the one under examination should be reported.

Baseline evaluation: All patients should undergo a comprehensive baseline evaluation as discussed in the general recommendations. Void diary and pad testing is of critical importance. In addition, we believe that uro-

dynamic studies are valuable in patient selection and should be performed on all patients undergoing surgery. The issue is discussed in more detail in the report of committee 7.

Conduct of the study: The exact surgical procedure should be described in such detail that it could easily be reproduced in another study. Discussion should include measures taken to assure that all subjects were treated in the same fashion, and that surgical technique did not change or evolve during the study. It is important to avoid doing studies "on the learning curve". Any new technique, especially surgical ones, first should be allowed to find its intended form before it is compared to established techniques ("gold standards") or subjected to other comparisons.

Analysis: A concerted effort should be made to follow-up every patient. Follow-up should be considered to be to the date of the last exam or complete data collection. Accounting for patients "lost to follow-up" must be detailed.

Reporting: Reports of successful treatment should be limited to those subjects with a minimum of one year follow-up. However, unsuccessful treatments should be reported as rapidly as possible, to avoid exposing many more patients to inadequate treatment.

RECOMMENDATIONS:

- **There is a great need for randomized clinical trials in surgical treatment of urinary incontinence. Reports of observational studies should follow the CONSORT flow-chart, which will help to define the degree of selection bias inherent in this type of research**
- **We recommend urodynamic testing in all subjects involved in surgical research. However, evidence does not exist to support recommendations for minimal testing or the use of specific tests. Research into the predictive value of preoperative urodynamic studies would be most valuable.**
- **Reports of successful treatment should be limited to subjects with a minimum of one year follow-up.**

V. CONCLUSION

One of the key themes of this Second International Consultation on Incontinence has been examining and classifying data by levels of evidence. The goal of this section has been to aid researchers in their efforts to produce research of high quality. High quality research will win "high grades of evidence", lead to new recommendations in future Consultations, and drive our efforts to understand the etiology of incontinence, treat it effectively, and, prevent its occurrence. Ultimately, good research is credible. Credibility creates impact. Credible research draws others to follow and expand on the work while simultaneously guiding clinical care of patients. Unfortunately, much of the published work in the field has not been credible and has not effectively changed patient care. In most cases, this has been due to preventable deficiencies in planning and data collection.

We cannot be discouraged by the fact that much of the clinical research that has been carried out in lower urinary tract disorders has been of low quality. Instead, we should identify and promote what is good and tenable, and build on that knowledge. Continued multidisciplinary cooperation anchored in preclinical activities is an absolute precondition for successful clinical research in lower urinary tract dysfunction in the future.

In summary,

- All quality research, be it prospective or retrospective, clinical or preclinical, begins with detailed planning—establishing a clear and relevant hypothesis, developing a trial of appropriate magnitude to accept or reject the hypothesis, and defining methods of adequate sensitivity and specificity to produce credible data.
- Clinical research in incontinence must include a broad range of baseline and outcome measures including anatomic and physiologic variables, urodynamic testing, voiding diaries and pad tests, symptom assessment, and quality of life measures. Economic outcome assessment should be included whenever possible. In each area, data must be collected using structured, reproducible methodology. Symptom assessment and other instruments must be validated for the population being studied.
- The CONSORT statement should be adopted as criteria for publication of randomized clinical trials by researchers, reviewers, and editors.
- Baseline urodynamic assessment is required in the neurogenic population and recommended in surgical trials. However, baseline urodynamic studies are highly desirable in all types of incontinence research. There is a great need to critically examine the predictive value of urodynamic testing in order to refine our tools as well as the diagnosis and treatment of patients.
- The primary goal of clinical research is to improve the care of patients; the ultimate goal is to understand the nature of disease and how treatments actually work. We can make this progress by collecting comprehensive baseline and follow-up data, and correlating outcome to baseline characteristics and observed changes during treatment.

REFERENCES

1. Abrams P, Blaivas JG, Stanton SL and Andersen JT: The standardization of terminology of lower urinary tract function. Scand J Urol Nephrol suppl 114, 5-19, 1988.
2. Bates P, Bradley WE, Glen E, Melchior H, Rowan D, Sterling A, Hald T. First report on the standardization of terminology of lower urinary tract function. Urinary incontinence. Procedures related to the evaluation of urine storage: Cystometry, urethral closure pressure profile, units of measurements. Br J Urol 48:39-42, 1976, Eur Urol 2:274-276, 1976, Scand J Urol Nephrol 11:193-196, 1976, Urol Int 32:81-87, 1976.
3. Bates P, Glen E, Griffiths D, Melchior H, Rowan D, Sterling A, Zinner NR, Hald T. Second report on the standardization of terminology of lower urinary tract function. Procedures related to the evaluation of micturition: Flow rate, pressure measurement, symbols. Acta Urol Jpn 27:1563-1566, 1977, Br J Urol 49:207-210, 1977, Eur Urol 3:168-170, 1977, Scand J Urol Nephrol 11:197-199, 1977.
4. Bates P, Bradley WE, Glen E, Griffiths D, Melchior H, Rowan D, Sterling A, Hald T. Third report on the standardization of terminology of lower urinary tract function. Procedures related to the evaluation of micturition: Pressure flow relationships, residual urine. Br J Urol 52:348-359, 1980, Euro Urol 6:170-171, 1980, Acta Urol Jpn 27:1566-1568, 1980, Scand J Urol Nephrol 12:191-193, 1981.
5. Bates, P, Bradley WE, Glen E, Melchior H, Rowan D, Sterling A, Sundin T, Thomas D, Torrens M, Turner-Warwick R, Zinner NR, Hald T. Fourth report on the standardization of terminology of lower urinary tract function. Terminology related to neuromuscular dysfunction of lower urinary tract. Br J Urol 52:333-335, 1981, Urology 17:618-620, 1981, Scand J Urol Nephrol 15:169-171, 1981, Acta Urol Jpn 27:1568-1571, 1981.
6. Abrams P, Blaivas JG, Stanton SL, Andersen TJ, Fowler CJ, Gerstenberg T, Murray K. Sixth report on the standardization of terminology of lower urinary tract function. Procedures related to neurophysiological investigations: Electromyography, nerve conduction studies, reflex latencies, evoked potentials and sensory testing. World J Urol 4:2-5, 1986, Scand J Urol Nephrol 20:161-164, 1986.
7. Rowan D (Ch.), James ED, Kramer AEJL, Sterling AM and Suhel PF (ICS working party on urodynamic equipment). Urodynamic Equipment: technical aspects. J Med Eng Technol 11, 2, 57-64, 1987.
8. Andersen JT, Blaivas JG, Cardozo L and Thüroff, J. Lower Urinary Tract Rehabilitation Techniques : Seventh Report on the Standardization of Terminology of Lower Urinary Tract Function. Int Urogynecol J, 3:75-80, 1992.
9. Bump RC, Mattiasson A, Bø K, Brukaber LP, DeLancey JOL, Klarskov P, Shull BL and Smith ARB: The Standardization of Terminology of Female Pelvic Organ Prolapse and Pelvic Floor Dysfunction. Am J Obstet Gynecol, 175:10-17, 1996.
10. Thüroff J, Mattiasson A, Andersen JT, Hedlund H, Hinman F Jr, Hohenfellner M, Månsson W, Mundy AB, Rowland RG and Steven K. Standardization of Terminology and Assessment of Functional Characteristics of Intestinal Urinary Reservoirs. Neuroul Urodyn, 15:499-511, 1996, Br J Urol, 78:516-523, 1996, Scand J Urol Nephrol, 30:349-356,1996.
11. Griffiths D, Höfner K, van Mastrigt R, Rollema HJ, Spångberg A and Gleason D. Standardization of Terminology of Lower Urinary Tract Function:Pressure-Flow Studies of Voiding, Urethral Resistance, and Urethral Obstruction. Neurourol Urodyn 16:1-18, 1997.
12. Mattiasson A, Djurhuus JC, Fonda D, Lose G, Nordling J and Stöhrer M: Standardization of Outcome studies in patients with Lower Urinary Tract Dysfunction. A report on general principles from the Standardization Committee of the International Continence Society. Neurourol Urodyn 17:249-253, 1998.
13. Lose G, Fantl JA, Victor A, Walter S, Wells TJ, Wyman J and Mattiasson A: Outcome measures in adult women with symptoms of lower urinary tract dysfunction. Neurourol Urodyn 17:255-262, 1998.
14. Nordling J, Abrams P, Ameda JT, Donovan J, Griffiths D, Kobayashi S, Koyanagi T, Schäfer W, Yalla S and Mattiasson A: Outcome measures for research in treatment of adult males with symptoms of lower urinary tract dysfunction. Neurourol Urodyn 17:263-271, 1998.
15. Fonda D, Resnick NM, Colling J, Burgio K, Ouslander JG, Norton C, Ekelund P, Versi E and Mattiasson A: Outcome measures for research of lower urinary tract dysfunction in frail older people. Neurourol Urodyn 17:273-281, 1998.
16. Blaivas JG, Appell RA, Fantl JA, Leach G, McGuire EJ, Resnick NM, Raz S and Wein AJ: Standards of Efficacy of Treatment Outcomes in Urinary Incontinence: Recommendations of the Urodynamic Society. Neurourol Urodyn 16:145-147, 1997.
17. Blaivas JG, Appel RA, Fantl JA, Leach G, McGuire EJ, Resnick NM, Raz S and Wein AJ: Definition and Classification of Urinary Incontinence: Recommendations of the Urodynamic Society. Neurourol Urodyn 16:149-151, 1997.
18. Blaivas JG: Outcome measures for urinary incontinence. Urology Feb; 51 (2A Suppl):11-19, 1998.
19. Spilker B: Guide to clinical studies and developing protocols. Raven press New York, 1984.
20. Senn S. Statistical Issues in Drug Development. Chichester: Wiley, 1997.
21. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991.
22. Senn S. Cross-over Trials in Clinical Research. Chichester: Wiley;1993.
23. Armitage P, Berry G. Statistical Methods in Medical Research, 3rd edition. Oxford: Blackwell Science, 1994.
24. International Conference on Harmonisation web site. Statistical considerations in the design of clinical trials. Available at: http://www.ifpma.org/pdfifpma/e9.pdf. Accessibility verified May 29, 2001.
25. CONSORT web site. Statistical considerations in the design of clinical trials. Available at: http://www.consort-statement.org. Accessibility verified May 29, 2001.
26. Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. JAMA 1994; 242: 1926-1931.
27. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996;313:570-571.
28. Altman DG, Schulz KF, Moher D, Egger M, Avidoff F, Elbourne D, Gøtzsche PC, Lang T, for the CONSORT Group. The revised CONSORT statement for reporting randomised trials: explanation and elaboration. Annals of Internal Medicine 2001; 134: 663-694.
29. Egger M, Jüni P, Bartlett C, for the CONSORT Group. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996-1999.
30. Moher D, Jones A, Lepage L, for the CONSORT Group. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001; 285:1992-1995.
31. Rennie D. CONSORT revised – improving the reporting of randomized trials. [Editorial] JAMA 2001;285:2006-2007.
32. Pocock SJ. Clinical Trials: a Practical Approach. Chichester: Wiley; 1983.

33. Black N. Why we need observational studies to evaluate the effectiveness of health care. Br Med J 312:1215-1218, 1996.
34. Peters TJ, Eachus JI. Achieving equal probability of selection under various random sampling strategies. Paediatric and Perinatal Epidemiology 1995;9:219-224.
35. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold; 2000.
36. Moher D, Schulz KF, Altman D, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285:1987-1991.
37. Matthews JNS. An Introduction to Randomized Controlled Clinical Trials. London: Arnold; 2000.
38. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. Br Med J 311:1145-1148, 1995.
39. Machin D, Campbell M, Fayers P, Pinol A. Sample Size Tables for Clinical Studies, 2nd edition. Oxford: Blackwell Science; 1997.
40. Sterne JAC, Davey Smith G. Sifting the evidence – what's wrong with significance tests? BMJ 2001;322:226-231.
41. Schwartz DS, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. Journal of Chronic Diseases 20:637-648, 1967.
42. Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. Int J Epidemiology 21:837-841, 1992.
43. Peters TJ, Wildschut HIJ, Weiner CP. Epidemiologic considerations in screening. In: Wildschut HIJ, Weiner CP, Peters TJ, eds. When to Screeen in Obstetrics and Gynecology. London: WB Saunders, 1996:1-12.
44. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-674.
45. Lewis JA, Machin D. Intention to treat – who should use ITT? [Editorial] British Journal of Cancer 1993;68:647-650.
46. Richards SH, Bankhead C, Peters TJ, Austoker J, Hobbs FDR, Brown J, Tydeman C, Roberts L, Formby J, Redman V, Wilson S, Sharp DJ. A cluster randomised trial comparing the effectiveness and cost-effectiveness of two primary care interventions aimed at improving attendance for breast screening. Journal of Medical Screening 2001;8:91-98.
47. Bland M. An Introduction to Medical Statistics, 2nd edition. Oxford: Oxford University Press; 1995.
48. Collett D. Modelling Binary Data. London: Chapman & Hall; 1991.
49. Collett D. Modelling Survival Data in Medical Research. London: Chapman & Hall; 1994.
50. Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. BMJ 300; 230-235, 1990
51. Egger M, Juni P, Bartlett C. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285(15):1996-9.
52. Bland M. An Introduction to Medical Statistics, 2nd edition. Oxford: Oxford University Press 1995.
53. Zar JH. Biostatistical analysis, 2nd edition. New Jersey: Prentice-Hall, 1984.
54. Brookes ST, Whitley E, Peters TJ, Mulheran P, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false positives and negatives. Health Technology Assessment 2001;5(33).
55. Moher D, Jones A, Lepage L. Use of the CONSORT statement: revised recommendations for improving the quality of reports of parallel-group trials. Lancet 2001;357(9263):1191-4.
56. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, for the QUORUM Group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. Lancet 1999;354:1896-1900.
57. Elbourne DR, Campbell MK. Extending the CONSORT statement to cluster randomised trials: for discussion. Statistics in Medicine 2001;20:489-496.
58. Berry H, Bloom B, Mace BEW, Hamilton EDB. Expectation and patient preference-does it matter? JR Soc Med 73:34-38, 1980
59. Weber AM, Abrams P, Brubaker L, Cundiff G, Davis G, Dmochowski RR, Fischer J, Hull T, Nygaard I, Weidner AC. The standardization of terminology for researchers in female pelvic floor disorders. Int Urodynecol J Pelvic Floor Dusfunct 2001;12(3):178-86.
60. Mattiasson A: Characterisation of Lower Urinary Tract Disorders-A New View. Neurourol Urodyn 2001; 20(5):601-20.
61. Ryhammer AM, Laurberg S, Djurhuus JC, Hermann AP. No relationship between subjective assessment of urinary incontinence and pad test weight gain in a random population sample of menopausal women. J Urol 1998;159(3):800-3
62. Ryhammer AM, Djurhuus JC, Laurber S. Pad testing in incontinent women: a review. Int Urogynecol J Pelvic Floor Dysft 1999;10(2):111-5
63. Sandvik H, Seim A, Vanvik A, Hunscaar S, A severity index for epidemiological surveys of female urinary incontinence: comparison with 48 hour pad weighing tests. Neurourol Urodyn 2000;19(2):127-45
64. Groutz A, Blaivas JG, Chaikin DC, Resnick NM, Engleman K, Anzalone D, Bryzinski B, Wein AJ, Noninvasive outcome measures of urinary incontinence and lower urinary tract symptoms: a multi-center study of micturition diary and pad tests. J Urol 2000 Sep; 164(3 Pt 1):698-701
65. Nygaard I, Holcomb R, Reproducibility of the seven-day voiding diary in women with stress urinary incontinence. Int Urogynecol J Pelvic Floor Dysfunct 2000;11(1):15-7
66. Kobelt G. Economic considerations and outcome measurement in urge incontinence. Urology. 1997;50(6A Suppl):100-7; discussion 108-10.
67. Groutz A, Blaivas JG, Rosenthal JE. A Simplified urinary incontinence score for the evaluation of treatment outcomes. Neurourol Urodyn 2000;19(2):127-35
68. Groutz A, Blaivas JG, Hyman MJ, Chaikin DC. Pubovaginal sling surgery for simple stress urinary incontinence: analysis by an outcome score. J Urol 2001 May;165(5):1597-600.
69. Groutz A, Blaivas JG, Kesler SS, Weiss JP, Chaikin DC. Outcome results of transurethral collagen injection for female stress incontinence: assessment by urinary incontinence score. J Urol 2000 Dec;164(6):2006-9
70. Raz S and Erickson DR : SEAPI QMM Incontinence Classification System. Neurourol Urodyn 11:187-99, 1992
71. Grady D, Brown JS, Vittinghoff E, et. al.: Postmenopausal hormones and incontinence: the heart and estrogen/progestin replacement study. Obstet Gyecol 2001;97:116-20.
72. Shumaker SA, Wyman JF, Uebersac J, McClish DK, Fantl JA. Health-related quality of life measures for women with urinary incontinence: The urogenital distress inventory and the incontinence impact quetionnaire. Quality of Life Research 3:291-306, 1994.
73. Rogers RG, Kammerer-Doak D, Villarreal A, 2001, Coates K, Qualls C. A new instrument to measure sexual function in women with urinary incontinence or pelvic organ prolapse. Am J Obstet Gynecol 2001;184(4):552-8.
74. Buchner DM and Wagner EH, Preventing frail health, Clin Ceriatr Med 1992;8:1-17

75. Katz S, Ford A, Moskowitz R, et al. The index of ADL: A standardized measurement of biological and psychosocial function. JAMA 185; 914-919, 1963.

76. Mahoney FI, Barthel DW. Functional evaluation: The Barthel Index. Maryland State Med J 14; 61-65, 1965.

77. Folstein MF, Folstein S, McHugh PR. Mini Mental State: A practical method for grading the cognitive state of patients for the clinician. J Psych Res 12; 189-198, 1975.

78. Nih policy and guidelines on the inclusion of children as participants in research involving human subjects, Release Date: March 6, 1998, National Institutes of Health, at: http://grants.nih.gov/grants/ guide/notice-files/not98-024.html

79. Farhat W, Bagli DJ, Capolicchio G, O'Reilly S, Merguerian PA, Khoury A, and McLorie GA. The dysfunctional voiding scoring system: quantitative standardization of dysfunctional voiding symptoms in children. J Urol 2000;164:1011-15.

80. Sureshkumar P, Craig JC, Roy LP, Knight JF. A reproducible pediatric daytime urinary incontinence questionnaire. J Urol 2001;165:569-73.

81. Wein AJ. Pathophysiology and Categorization of Voiding Dysfunction. In Campbell's Urology, 7th edition, Walsh PC et al eds. WB Saunders, Philadelphia, 1998. 917-26

82. McGuire EJ, Woodside JR, Borden TA. Upper tract deterioration in patients with myelodysplasia and detrusor hypertonia: a follow up study. J Urol 129:823-6;1983

83. U.S. Food and Drug Administration – Center for Devices and Radiologic Health. Guidance Documents and Reports. Draft guidance for preclinical and clinical investigations of urethral bulking agents used in the treatment of urinary incontinence. Published November 29, 1995. http://www. fda.gov/cdrh/ode/oderp850.html

84. Collier J and Dwight: Medicines and the NHS. Which Limited, 1997.

85. Collier J: Confusion over use of placebos in clinical trials. Editorial. Br Med J 311: 821-22, 1995.

86. Jones B, Jarvis P, Lewis J, Ebbutt A: Trials to assess equivalence: the importance of rigorous methods. Br Med J 313:36-39, 1996.

87. Herbison P: Letter to editor. Neurourol Urodyn 17: 513-14, 1998.

88. Henry D, Hill S: Comparing treatments. Br Med J 310:1279, 1995.

89. Davidoff F, DeAngelis CD, Drazen JM, Hoey J, Hojgaard L, Horton R, Kotzin S, Nicholls MG, Nylenna M, Overbeke AJPM, Cox HC, Van Der Weyden MB, Wilkes MS: Sponsorship, authorship, and accountability. JAMA 2001;286(10):1232-34.