CHAPTER 3

Committee 25

Research Methodology

Chairman

C. PAYNE (USA),

Members

J. BLAIVAS (USA),

J. BROWN (USA),

M. HIRSCH (USA),

J. KUSEK (USA),

T. PETERS (UK),

W. STEERS (USA),

M.L. STOTHERS (CANADA),

P. VAN KERREBROECK (THE NETHERLANDS),

A. WEBER (USA)

CONTENTS

Acronyms used in Research Chapter:

AUA:	American Urological Association
BOO:	bladder outlet obstruction
BPE:	benign prostatic enlargement
CI:	confidence interval
CONSORT:	Consolidated Standards of Reporting Trials
DMSC:	Data Monitoring and Safety Committee
FDA:	Food and Drug Administration
HRQOL:	Health related quality of life
HT:	hormone therapy
IC:	interstitial cystitis
ICCRN:	Interstitial Cystitis Collaborative Research Network
ICI:	International Consultation on Incontinence
ICS:	International Continence Society
ISD:	intrinsic sphincter deficiency
LUT:	Lower Urinary Tract
LUTS:	Lower Urinary Tract Symptoms
NIDDK:	National Institute of Diabetes, Digestive, and Kidney Diseases
NIH:	National Institutes of Health
PBS/IC:	Painful Bladder Syndrome (including interstitial cystitis)
PFDN:	Pelvic Floor Dysfunction Network
POP-Q:	Pelvic Organ Prolapse Quantification System
RCT:	Randomized Controlled Trial
SE:	standard error
STARD:	Standards for Reporting of Diagnostic Accuracy
SUFU:	Society for Female Urology and Urodynamics
UITN:	Urinary Incontinence Treatment Network
UK:	United Kingdom

Research Methodology

C. Payne,

J. Blaivas, J. Brown, M. hirsch, J. kusek, T. Peters, W. Steers, M.L. Stothers, P. Van Kerrebroeck, A. Weber

I. INTRODUCTION

The International Consultation process focuses on exploring the current knowledge base in order to guide the clinical care of patients by using the highest possible degree of evidence. The assignment for this committee is slightly different. The charge is to define research methodology that will guide today's investigators in producing exceptional work—the kind of data that provides the convincing evidence needed to direct clinical practice, the kind of data that stimulates other investigators and generates new sponsorship, the kind of data that lead to a deeper understanding of physiology and pathophysiology.

Urinary incontinence and lower urinary tract dysfunction comprise a group of common diseases, and far more knowledge of their origin, diagnosis, treatment, and ultimately prevention is needed. Clinical research is a precondition for any progress in these areas. The committee provides here general recommendations for good research practice, including principles of trial design and statistical methodology. In addition, specific recommendations applicable to types of treatments and studies of different groups of patients are presented. 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. This report includes new sections on pelvic organ prolapse, fecal incontinence and painful bladder syndrome/interstitial cystitis in keeping with the broad mission of the ICI to unify and promote research efforts throughout the spectrum of pelvic floor disorders.

The Oxford Centre for Evidence-based Medicine Levels of Evidence and Grades of Recommendation (http://www.cebm.net/levels_of_evidence.asp) are difficult to apply to this section. In order to comply with the spirit of the Consultation the recommendations in this report are graded as follows:

- *High*: Supported by strong evidence (multiple strong publications)
- *Medium*: Supported by moderate evidence (limited/moderate level publications)
- *Low*: Expert/Panel opinion

The report highlights published guidelines produced by the International Continence Society (ICS) [1-14, 193] and Society for Female Urology and Urodynamics (SUFU) [15-17]. Consistent use of the methodology endorsed by these groups will facilitate incontinence research by producing high quality studies. Consistent use of the approved terminology will similarly facilitate communication about research.

The aim of clinical research is clear—namely, to validate treatments that will 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 coupled with the marked demographic trend toward an older society is producing an explosive increase in demand for effective incontinence therapies. At the same time, evidence about the etiology and risk factors for developing incontinence, the optimal treatment strategies, and effective prevention is deficient. The quality of research is of the utmost importance. While there is ample evidence that urinary incontinence is a troublesome disease creating a major impact on patient's quality of life, investigators must vie for research funding with

those studying heart disease, cancer, and innumerable other diseases. Only the highest quality work will be successful in today's competitive environment.

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.

II. GENERAL RECOMMENDATIONS

1. The planning phase of a clinical study on incontinence

The planning stage of both prospective and retrospective studies requires the same deliberate approach. Having formulated a general research question, the investigator reviews previous and, if possible, ongoing work in the field to determine how the research question fits with the known body of knowledge in that area. A thorough knowledge of related clinical work is the cornerstone of protocol development. Care should be taken to identify studies that are well designed and clinically relevant. The research reviews provided by the Cochrane Incontinence Group (http://healthsci.otago.ac.nz/dsm/wch/obstetrics/cure) provide an excellent starting point for most major incontinence topics. This collection of carefully scrutinized data allows researchers to focus their research on key questions. Following the recommendations made by the Cochrane Group will help to ensure that future studies will be interpretable in the context of past work.

Based on the literature review, the investigator clearly elucidates and documents the primary research question, summarizes the background information, and formulates the rationale, objectives and hypotheses for the study.

A rule of thumb for all research is that one should seek the least complex approach capable of answering a given hypothesis or question. The project must provide a convincing answer to the question in an efficient manner. At the same time, it is ideal that an effort be made to discover how a treatment works, not just whether it works. In terms of designing the study, this may be thought of a balance between breadth and depth [18]. Although the number of questions in a single study should be limited, it is still relevant to record as many as observations as is possible without jeopardizing recruitment or retention with onerous demands.

Once the concept of the study has been clearly defined, the search for an appropriate funding agency commences. The chance of a successful application increases when the investigator ensures that the application meets the mandate of the funding agency.

2. Study design

The type of study and other aspects of study design are the framework within which the study objectives are met. An initial decision must be made as to whether the study will be *observational* or *experimental*. Experimental studies are where the investigators control the process by which it is decided which treatment a participant should receive, while in observational studies the treatment decisions are not in any way influenced by the investigators and the research study. The strength of the scientific evidence arising from various experimental study designs are ranked as follows:

Strength of study design (ranked in descending order of strength)

a) Randomized controlled clinical trial(s)

i. Double-blinded: Neither the participants nor the investigators (in particular, those responsible for outcome assessment) know which subjects are receiving the active treatment while the study is in progress.

ii. Non-blinded (open): The investigators and/or participants know which treatment is being given.

The randomized controlled trial (RCT) is the gold standard study design. The participants are assigned to a particular treatment group by a mechanism designed by the investigators and based on a chance allocation to the various treatments. Provided that adequate concealment is maintained, neither the patient nor the investigator can influence to which group any particular participant will be assigned. This provides protection from allocation bias by the investigator and/or subjects. RCTs are expensive to conduct and can occasionally be ethically problematic. However, the central importance of the RCT in terms of influencing decisions about patient care should and will continue.

b) Non-randomized controlled clinical trial(s)

This category includes trials in which the basis for treatment allocation is known to the investigator prior to obtaining informed consent (for example, day of clinic appointment). This type of study design provides no confidence that the treatment and non-treatment groups are comparable.

c) Case series

Case series are studies that describe the outcome when all subjects receive the treatment being investigated. These are the weakest form of study design, but they may be the only available or practical approach, particular for rare diseases, when treatments in the study arms are markedly different in nature, or when the therapy became established prior to the acceptance of RCTs in medical research. Because these designs do not have internal controls, external observations must be used for comparison, raising questions about patient selection and comparability with other populations.

Properly planned and executed, the RCT is the optimal approach to limiting allocation bias [19]. RCTs compare outcomes in groups of subjects for whom treatment was allocated by chance. Using this approach, treatment groups will not vary in any systematic fashion, and comparisons between them will be unbiased [20]. Subject assignment must be concealed during enrollment (for example, by separating allocation from the process of recruiting subjects, and by using remote randomization such as by telephone), and wherever possible treatment allocation must be concealed during the trial (for example, using blinding with or without placebo). In some studies, blinding of subjects and health care providers may be impossible or undesirable, for example in trials of some surgical procedures or health care delivery methods. In all cases, however, the personnel collecting outcome data should be blinded to the subjects' treatment allocation.

Observational studies include a variety of designs, from cross-sectional descriptive studies in which the primary purpose is estimation of the prevalence of incontinence in a defined population, to case-control studies and long-term prospective or retrospective cohort studies useful in studying rare diseases. Observational studies may be purely descriptive (case series), or they may be analytic when designed with a control or comparison group. Observational studies can contribute useful information on many aspects of health care [21], and may be necessary precursors to a randomized trial. However, all comparisons based on observational data have a common limitation – the inability to ensure that one is comparing like with like. In particular, it is not possible, even with advanced statistical methods, to eliminate the bias resulting from the effects of the selection process, whether induced by patient or clinician.

Although the classical RCT involves parallel groups, other options are possible and may overcome some of the limitations of the classical approach [22]:

- Parallel Trials: These designs offer one group of subjects the treatment under study, and a parallel group a placebo or alternative treatment. Within drug trials the dosages may be held constant or varied to either maximize clinical benefit or minimize side effects. Sophistications around the basic design can in some circumstances be worth considering – for example, factorial trials where two or more interventions can be investigated simultaneously [23, 24], and cluster randomized trials whereby groups of participants rather than individuals are randomly allocated to the trial arms [25]. This strategy might be employed when studying an intervention requiring policy changes in an institution; thus a hospital, clinic, or health care system may be the unit of randomization.
- Crossover Trials: Subjects receive both the treatment being studied and the placebo/alternative treatment, with the order in which the treatments are received being randomly assigned. The benefit of crossover studies is that they eliminate the effect of variation between groups of participants seen in parallel trials. 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 change in symptoms in response to treatment. The duration of treatment effect is critical in determining whether the crossover study design is appropriate – too long an effect, and the disease state may become unstable before the patient has completed all arms of the study; too short, and it may not be possible to detect the effect during the period of data collection. Carryover effects may occur, in which the results of the first treatment are prolonged and affect the results of the second treatment. To avoid this, a washout period should be planned, in which participants receive either placebo or no treatment. A run-in period in which signs and symptoms are monitored may be necessary before treatment commences to ensure that only those whose disease state is stable are included. Given these features and limitations, this design is unlikely to be widely applicable in studies of interventions for incontinence.

- 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 design may be appropriate when one treatment is considerably more cost-effective, offers a better quality of life, or is less toxic or time consuming for the patient while producing a similar clinical outcome. The observed difference in outcome between two treatments should be clinically unimportant and be accompanied by a narrow confidence interval. An equivalence trial can be a powerful design when appropriately employed, and is not the same as failing to find a difference between two groups. Clinically unimportant differences may be quite small, necessitating large sample sizes [26, 27]. Stringent statistical methods are needed to ensure that there is adequate power to exclude an important difference between treatments, and that failure to demonstrate a difference is not merely a consequence of poor study design and procedures. Particular caution is required in applying these trials in the context of (especially pragmatic) trials where non-trivial proportions of participants switch from one intervention to another.

Drug trials are categorized according to the following definitions [26] [28].

- **Phase I studies:** The first studies of a drug in humans, often open and uncontrolled, concentrating on safety and frequently but not exclusively carried out in healthy volunteers. Pharmacokinetic and tolerance information is obtained from Phase I trials.
- **Phase II studies:** The first attempts to investigate treatment efficacy, often the first use of the drug in subjects and focusing on short-term outcomes. A common objective of Phase II studies is dose finding in terms of efficacy. Two sub-types may usefully be distinguished: Phase IIA studies where single treatments are considered in relation to a minimum response prior to further investigation; Phase IIB where direct comparisons are made between interventions, albeit on a small scale and not necessarily involving randomization [29].
- **Phase III studies:** Large-scale, authoritative randomized studies performed once the most likely effective and tolerated treatment regimens have been established. The objective is often to establish that the intervention is suitable for registration with the appropriate regulatory authority. Trials are conducted after submission of a new drug application (NDA), but before the product's approval for market launch. Phase IIIB trials (between submission for approval and receipt of marketing authorization) may supplement or complete earlier trials, or seek different kinds of information (for example, quality of life or marketing). Phase III trials are also used to investigate the effectiveness and cost-effectiveness of various interventions – that is, non-drug including organizational issues – and not necessarily with reference to regulatory authorities. All Phase III trials should be subject to a formal sample size calculation – for instance to obtain sufficiently precise estimates of the comparisons between treatments or to have a reasonable chance (power) of detecting a difference if one exists (see section II C 7 below).
- **Phase IV studies:** These investigations are usually carried out after registration, to investigate the drug's safety and efficacy in different populations. Such postmarketing surveillance studies are typically larger and simpler than regulatory studies; they may lack a control group and are often conducted using surveys.

Precisely which study design to choose to answer a given primary research question depends on a number of factors, including the ability to recruit sufficient participants for a particular design (see Statistical Considerations, below), the natural history of the disease, the treatment itself, and patient endpoints. Patient-related endpoints may be short-term, such as changes in signs or symptoms, or longer-term, such as increased survival.

Once the sample size required to answer the primary research question has been calculated, it is usually obvious whether the study can be performed at a single institution, or whether a multicenter study will be required. Single institution studies have the benefit of being less complicated from a logistical perspective.

While multicenter trials are more complex to manage and are usually more expensive, they provide larger numbers of participants in a shorter period of time, and increase the generalizability of research findings.

Common mistakes that can occur during the planning and conduct of a study are described in **Figure 1**.

Figure 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

3. Study conduct and statistical considerations

Research must not only be planned early, but also planned often. All issues should be addressed at the start of the planning process, and many will need to be revisited 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 [27]. Regarding presentation, the Consolidated Standards of Reporting Trials (CONSORT— http://www.consort-statement.org) statement provides guidelines for reporting the design, detailed methods, and results of RCTs [30]. The original statement [31, 32] has since been revised with the aim of improving clarity and, where appropriate, increasing flexibility and coverage of sophisticated designs [33-37]. Many of the points discussed here relate to those guidelines, which should be closely followed throughout the design, conduct, analysis and presentation of RCTs as is required by most leading medical journals. The 22 item checklist is presented in **Figure 2**. For studies of diagnostic tests the Standards for Reporting of Diagnostic Accuracy (STARD— http://www.consort-statement.org/stard-statement.htm) [38] statement fills the same role.

a) Sampling strategies

Whether a study is analytic or descriptive (that is, whether or not a comparison is involved), the first practical issue to resolve is the selection of participants. A study may require a sample that is representative of the community overall or one representative of patient groups suffering the condition/disease. In principle, this is achieved by taking a simple random sample from a known population. In practice, a list of all eligible 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 [39].

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 subjects in an RCT (and indeed there are similar issues in trials relating to stratification and clustering) [25], 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, RCTs are drawn from a subset of the population, often limited to those with access to academic medical centers plus the willingness and ability to participate. Where this is the case, it is crucial to provide descriptive information about the study sample, so that broad representativeness can be judged. 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 [33, 34, 40]. Nevertheless, "the basic logic of clinical trials is comparative and not representative" [26]. In other words, the principal benefit of conducting a randomized trial is to provide groups that allow valid comparisons to be made.

CONSORT **Checklist of items to include when reporting a randomized trial**

PAPER SECTION And topic	Item	Description	Reported on Page #
TITLE & ABSTRACT	1	How participants were allocated to interventions (*e.g.*, "random allocation", "randomized", or "randomly assigned").	
INTRODUCTION Background	2	Scientific background and explanation of rationale.	
METHODS Participants	3	Eligibility criteria for participants and the settings and locations where the data were collected.	
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered.	
Objectives	5	Specific objectives and hypotheses.	
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (*e.g.*, multiple observations, training of assessors).	
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules.	
Randomization -- Sequence generation	8	Method used to generate the random allocation sequence, including details of any restrictions (*e.g.*, blocking, stratification)	
Randomization -- Allocation concealment	9	Method used to implement the random allocation sequence (*e.g.*, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.	
Randomization -- Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.	
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. When relevant, how the success of blinding was evaluated.	
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s); Methods for additional analyses, such as subgroup analyses and adjusted analyses.	
RESULTS Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.	
Recruitment	14	Dates defining the periods of recruitment and follow-up.	
Baseline data	15	Baseline demographic and clinical characteristics of each group.	
Numbers analyzed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by "intention-to-treat". State the results in absolute numbers when feasible (*e.g.*, 10/20, not 50%).	
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (*e.g.*, 95% confidence interval).	
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory.	
Adverse events	19	All important adverse events or side effects in each intervention group.	
DISCUSSION Interpretation	20	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.	
Generalizability	21	Generalizability (external validity) of the trial findings.	
Overall evidence	22	General interpretation of the results in the context of current evidence.	

Figure 2

b) Randomization and stratification

Randomization is the process of allocating subjects to groups by chance [19, 20]. 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 minimizing bias.

In order to minimize bias, the randomization process must be concealed from those recruiting subjects to the trial [33, 40]. This can be achieved most effectively by the use of central telephone randomization. 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.

1. 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.

2. 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. In any case, blocking prevents serious imbalances in characteristics across groups when used in conjunction with stratification as described below.

Most disease states have factors known to influence the outcome of treatment, for example, pre-and postmenopausal or male and female. A form of randomization that accounts for such factors is called *stratified randomization* [19, 20]. Stratified randomization ensures equal distribution of subjects with a particular characteristic in each group when blocking is employed within strata. 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 [19, 20].

c) 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 participants, clinicians, regulatory bodies, and health care purchasers). In almost all situations, the outcome set should include both a dimension representing the viewpoint of the patient (including symptom frequency, severity, and quality of life) as well as an appropriate objective 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 statistical corrections should be considered as described below. Analyses of secondary outcomes are best viewed as exploratory, i.e., as hypothesis-generating exercises for which independent confirmation is essential.

d) 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. 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. 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 subjects. 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 criteria (e.g., avoiding nephrotoxic drugs in subjects with renal insufficiency). All criteria should be defined precisely enough to allow the study to be reproduced by other groups of researchers.

e) Informed consent

Peer review of protocols by a multidisciplinary team may include members of the scientific community, clinicians, pharmacists, the public/patient groups, the legal profession and individuals who can provide an ethical perspective. 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 extreme, they involve exhaustive pages of information, which explain every alternative treatment with 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 preferred 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, written using terms that are meaningful to potential participants [41, 42].

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 interim analyses to ensure that a treatment is not producing unacceptable levels of side effects and/or efficacy. Rules for stopping the study should be agreed upon, prior to the start of the trial, and should be subjected to independent statistical and clinical scrutiny, usually through a Data Monitoring and Safety Committee (DMSC). It is important to note that interim analyses (in particular those based on efficacy) will have implications for the study power – specifically, a larger sample size will be required eventually to achieve the same power/precision compared with if such analyses are not conducted (see section III 3g below). Specialist statistical advice and support will be essential to address these issues [43-45]. 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, and so an important role for the DMSC is to provide implicit reassurance on this point. Investigators should not be aware of the results of interim analyses, however, since this may cause bias by influencing how vigorously any given patient is recruited into the study. Nevertheless, emergency procedures for unblinding a patient's allocation are required in case of a severe side effect or concomitant serious illness.

f) Bias, blinding and effects on validity

Any part of the human ensemble in an RCT can introduce bias (systematic error), which can result in erroneous conclusions regarding treatment effects. Bias can occur in every aspect of an RCT 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 consciously or uncons-

ciously 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 [33, 40]. While blinding is important its effect is lower than that of concealment of allocation [46].

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 the participants and the research personnel responsible for the measurement and assessment of outcome [20, 33]. 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 the final data analysis. This may be justifiable for the primary analysis of trial data, but at some pointing proceedings there is often a strong case for unblinding the data analyst. For example, another opportunity for bias occurs if an appreciable number of subjects drop out or withdraw from a study, in the sense that they fail to provide outcome data. Such attrition can be particularly problematic if it 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, which may well necessitate unblinding in the final analysis.

g) 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 [47-49]. 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.

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 subjects; 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 [27], 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 [20]. This is the basis for the adage that "the absence of evidence is not evidence of absence" [26]. A planned unequal allocation to the trial groups also requires an inflation of the sample size [20], 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 [26, 48].

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. For example, a study might propose that a 20% difference in incontinence episode frequency is a clinically meaningful response. Ideally, such an assumption would be based on surveys of patient behavior but in practice the decision is often arbitrary. In any case, 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 [50]. 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 [49]. 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 that 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.

In all cases, appropriate adjustment for attrition (loss to follow-up) should be performed. This is common-

ly achieved by simply increasing the planned sample size in proportion to the anticipated attrition (i.e. to predict the reduced effective sample size that will be available for the analysis). To adhere to the 'intention-to-treat' principle described in section II 3h below, however, it might be more consistent if this adjustment were achieved by a realistic reduction in the target difference to account for any dilution of this that would result from assumptions made about missing follow-up data.

h) Pragmatic and explanatory trials

There is an important distinction between *pragmatic* and *explanatory* trials [51, 52], and correspondingly, between *intention-to-treat* and *per-protocol* approaches to data analysis [26, 48]. This distinction has a number of facets. For example, data from pragmatic trials 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 (generalizability).

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 [26]. 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 [53].

As regards other aspects of the distinction, 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 in comparison with explanatory trials. In addition, pragmatic trials may involve a wide variety of outcome domains, including patient-completed questionnaires, and an economic evaluation of outcomes. As a result 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; therefore, full data analysis ideally incorporates both intention-to-treat and per-protocol approaches [26]. The primary analysis, though, should follow the intention-to-treat principle.

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 at a fixed time relative to randomization should, however, 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 [33, 40]. 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 [54], 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 [55]. 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 [54]. Sensitivity analyses can be used to test the exclusion of, or assumptions about, missing values; practical examples of such analyses are becoming more common [56].

i) Data analysis

This section will not contain any technical details of statistical methods, which are available in standard

texts [20, 57, 58], 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 (i.e., allowing for the possibility that the difference could have been in either direction), 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% [33]; indeed, the p-value is better considered a measure of the strength of evidence against the null hypothesis, on a continuum or 'shades-of-grey [50, 59]. 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 [33].

Data analysis for numerical outcome variables may use parametric or non-parametric methods. Simple parametric methods require that the data follow a normal or Gaussian distribution, while non-parametric methods do not have this requirement. Strictly speaking, the distributional assumption relates not to the raw data but to what are termed 'residuals' – that is, the outcome variable after the effects of for example the treatment effects and baseline variation have been accounted for. For example, consider a comparison of mean urinary flow rates between two groups of men reporting that they either have or have not experienced a urinary symptom such as incomplete emptying of the bladder. If this comparison involved an unpaired t-test then the assumption of a normal distribution relates not to the urinary flow rates amongst all men, but to the distributions within each symptom group separately. As noted above, to cover simple and more complex analyses, in general this concept relates to the statistical residuals from the relevant regression model. In addition, parametric methods are extremely robust to the assumption of a normal distribution since it relates to the mean value rather than individual values.

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 [22]. Binary outcome variables can be analyzed using chi-square tests and confidence intervals for comparing proportions, and multiple logistic regression [60]. 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 [61].

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 [33, 40], and it is now considered good practice for the trial team to draw up a detailed analysis plan in advance for approval by the Trial Steering Committee, which includes independent members one of whom is a statistician/trials methodologist. 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 of outcomes and interim analyses involve considerably more complex methods of planning and analysis, for which expert help is essential [62].

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. In order to reflect representativeness accurately, this should include all eligible patients: in practice there is a tendency for researchers to avoid approaching certain potentially eligible patients, for any of a wide variety of reasons, and this induces a subtle investigator bias. The presentation of this information is facilitated by use of the CONSORT flow diagram [33, 40] (**Figure 3**)—indeed, its use is associated with improved quality of reporting of trials generally [34]. Descriptive statistics should also 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.

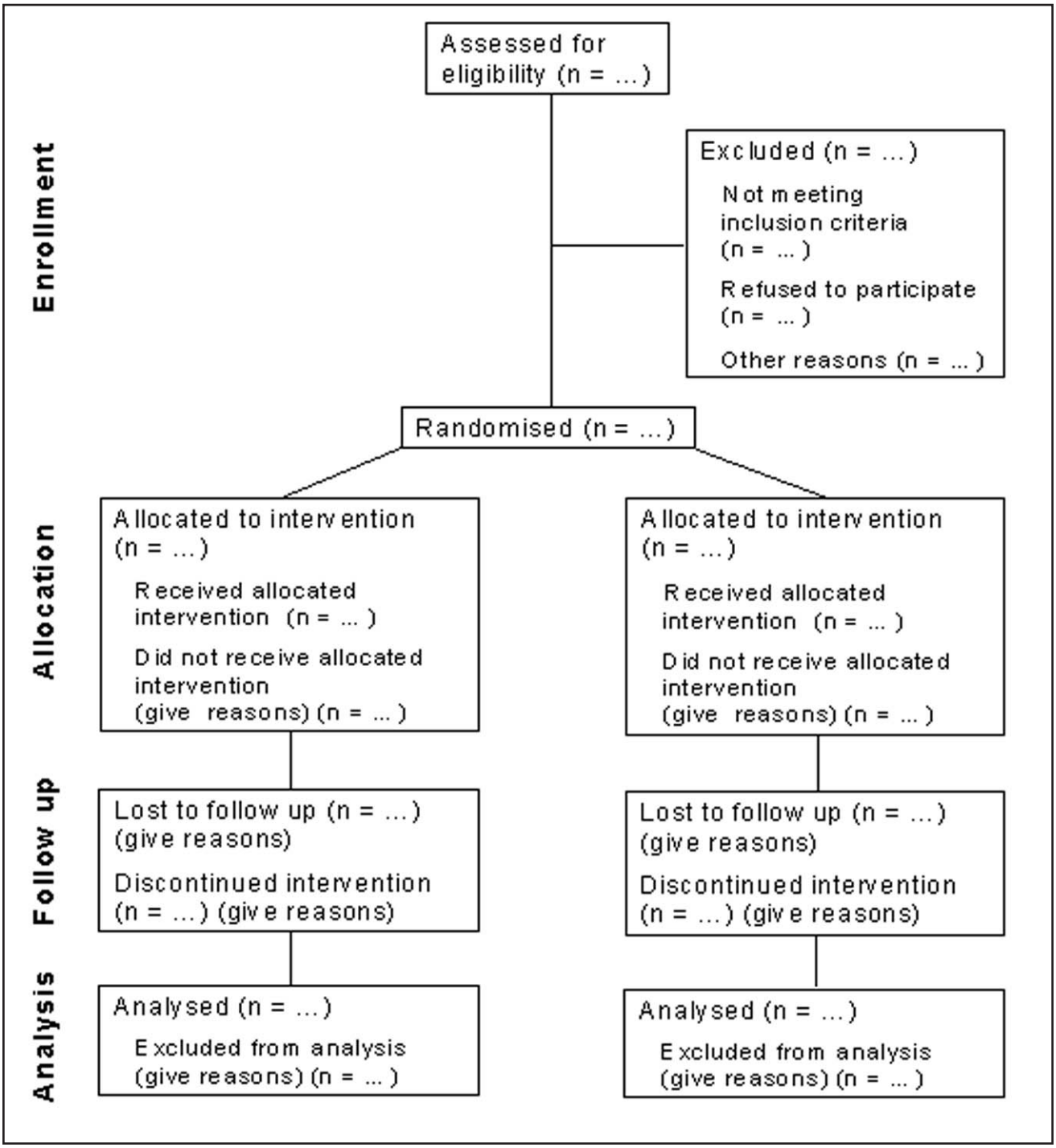

Figure 3 . Flow diagram of the progress though the phase of a randomised trial

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, 26, 48]. 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, 26, 48]. 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 [26]. Investigators should consider stratifying the randomization on any strongly prognostic variable (for reasons of efficiency rather than bias). Since there are practical limitations as to how many variables a trial can stratify for, as indicated above in section C 2, a technique known as minimization may also be considered [19, 26]. Any variables stratified or minimized at randomization should be allowed for in the analysis [26].

The next stage of data analysis is to perform the primary (comparative) analyses for the outcome variables. First, though, it is essential to derive and report actual numeric data – even if simply in the form of descriptive statistics – rather than just reporting for instance a percentage change, even if the latter are relevant and provided as well. Graphs can be misleading, especially when sub-sections of the scales are magnified, and should be used to supplement or clarify the numerical data, not to replace it. Primary outcomes should initially be analyzed by intention-to-treat comparisons of the groups as randomized, both using hypothesis tests for statistical significance and 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.

The most commonly used procedure for multiple testing of many outcomes is the Bonferroni correction [20, 26, 57]. 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 [26]. 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 [26]. 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 [63].

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 multi-center trials [27]. The correct approach for continuous outcome variables is to use the (regression-based) technique known as the analysis of covariance [26, 48]; 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 [26, 48]. 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 [27].

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 the investigation of different intervention effects across age, ethnic, or disease severity groups. Subgroups should be analyzed by using appropriate interaction terms in regression models [33, 48]. Using interaction terms rather than performing repeated, separate, subgroup-specific analyses considerably reduces the risk of false positive findings [64, 65]. Subgroup analyses should be carried out sparingly, specified in advance (preferably with a clinical rationale), and above all should not be reported selectively [33, 64, 66]. This last point relates not just to subgroup analyses but to all stages of reporting randomized trials. Pre-specification of the primary outcomes and clear statements about all the outcomes considered is essential to avoid selective reporting.

j) Reporting of randomized controlled trials

The CONSORT statement is specifically designed to provide standards for reporting RCTs [33, 40]. Adherence to these guidelines and the use of flow diagrams in particular is associated with improved quality in reporting of RCTs [67]. Errors in presentation of statistical information are extensively covered in many textbooks [20, 57]. 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 [68]. However, such guidelines are not a panacea [36]; deficiencies in reporting are still common [67].

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 participant but occasionally an aggregate group of subjects will be allocated jointly in a cluster randomized design [25]. Cluster randomized designs are also now the subject of separate reporting guidelines [37], and involve particular complications in terms of data analysis [69]. 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, 26]. Results should include a trial flow diagram, with numbers and reasons for the exclusion of eligible subjects, the number randomized, and subsequent losses to follow-up [34]. Protocol deviations should be described and explained [48]. 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 [33].

k) 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 from the outset. Statistics has been described as a combination of mathematics, logic and judgment [26], and this applies to all phases of RCTs. 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.

Recommendations on Study Conduct and Statistical Methods

- The role of quality RCTs as providing the strongest level of evidence in incontinence research should be fully acknowledged by researchers, journal reviewers, and editors. **HIGH**
- Careful attention to the planning and design of all research, especially RCTs, is of the utmost importance. **HIGH**
- Appropriate expertise in biostatistics and clinical trial design should be employed at the design phase of a RCT and thereafter on an ongoing basis. **HIGH**
- The design, conduct, analysis and presentation of RCTs must be fully in accordance with the CONSORT guidelines. **HIGH**
- Reporting studies of diagnostic tests, including urodynamics, should follow the STARD statement guidelines. **HIGH**

3. Outcomes Research in LUTS Including Incontinence

No single measure can fully express the outcome of an intervention. While every clinical trial must focus on a few primary endpoints, 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, and by how much. 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 simultaneous consideration of outcomes, 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, detailed data is needed 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. Perceptions of the patient, doctor or therapist are frequently at variance and this must be reported. Participants' expectations may also influence the outcome of a study [70].

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 [11]. 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

a) Baseline clinical and demographic data:

Data collection in clinical research begins with complete demographic description of the study participants including age, race, sex, duration of symptoms, prior treatments, comorbidities, medications. 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 [71]. While few trials will be large enough to analyze the effect of these demographic factors on outcome, the potential future use of meta-analysis makes a complete database valuable.

b) Observations:

1. Participants' observations and Subjective measures:

Validated patient completed symptom questionnaires and other validated instruments are recommended for all trials for LUTS and incontinence (see report from Committee 6). 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. 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

Measures should include the frequency of the symptom (e.g., daily urge incontinence), the severity of the symptom (e.g., pads are saturated) and the impact or bother produced by the presence of the symptom (e. g., much greater for the individual who works in a public setting). There is no one instrument covering all of these areas which has 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. For a detailed discussion of available instruments see the report of Committee 6.

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 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 the key factors determining success or failure, for example do patients who continue to leak after a bladder neck suspension do so because of technical factors (recurrent hypermobility) or due to an inherent limitation of the procedure (Type III stress incontinence). Similarly, reports of biofeedback training for incontinence provide data about continence after intervention but little information about muscular function is provided. Do participants 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. Only complete evaluation of both structure and function will ultimately lead to an optimal classification of LUT disorders.

Recommendations on Observations During Incontinence Research

- One or more high quality, validated symptom instruments should be chosen at the outset of a clinical trial to accurately define baseline symptoms and any other areas in which the treatment may produce an effect. **HIGH**
- Observations of anatomy should be recorded using standardized, reproducible measurements. **HIGH**
- Pelvic muscle and voluntary sphincter function should be reported using a quantifiable scale. **HIGH**
- All observations should be repeated after intervention and throughout follow-up and their relationships with primary clinical outcome measures investigated. The duration of follow-up has been inadequate
- in many studies. Given the nature of the disorder, "short term" follow-up in all types of incontinence trials should begin with all participants having reached one year. RCTs should be extended into cohort studies whenever practical. **HIGH**

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. Recorded data may include fluid intake, voiding frequency (diurnal and nocturnal), frequency of incontinence episodes by type, pad use, and voided volumes. Accuracy depends on proper training of the subjects. 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. Longer diaries are more reliable but have decreased subject compliance; the ideal length is not known. 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 [72-76]. The period of time the diary was used should be noted [77].

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 provide additional insight. 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 or provocative tests, generally performed under standardized conditions as office tests, and long-term tests, generally performed at home over 24 to 48 hours. 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 compliance. For short-term tests, the experimental conditions must include standardized bladder volumes and the tasks must be described in detail. The pad test quantifies incontinence; therefore it can provide a 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.

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 ; in most instances there is inadequate information about reproducibility, accuracy, sensitivity, specificity, and predictive value.

Participants 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. A primary research goal should be to collect data to determine the predictive value of urodynamic testing prior to intervention.

It would be ideal to use urodynamic studies routinely to accurately characterize baseline lower urinary tract function and dysfunction and the change after treatment. However this is impractical in most RCTs and unnecessary to answer the basic questions involved. At the same time, urodynamic tests are among the best tools currently available to understand the basic physiology and mechanisms of disease; these tests must somehow play a role in research. While routine incorporation of urodynamic testing is not warranted in clinical research, hypotheses driven inquiries aimed to refine the studies, define the utility of specific tests in different patient populations, and to develop new and better tools would be most welcome.

Interpretation of urodynamic signals remains an art—namely 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. Procedures for performing urodynamic studies must be carefully standardized in trials 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 (e.g., investigators will be biased to read baseline data to allow inclusion in study protocols). In multicenter studies, this may be minimized by using a central reader for urodynamic tracings, after detailed annotation by the primary observer.

Recommendations on Tests used in Incontinence Research

- Clinical trials of incontinence and LUTS should include bladder diaries as an essential baseline and outcome measure. **HIGH**
- The diary should include measured voided volume (for at least one day). **HIGH**
- Pad tests should be considered in clinical trials when practical. Continued research into the validity, sensitivity and utility of pad tests is needed before stronger recommendations can be made about the role of such tests as a primary outcome measure. **MEDIUM**
- Urodynamic studies have not been proven to have adequate sensitivity, specificity or predictive value to justify routine use of testing as entry criteria or outcome measures in clinical trials. Most large scale clinical studies should enroll subjects by carefully defined symptom driven criteria. **HIGH**
- High quality, hypothesis driven research into the utility of using urodynamic studies to define patient populations or risk groups within clinical trials is greatly needed. **HIGH**
- In all trials employing urodynamics, standardized protocols (based on ICS recommendations) should be defined at the outset. In multicenter trials, urodynamic tests should be interpreted by a central reader to minimize bias unless inter- and intrarater reliability has already been established by standardized procedures within the trial. **HIGH**

d) Follow-up

Minimal standards for evaluation of treatment outcomes in urinary incontinence have been presented [15] in a report approved by the AUA and SUFU. The 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 subjects 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 decrease the frequency of incontinence episodes, but also how a treatment affects a patient globally. 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 touched upon. Generic HRQOL instruments such as the SF-36 are designed to be used across groups by having established age and gender norms. Condition-specific instruments such as the Leicester Impact Scale for lower urinary tract symptoms [78] 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 "bother" related to 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 sleep disturbance. A practical approach to assessing HRQOL is to combine a multidimensional generic and/or condition-specific instrument with dimension-specific instruments appropriate for the trial.

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 on Health Related Quality of Life in Incontinence Research:

- Research in incontinence and LUTS should include both generic and condition-specific validated HRQOL instruments whenever practical and appropriate. **HIGH**
- Changes in HRQOL after therapy should be considered in relation to changes in individual symptoms, and with physiologic and anatomic outcome measures. **LOW**

f) Socioeconomic Data as Outcome Measures [79]

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 difficult to determine 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 in an unforeseen manner.

In health and medicine economic analyses are descriptive and/or comparative. Descriptive data include the socioeconomic cost caused by the disease and its current treatment, whereas comparative data provide an economic evaluation of different treatment strategies and interventions where costs are compared to health outcomes.

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 several 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 actually a subgroup of CEA, where outcome is expressed as a single measure incorporating survival and quality of life, usually quality-adjusted life years (QALY). Cost utility analysis allows comparisons of treatments in different diseases.
  - Cost Consequence Analysis (CCA) is another

variation of CEA presenting a comprehensive list of the expected costs and health outcomes in a tabular form without totaling the various costs or placing value on the outcomes. This disaggregated approach provides greater flexibility to users in applying the relevant findings to other settings [80].

- 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 (monetary). 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, comprehensive coverage of all the relevant dimensions of disease would be incorporated into economic evaluations.

Recommendations on Socioeconomic Outcomes in Incontinence:

- The type of economic evaluation (or at least the general strategy to be adopted) should be chosen before starting a trial, based on the specific objectives to be addressed. Analysis is based on intention-to-treat, and attrition must be handled in the same way as for the primary outcomes analysis of the trial. **HIGH**
- Very few economic evaluations have been done in the field of urinary incontinence and more are greatly needed to understand the cost implications of therapeutic decisions. Researchers should consider both a condition-specific outcome measure to use with the economic evaluation, as well as a quality of life instrument with utility properties to allow for comparison with other diseases. **MEDIUM**
- The assumptions underlying any economic investigations should be subject to sensitivity analysis. **HIGH**

III. CONSIDERATIONS FOR SPECIFIC PATIENT GROUPS

1. Men with LUTS, including incontinence

Four unique factors influence research on lower urinary tract symptoms in adult men:

1. the confounding influence of the presence of the prostate
2. the likelihood of bladder outlet obstruction (BOO)
3. the rarity of sphincteric incontinence, and
4. the rarity of any kind of incontinence at all in young and middle aged men except under two special circumstances – neurogenic bladder and after treatment for prostate cancer.

The presence of the prostate complicates research because of its known effect in causing bladder outlet obstruction (with or without benign prostatic enlargement—BPE), its propensity for developing prostate cancer and the effects of treatment for both conditions, resulting in sphincteric incontinence (after prostate surgery) and incontinence due to detrusor overactivity (after radiation based therapies). Further, prostate size itself may influence outcomes in some therapies of urinary incontinence, for example the type of surgical intervention in men with prostatic obstruction and urinary incontinence. Overall, about 2/3 of men with LUTS have urethral obstruction and over 50 % have detrusor overactivity, although a much smaller number have urinary incontinence due to detrusor overactivity [81]. The rarity of incontinence in men suggests that gender specific bother and quality of life instruments will be necessary.

a) The presence of the prostate:

Aside from obstruction (see below), the prostate impacts on outcomes research in incontinence in three ways – prostatic size, the possibility of undiagnosed malignancy, and prior therapy (prostatic surgery and other therapies for BOO, BPE or prostate cancer). If prostate size is believed to be a variable that could affect outcomes, measurement of prostate 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 of the treatment's mechanism of action, but of course, prostate volume should be

measured at the same time as the primary outcome measurements for incontinence are determined. The association between outcome and change in prostate size should be reported. Consideration should be given to stratifying participants by prostate volume when there is suspicion that response to therapy may be size dependent.

Since the presence of undiagnosed malignancy might affect outcomes, participants should be screened for prostate cancer by digital rectal examination and measurement of serum PSA and appropriate disposition should be made based on the outcome of such testing. A careful history of past surgery and other treatment of BPE, BOO, and prostate cancer should be obtained and dealt with based on the inclusion and exclusion criteria for the study.

b) Likelihood of bladder outlet obstruction (BOO):

Insofar as about 2/3 of men with LUTS have BOO, it is important in any research protocol to screen for its presence. At the least, uroflow and measurement of post-void residual urine should be recorded pre-treatment and the effect of therapy on these parameters should be documented simultaneously with assessment of the primary outcome variables. More sophisticated urodynamic studies add expense and have not yet been shown to predict response to treatment; while synchronous pressure-flow studies are generally desirable and should be included whenever feasible there should be a focus on hypothesis driven research that will produce clear evidence about the utility of these studies. Results should be presented as stated in the ICS 1997 Standardization Report on Pressure Flow Studies of Voiding, Urethral Resistance and Urethral Obstruction." 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.

c) The rarity of sphincteric incontinence:

The rarity of sphincteric incontinence in men has one practical consequence – any man with sphincteric incontinence should undergo an exhaustive neurologic evaluation unless it is consequent to prostatic surgery or a neurologic condition known to cause sphincteric incontinence, such as a thoracolumbar neurologic lesion.

d) The rarity of any kind of incontinence at all in young and middle aged men except as noted above:

Young and middle aged men only rarely have incontinence unless they have a neurologic condition, prior prostatic surgery or severe bladder outlet obstruction. With advancing age, there is a gradually increasing incidence of detrusor overactivity and incontinence. For this reason, it may be important to use gender and even age specific quality of life and bother scores when assessing these outcome measures.

Recommendations for Research in Men:

- High quality, symptom and bother scores (e.g., IPSS, ICS male, DAN-PSS) validated in men should be employed when assessing outcome in male incontinence research. **HIGH**
- Uroflow and measurement of post-void residual urine should be recorded pre-treatment and the effect of therapy on these parameters should be documented simultaneously with assessment of the primary outcome variables. **HIGH**
- Measurement of prostate size (or at least Prostatic Specific Antigen) should be performed before and after treatment (synchronous with other outcome measures) whenever prostate size is expected to change due to the treatment. **HIGH**
- Participants should be stratified by prostate size at randomization when size is considered to be a potentially important determinant of treatment outcome. **LOW**
- Although bladder outlet obstruction is a urodynamic diagnosis, synchronous detrusor pressure uroflow and videourodynamic studies have not been shown to predict treatment outcome in the general population. Further research in this area would be highly desirable in order to define relevant subgroups of men who may respond differently to various treatments. **LOW**

2. Women with LUTS and incontinence

We concur with the 1997 Urodynamics Society recommendations for outcome research in women [15, 16]. We also refer to the ICS recommendations for outcome measures in women with lower urinary tract dysfunction [12] and the Proceedings of the NIH Terminology Workshop for Researchers in Female Pelvic Floor Disorders [71]. Unique factors influencing research on lower urinary tract symptoms in adult women include:

a) Hormonal effects

b) Obstetric history

c) Pelvic organ prolapse

d) Gender related outcome measures

e) Sexual functioning

All of these potentially confounding variables can affect the outcome of treatment of incontinence.

a) Hormonal effects:

Our knowledge of hormonal influences on the lower tract remains limited. Recent RCTs and prospective cohort studies have demonstrated that (HT) does not improve or may worsen incontinence [82-84]. It therefore seems appropriate that information about menstrual and hormonal status should an integral part of the baseline history. Studies designed to examine the influence of hormones on incontinence should include menopausal status (premenopausal, postmenopausal without HT, post-menopausal with HT), whether or not oophorectomy has been performed, and the type, dosage and route of administration of HT if used.

b) Obstetric History:

The unique influence of vaginal childbirth on the structure and function of the female pelvis remains incompletely understood. That childbirth may lead to incontinence and pelvic organ prolapse is indisputable; the potential effect of childbirth on treatment of incontinence has yet to be determined. 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. Potentially confounding variables include: number and route of deliveries (vaginal/Cesarean), use of forceps or suction devices, infant birthweight, duration of second stage of labor, use of midline versus mediolateral episiotomy, obstetric analgesia, and obstetric complications such as lacerations and fistulae.

c) Pelvic Organ Prolapse:

The effect of pelvic organ prolapse on lower urinary tract function remains controversial and understudied. It has been suggested pelvic organ prolapse may affect lower urinary tract function in at least four ways. It may cause urethral obstruction, it may mask sphincteric incontinence, it may cause urgency and urge incontinence, and/or it may diffuse pressure transmission, making it more difficult to void by abdominal straining. Some of these effects may be immediately reversible with reduction of the prolapse. For all of these reasons, it is essential to include assessment of pelvic organ prolapse in incontinence research. Methods used for the assessment of pelvic organ prolapse should be stated and reliability and validity data should be provided if available or their absence indicated; the Pelvic Organ Prolapse Quantification System (POP-Q) [9] is recommended as discussed by committees 5 and 17. Further, whenever prolapse is present, instruments for assessing incontinence and LUTS should be reported with the prolapse reduced and again at its full extent whenever possible. In either event, prolapse should be graded at the same time as the outcome assessment for incontinence and LUTS is performed.

The degree of urethral mobility is considered an etiologic and prognostic factor in women with urinary incontinence, although its precise role has not been defined. At present, there is neither a well accepted method of assessment nor a classification system. Common methods of assessment include measurement of the Q-tip angle (cotton swab test), magnetic resonance imaging, ultrasound, and cystogram. In the absence of a classification system, it is recommended that data be presented as a continuum, such as the Q-tip angle, not as a dichotomous normal versus abnormal classification, at least until the terminology is better defined. Methods used for the assessment of urethral mobility should be stated and reliability and validity data should be provided if available or their absence indicated.

d) Definition of Outcomes Measures for LUTS & Incontinence:

Treatment of urinary incontinence in women may have broad ranging effects on lower urinary tract function, prolapse, sexual function, and bowel function. It is therefore important that a broad perspective of outcome be presented. The NIH Terminology Workshop for Researchers in Female Pelvic Floor Disorders recommendations [71] for stress incontinence treatment define outcomes as cure/ improved/failed in terms of 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 strength of the definitions proposed by NIH is the amalgamation of subjective and objective measures as well as the recognition of the potential adverse effects of incontinence therapy. However, there is as yet little experience with this system. 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 relia-

bly 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. It may not be appropriate to report a woman cured of stress incontinence as failed if she develops urinary tract infections or a rectocele several years later. 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. In any case, 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 in the materials and methods section. Further, possible (and likely) discordant outcomes should be described and categorized. For example, a woman might state that she is cured of incontinence, have a negative pad test and diary, yet stress incontinence might be demonstrated on a stress test with a full bladder.

An alternative view of a global outcome measure has been described [85]. Here, incontinence is first assessed and the patient declared cured, improved or failed based on outcome measures derived from the criteria listed below. If the patient is not cured (i.e. dry) the reasons ascribed are defined as sphincteric incontinence, detrusor overactivity incontinence, extraurethral incontinence (fistula) or incontinence of undetermined etiology. The system can be used for studies of both stress and urgency incontinence. Once continence status has been determined, other LUTS should be described using LUTS outcome instruments. In all instances, methods used for the assessment of incontinence and LUTS should be stated and reliability and validity data should be provided if available or their absence indicated. In such a scheme, continence could be cured yet the patient might experience de novo or persistent LUTS that mitigate against the patient considering herself cured or improved. Examples of outcome instruments that take these objective, semi-objective and subjective measures into account are the SEAPI-QMM system [86] and the Simplified Urinary Outcome Scores [87].

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.

e)Sexual function:

Urinary incontinence, LUTS and the treatment of these disorders all have potential effects on sexual function yet little is known about the impact of incontinence treatment on sexual function. It is therefore appropriate that sexual function be considered one of the domains for investigation in all types of incontinence research. Validated instruments that deal with sexual function in women with LUTS and incontinence include the Incontinence Impact Questionnaire (IIQ) [85] and the Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire (PISQ) [88].

A final issue relating to research methodology in female incontinence is defining the population for studies of stress incontinence. One group has described a clinical algorithm that might be used to select patients without urodynamic testing [89].

RECOMMENDATIONS FOR RESEARCH IN WOMEN:

- Specific information about menopausal status, hysterectomy, parity/obstetric history, and hormonal status should be included in baseline clinical trial data. **HIGH**
- Standardized assessment of pelvic organ prolapse should be performed before treatment and synchronous with other outcome assessments in all surgical trials and whenever relevant. **HIGH**
- Strict criteria for cure/improve/fail of incontinence should be defined based on patient perception as well as objective and semi-objective instruments such as validated questionnaires, diaries and pad tests. **HIGH**
- De-novo or persistent LUTS should be evaluated concomitantly with other outcome assessments and are best done with validated instruments. There may be a role for urodynamics in defining the etiology of these problems. **MEDIUM**
- Assessment of the impact of treatment on sexual function should be performed synchronously with other outcome assessment when appropriate. **MEDIUM**

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" [14]. In addition, please refer to the full report of Committee 13 regarding conservative treatment in the elderly. Frailty is defined as "a state of reduced physiological reserve associated with increased susceptibility to disability [90]." There is 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. Consequently, the frail older population is a heterogeneous group residing in a broad range of care settings, with multiple medical conditions related and unrelated to the lower urinary tract, and often on numerous medications. There are a number of unique and pertinent research issues for this population.

a) Prevalence, natural history, and risk factors

There remains a paucity of research on the prevalence and incidence of urinary incontinence to assess the burden of disease in the frail older population. This information is vitally important to estimate health care costs and direct resource planning. Less information is available on risk factors and this lack of established risk factors limits preventive efforts and highlights the need for increased epidemiologic research.

Representative samples of well-defined target populations are necessary in different care settings including the community (independent and homebound), acute inpatient, and nursing home or institutionalized (bedfast and non-bedfast). Well-conducted data collections are important to provide careful measurement of a wide variety of potential risk factors and a large sample to allow for statistical adjustment for multiple potential confounding variables to identify independent risk factors for incontinence. Cross-sectional studies are efficient to determine prevalence and identify potential risk factors while longitudinal studies are needed to estimate incidence, understand the natural history, and to define causality of risk factors. There have been few longitudinal studies of incontinence in this group.

Unfortunately, most prior studies were small, did not differentiate functional or cognitive impairment, failed to identify medical conditions or medication use. Most of the analyses were not adjusted for potential confounding variables as mentioned or age, body weight, parity, etc. It is important to include previously reported risk factors and potential risk factors. Validated instruments should be used when available. Otherwise, a detailed description of how the risk factor was measured is required to assess how well the variable represents the area of interest. In the frail elderly, important variables include:

- Demographic information: Advancing age, white race, and women [84, 91-93] are at an increase risk of incontinence and each of these variables should be adjusted for in most analyses**.**
- Reproductive History: Childbirth is an established risk factor for incontinence [94],[95] and it is reasonable to collect data on the number of births. However, in women over 60 years of age, as chronic medical illnesses become more prevalent and impact incontinence, risk profiles change and parity may not remain a significant risk factor [93, 96-98]. Prior hysterectomy has also been suggested as a potential risk factor for incontinence in older women [98, 99]
- Medical Conditions: Medical conditions related and unrelated to the lower urinary tract have been shown to increase risk of incontinence in older women and are especially important to assess in the frail older population [93, 96-98].
- Medication Inventory: Certain medications may exacerbate incontinence and therefore a complete medication inventory is essential [100-103].
- Physical function: Mobility is often impaired in the frail elderly and impacts urinary control [104], therefore mobility should be assessed using validated instruments such as the Bartel Orcats or ADL scales [105, 106]. Data on walking aids or wheelchairs, gait speed, and manual dexterity may also be collected.
- Cognitive function: Cognitive function impairment and/or dementia increase risk of incontinence [104]. The Mini-Mental Status Scale Examination [107] assesses global cognitive function and the Confusion Assessment Method (CAM) [108] is a standardized assessment for delirium. A battery of neuropsychological tests to measure subtle impairments in cognitive function include the Buschke Selective Reminding Test (verbal learning and memory) [109]**,** the Digit Symbol (incidental memory, visual scanning and motor speed) [110], and the Trails A (attention and visual) [111].
- Environmental factors: In the frail elderly, envi-

ronmental factors may also contribute to incontinence: toilet access, usual continence care in the facility, and a description of caregivers and their training may be useful.

b) Outcome measures

An outcome measure should be reproducible (test-retest), accurate (sensitive and specific), feasible (balance of risks, costs, acceptability, ease of use), and sensitive to change over time. Additionally, the outcome measure should be clinically relevant and meaningful. It must be acknowledged that almost all outcome measures used in the study of incontinence that have been shown to be reliable and valid in the community dwelling population require separate validation for use with the frail elderly. Whatever outcome measures are chosen should be described in terms of applicability to the frail elderly.

Commonly used self-reported measures of frequency of urinary symptoms, severity, or level of bother may not be possible in the cognitively impaired frail elderly patient. Similarly, voiding diaries that have been shown to be to valid and reliable in assessing urinary frequency, nocturia, and incontinence episodes by type [75, 76, 112, 113] may not be feasible or reliable. Motivated and trained staff, caregivers, or family members may be able to adequately collect diary data; however, this has not been validated.

In nursing home or inpatient settings, wet checks by staff at set intervals have been used in a number of studies. There are limitations to the measurement including visually determining what is "wet" because of new absorbent materials and staff reports not always being reliable or valid due to underreporting [104], [114]. To overcome the limitation of defining wetness and underreporting, 24 hour pad weighing tests [115, 116] may be used. Pad weighing tests and wet checks are feasible and can provide important outcome data if staff are well trained and checks are often [117].

The usefulness of cystometry, simple or complex, as an outcome measure in the frail elderly remains unclear. Cystometry is invasive, difficult to perform, has poor reproducibility, and has not been shown to be clinically useful by demonstrating the improved outcomes in randomized controlled trials [115, 118]. A post-void residual volume is useful as a screening tool prior to an intervention that may exacerbate urinary retention (pharmacologic or surgical). It is also a useful outcome measure of adverse events in intervention studies by demonstrating the development of urinary retention. It is easily performed by ultrasound or catheter and has been shown to be reproducible and accurate [119-123].

Although a primary outcome is needed for sample size estimation, it is useful to have several outcome measures that assess different aspects of urinary incontinence.

Evaluation of incontinence bother and effect on quality of life is pertinent to the patient and may also be important from the perspective of the staff, caregivers, and family members. New outcome measures specific to the frail older population such as increased socialization or decreased caregiver burden need to be developed. Having multiple outcomes can provide a more detailed description of the effect on urinary incontinence.

c) Intervention trials

Prior to initiation of pharmacologic or surgical intervention trials in the frail elderly, careful consideration of the risks and benefits is important because of the increased risk of adverse side effects or events. In addition to the urinary incontinence outcome measures mentioned above, extensive outcome measures that will detect adverse effects from the intervention are important to demonstrate that the beneficial effect of the intervention outweighs the adverse events.

For example, in trials of new medications using a battery of neuropsychological tests to measure subtle impairments in cognitive function would be important but thus far not been done. In intervention trials of new medications or operations, clinically significant outcome measures (global patient satisfaction with improvement and consideration of staff, caregiver, and family satisfaction perspectives) that demonstrate substantial effect sizes (clinical significance) rather than "statistically significant" improvements are particularly important in the frail elderly.

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 validation of 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 for Research in Frail Older and Disabled People:

- There is a need for validation of all instruments and procedures used in incontinence research for the population of frail elderly as well as development of new study measures in multiple domains of incontinence. These measures need to be evaluated for reproducibility, accuracy, feasibility, and effects on clinical decisions and outcomes. **HIGH**
- Clinically important outcome measures and relationships of outcome to socioeconomic costs are essential to establishing the utility of treating urinary incontinence in this population. **HIGH**

4. Incontinence in children

The conduct of clinical research in children is generally more difficult than in adults. 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 [124] 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 a pediatric population 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—principal 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 who 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: neurogenic (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 are often on medication beginning at a very young age and continuing for many years; 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. While children should always be offered the standard of care when such exists, so few treatments have ever been studied properly in children that 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 [125] and daytime incontinence [126]; 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 for Research in Children:

- 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. **HIGH**
- Research is needed to develop standardized outcome measures including validated, age-specific symptom and disease-specific quality of life outcome measures. **MEDIUM**
- 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. **LOW**

5. Neurogenic lower urinary Tract dysfunction

In the past, renal failure was a leading cause of death in the spinal cord injured population and a feared complication of many neurologic conditions. Modern neurourologic care is generally successful in maintaining renal function and preventing other upper urinary tract complications, affording social continence, and advancing 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 the General Recommendations above and focus on the specific characteristics of the neurogenic patient. Specific discussion of treatment in the neurogenic population is contained in reports from committees 12. Reports from committees 3, 7, and 8 are also relevant to this population. Statistical methods and research outcomes are applicable 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.

It is difficult to find a classification system of neurogenic voiding dysfunction as a base for research that is satisfactory for each of the three aims. The published systems have been reviewed in detail [127]. 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 method that meets the broad needs of classification in this group. Typical or classic cases are often well described but it is especially difficult to describe mixed and incomplete lesions. Thus, 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 that apply to other than a narrow segment of the neurogenic population. Ideally a broad population of potentially relevant participants would be enrolled in research studies with full characterization of both the neurologic condition and the nature of the lower urinary tract dysfunction so as to allow for subgroup analysis.

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 neurogenic patient. Baseline data collected by history in subjects with neurogenic lower urinary tract disorders should include:

- bladder volumes by diary or examination (maximum voided volume, post voiding residual urine, total capacity);
- 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.

Issues such as mobility, independence in activities of daily living, cognition, 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 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, often with combined 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 subjects with neurogenic detrusor overactivity and coordinated sphincters will be easier to interpret than one of neurogenic urge incontinence or multiple sclerosis).

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 normal or an overactive filling phase, compliance is also of major importance. Elevated detrusor leak point pressure predicts upper urinary tract deterioration in children with myelomeningocele [128] and is presumed to be 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 pdet and Q curves) because of the possibilities of functional obstruction and impaired contractility, which are uncommon outside of the neurogenic population. Because the bladder and sphincter may be dyssynergic, assessment of sphincteric activity is

essential. This may be accomplished by surface electromyography of the pelvic floor, needle electrodes, fluoroscopy, ultrasound, or direct measurement of urethral pressure.

Recommendations for Research in Neurogenic Lower Urinary Tract Dysfunction:

- 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. **MEDIUM**
- Change in detrusor leak point pressure should be reported as an outcome as appropriate, and can be considered a primary outcome for spina bifida subjects. **HIGH**
- An area of high priority for research is the development of a classification system to define neurogenic disturbances. Relevant features could include the underlying diagnosis, the symptoms, more precise documentation of the neuromuscular lesion by clinical neurophysiologic testing, and the nature of the urodynamic abnormality. **LOW**
- It may sometimes be appropriate to group participants with urodynamically similar neurogenic bladder disorders of different etiologies in a clinical trial. However, great caution must be used if subjects with progressive disease (e.g., multiple sclerosis) are grouped with subjects having a stable deficit (e.g., traumatic spinal cord injury). **HIGH**

6. Fecal incontinence

Men and women with urinary incontinence frequently have coexistent problems with the posterior compartment, such as fecal incontinence, fecal urgency, constipation, chronic pain (such as levator syndrome or proctalgia fugax), solitary rectal ulcer syndrome, or rectal prolapse. Discussion of all these conditions is beyond the scope of this chapter, which will focus on research methodology in the study of fecal incontinence. Fecal incontinence has been among the least studied of all pelvic floor disorders. There is a high degree of coexistence of urinary symptoms and fecal incontinence [129]; this warrants focusing more attention to comprehensive evaluation and management of pelvic symptoms. Many definitions for fecal incontinence have been created for use in clinical research, but there is still no consensus on a single instrument or group of instruments that is ideal for assessment of outcomes. Therefore, the following comments are by necessity non-prescriptive; we anticipate that future research findings will guide the development of more specific recommendations.

For the purposes of this discussion, fecal incontinence includes impaired ability to control either stool (liquid or solid) or gas. It is important to emphasize that fecal incontinence is a symptom and, as such, must be measured through subjective assessment [130]. The subjective evaluation of fecal incontinence in clinical research requires measurement of at least two aspects of the condition: severity and impact. Severity can be assessed by either grading scales or summary measures [131-133], and includes a component of frequency of episodes over a specified period of time. Fecal urgency, while not an integral part of the definition of fecal incontinence itself, may have a marked impact on quality of life as patients restrict their activities to avoid fecal incontinence [133]. The impact of fecal incontinence is best measured by disease-specific quality of life instruments [134, 135]. The need for one standardized system of evaluation and quantitation of symptoms (which may include questionnaires, diaries, and quality of life assessment) is a high priority for further research in this field.

In contrast to urinary incontinence, the objective demonstration of fecal incontinence is not a component of current definitions. Physical examination should include screening for pelvic organ prolapse (in some studies such as surgical studies of fecal incontinence, the standardized quantification system for staging of prolapse [9] should be used); rectal examination; assessment of pelvic muscle function; and screening pelvic neurologic examination. Other specific test results are not included in the definitions. However, certain tests may be useful in identifying different subcategories of fecal incontinence, such as a specific sphincteric defect versus generalized atrophy. If any component of anorectal anatomy or function could reasonably be expected to change with treatment, consideration should be given to performing such tests before and after intervention. For example, if anal ultrasonography is used to detect sphincter defects preoperatively, it would be beneficial to obtain anal ultrasound postoperatively as well, to study the association between anatomic change and change in subjective assessment (i.e., symptoms).

In defining the impact of interventions on fecal incontinence, cure is defined as complete resolution of the symptom. Failed treatment (persistence or recurrence) is defined as no improvement or worsening of symptoms. Without evidence, improvement cannot be specifically defined at this time but may include a favorable change in symptoms related to severity or impact or both. Simply reporting a statistically significant average improvement in grading or summary scores in a group of subjects is not very informative; measuring within-patient change is more informative as to the true impact of the intervention. Further research is needed to develop clinically meaningful levels of improvement after intervention.

Recommendations for Research in Fecal Incontinence:

- Due to the high concordance of fecal and urinary incontinence, and the potential for urinary incontinence therapy to affect bowel function, data on fecal incontinence should be collected at the outset and during trials of urinary incontinence whenever practical. **HIGH**

7. Painful bladder syndrome (including Interstitial Cystitis)

The prevalence of Painful Bladder Syndrome (including interstitial cystitis) (PBS/IC) has been estimated at 52 to 67 per 100,000 adult women in the United States [136]. The importance of the condition is underscored by its inclusion in this Consultation for the first time. The report of committee 21 reviews the knowledge base for this disease in detail.

There is a great controversy over the definition of PBS/IC and the appropriate population for research studies. In 1987 the NIDDK sponsored consensus conference resulting in a research definition of interstitial cystitis [137]. The definition encompasses inclusion criteria that describe the syndrome and exclusion criteria that serve to create a relatively homogeneous patient population. One study examined the performance of the NIDDK criteria [138]. The authors found that 90% of the subjects meeting NIDDK criteria were felt to have interstitial cystitis by experts. However, over 60% of participants diagnosed with interstitial cystitis by the same experts did not meet the strict criteria. Therefore, use of the strict NIDDK criteria may exclude 2/3 of appropriate subjects and diminish the impact of trials because the patient population is not representative of the PBS/IC population at large. Such criteria could also select for patients with more severe/chronic disease who may be less likely to respond to intervention. It has been suggested [139] that very inclusive criteria be used in PBS/IC trials to improve the generalizability of results.

One possible direction could come from the new ICS definition of "painful bladder syndrome" which is defined as, "the complaint of suprapubic pain related to bladder filling, accompanied by other symptoms such as increased daytime and night-time frequency, in the absence of proven urinary infection or other obvious pathology" [140]. Cystoscopic findings are not part of the definition. It is further proposed that, "interstitial cystitis is a specific diagnosis and requires confirmation by typical cystoscopic and histological features". This issue may not easily be settled as there appear to be dramatically different worldwide perceptions as to the true nature of PBS/IC as manifest by willingness of different investigators to make the diagnosis [141]. For now, it would appear that there is much to be gained from trials with both inclusive and restricted entry criteria and there will be some that will be most appropriate for each strategy. Ultimately, antiproliferative factor and associated bladder growth factors may be useful in defining a homogeneous and biologically meaningful patient population for research.

Once the population to be studied is decided there has been considerable progress in defining appropriate methods of conducting clinical trials. Key clinical trial design issues have been reviewed [139]. The Interstitial Cystitis Collaborative Research Network (ICCRN, http://porter.cceb. upenn.edu:7778/servlet/page?pageid=234,238& dad=portal30& schema= PORTAL30), a group funded by the NIDDK, is composed of 10 centers in North America and a Data Coordinating Center. The ICCRN has completed randomized clinical trials using both oral [142] and intravesical agents [143]. These provide excellent templates for the investigator in the planning phase of a project. The current project is focused on recruiting newly diagnosed patients in order to learn more about the natural history of the disease and the response of such patients to treatment with amitriptyline. There is general agreement that the primary outcome must be patient driven and the ICCRN has used the Global Response Assessment. The Global Response Assessment asks subjects to rate their symptoms, as compared to baseline, on a seven-point centered scale: markedly worse, moderately worse,

slightly worse, no change, slightly improved, moderately improved, and markedly improved. Typically those responding moderately or markedly improved are considered responders. A full spectrum of objective and subjective secondary endpoints will be required to fully characterize the treatment effect. Special consideration should be given to examination of specific subgroups such as patients with Hunner's ulcers, newly diagnosed patients, and male patients.

Recommendations for Research in Interstitial Cystitis and Painful Bladder Syndromes:

- The patient population for PBS/IC trials must be carefully defined. When appropriate, relaxed entry criteria should be used to reflect the full spectrum of the PBS/IC patient population. **MEDIUM**
- The primary endpoint of PBS/IC trials should be patient driven and the Global Response Assessment is recommended. A rich spectrum of secondary endpoints will be useful in defining the effect of treatments. **MEDIUM**
- Biomarkers, especially antiproliferative factor, hold promise for defining a biologically distinct group of participants in future research trials. **HIGH**

8. Pelvic prolapse

For the purposes of this discussion, pelvic organ prolapse includes anterior vaginal prolapse (previously known as cystocele), apical or uterine prolapse, posterior vaginal prolapse (previously known as rectocele), enterocele, and perineal descent; it does not include rectal prolapse. Ideally, for clinical research purposes, prolapse would be defined by three components: (1) by the presence and severity of symptoms, with some indication of bother or impact on quality of life; (2) by signs obtained at physical examination; and (3) by testing, depending on specific study goals. However, singly or in combination, all three components are severely limited in their capacity to distinguish "normal" from "abnormal" regarding prolapse.

Most pelvic symptoms are highly nonspecific and do not show strong associations with the location (anterior, apical, or posterior compartment) or stage of prolapse [144, 145]. One exception to this is the patient's awareness of an actual bulge or protrusion, which has high positive and negative predictive values for Stage III or IV prolapse (but not the affected compartment of prolapse). As a consequence, specific symptoms cannot currently be required in the definition of prolapse (with the possible exception of tissue protrusion). By the same token, the resolution of specific symptoms cannot be required in the definition of "cure" after treatment; however, surgeons should state which symptoms (other than the local bulge) that he/she plans/hopes to cure/improve. Nevertheless, it is essential to include a comprehensive survey of pelvic symptoms to be able to describe what symptoms are present at baseline and which symptoms change with treatment. For example, the Pelvic Floor Distress Inventory (PFDI) includes subscales on prolapse, urinary function, and coloanal function; its companion, the Pelvic Floor Impact Questionnaire, surveys the degree to which symptoms impact quality of life [146]. These condition-specific instruments are relatively new and, although validated, have not yet been shown to be sensitive to change with treatment; however, their comprehensiveness makes them attractive for inclusion in studies where women may have or develop different pelvic floor disorders. Sexual function should be specifically assessed, especially in studies of surgical treatment for prolapse. The ICI questionnaire is one instrument that covers all of these areas of interest although experience with its use is limited.

Unfortunately, the relationship between symptoms and anatomy as measured by the POP-Q are poorly understood. The point at which symptoms may be attributed to the anatomical prolapse is not known. The POP-Q system is a validated, quantitative system with excellent inter- and intra-rater reliability. It includes measurement in centimeters of six vaginal sites relative to the hymen, plus three other measurements for total vaginal length, perineal body, and genital hiatus [9] as described in chapters 5 and 17.

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 magnetic resonance imaging). 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.

At the time of the 1999 NIH Terminology Conference [71], prolapse was defined based on the POP-Q as any prolapse greater than Stage 0. However, research findings since then have challenged that definition. Stage II prolapse can be found in up to 48% of women presenting for preventive health care; Stage III prolapse, in 2-4% [147-150]. The POP-Q staging system currently categorizes women with prolapse at and one centimeter above and below the hymen as Stage II; this combines many women who are asymptomatic (perhaps pre-clinical) with women who are symptomatic. Until further research becomes available, we recommend that the physical examination for research subjects in studies of prolapse include: (1) use of the standardized POP-Q system for prolapse staging; (2) rectovaginal and anal sphincter examination; and (3) assessment of pelvic muscle function.

Although clinical experience strongly supports that prolapse develops gradually over years, it is not known whether Stage II prolapse predicts future support loss or, if it does, in how many women and over what time course. This is particularly controversial in choosing a cutoff for what constitutes clinically significant persistent or recurrent prolapse after surgical treatment. The location of prolapse is important in this choice as well. While both Stage II anterior and apical vaginal prolapse may be asymptomatic, most surgeons would not be willing to accept Stage II apical prolapse as a successful surgical treatment, yet Stage II anterior (or posterior) asymptomatic vaginal prolapse is often considered to be acceptable. While planning clinical trials, the goal of the surgical treatment (restoration to Stage 0, restoration to Stage 1, etc. with or without symptoms) should be clearly specified. Durability of all types of prolapse surgery requires longitudinal, long-term follow-up, which is generally not available in the current literature. The utility of supplementary materials is also not known, and is discussed further in Chapter 17 (Pelvic Organ Prolapse).

Primarily due to lack of evidence, prolapse is not currently defined based on specific test results. We recommend that further research be performed to investigate the usefulness of various tests (for example, imaging by X-ray contrast or ultrasound) in determining definitions and outcomes of prolapse treatment. In a specific study, if an aspect of anatomy will be directly influenced by surgery, it may be reasonable to perform imaging before and after surgery to assess whether any change seen on imaging is correlated with changes in symptoms and physical findings. It is equally important that key factors to patient satisfaction be identified in order to develop better patient driven outcome measures.

Recommendations for Research in Pelvic Organ Prolapse:

- It is critically important to determine what constitutes clinically significant prolapse. This must include:
  - Development of patient reported outcomes
  - A focus on Stage 2 prolapse—its natural history and treatment outcomes **MEDIUM**

9. Nocturia

While nocturia has often been viewed as just one of the manifestations of lower urinary tract dysfunction, there has been a recent appreciation that it is an important independent symptom that does not "belong to" another clinical problem. Waking at night to void can be the result of abnormal fluid intake or other behavioral issues, cardiac conditions, peripheral venous disease, and sleep disorders as well as lower urinary tract dysfunction. This complexity makes clinical research in nocturia very difficult; the fact that the condition is affected by so many different issues mandates a multidisciplinary research team. Although nocturia is one of the most bothersome of symptoms there has been little progress in research outside of studies in pediatric nocturnal enuresis. Only one randomized controlled trial in adults was identified in a literature search. It found that desmopressin was significantly more effective than placebo in reducing nocturia in a population of adult women with at least two voids per night [151]. The ICS Standardization sub-committee report on nocturia should greatly facilitate research efforts; it defines the relevant terms, introduces a flow-chart for evaluation, and presents tables of the causes for the various subtypes of nocturia [152]. Potential investigators should carefully review this document in the planning stage of research.

Recommendations for Research in Nocturia:

- Population or community based research is needed to define the epidemiology of nocturia and how the symptom relates to normal aging. **MEDIUM**
- Clinical research in treatment of nocturia should

begin with classification of subjects by voiding diary categories—polyuria, nocturnal polyuria, and apparent bladder storage disorders. If desired, those with low bladder capacity can be further divided into those with sleep disturbances and those with primary lower urinary tract dysfunction. **HIGH**

- The impact of nocturia on falls and fractures deserves further investigation **MEDIUM**

IV. CONSIDERATIONS FOR SPECIFIC TYPES OF 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. Although these treatments are generally very safe and applicable to most incontinent patients, there should be no compromise in the quality of clinical research.

The type of therapy must be defined with sufficient detail for other investigators to 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 characterized. The usual clinical outcome measures suffice. In order to progress in our understanding of these treatments it is important to consider clinical outcome alongside 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. Recent studies using a standardized self-help booklet are commended as important steps in defining this issue [153, 154].

It can be 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 to demand the 'most optimal and possible level of blinding'. A relevant control group is established, allocation to treatment groups is concealed, as many persons as possible are blinded (in particular, those individuals recording outcome measures), and appropriate measures surrounding this issue are discussed in the manuscripts.

Recommendations for Behavioral and Physiotherapy Research:

- There is a great need for long term data to define the durability of effect for all conservative treatment modalities in all population groups. **HIGH**
- Intervention protocols must be detailed to the degree that the work can easily be reproduced. **HIGH**
- A structured examination of pelvic floor neuromuscular function should be included before and after treatments that are aimed at pelvic muscle training. **HIGH**
- More work is needed to separate the specific and non-specific effects of treatment. **MEDIUM**
- Consider a variety of methodologies in development, evaluation, and interpretation of these interventions. **MEDIUM**

2. Device Trials

In the United States, devices for urinary incontinence are regulated by the Center for Devices and Radiological Health (http://www.fda.gov/cdrh), a branch of the United States Food and Drug Administration (FDA). Although urethral 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 these treatments, yet other devices used in incontinence therapy may elude careful scrutiny. For example, materials for reconstructive surgery are primarily approved based on biocompatibility testing. However, implantation in less than a sterile environment (e.g. vagina), placement during concomitant operations (e.g. hysterectomy, prolapse repair) and effects of biofilms adjacent to the urinary tract could pose unique conditions with subsequent complications in the long term. Very little is known about the effectiveness of such devices at the time of approval.

Detailed guidelines for studies on intra-urethral urethral bulking agents were published in 1995 [155]. Guidelines for implantable devices such as used for neuromodulation fall under different criteria. Guidelines for engineered tissues and cell therapy are evolving. Any researcher considering investigation of bulking agents should be familiar with the FDA document, which outlines the entire conduct of studies from design through outcome measures. For the most part, these guidelines follow the general recommendations. However, the document is now dated and some specific issues invite comment:

1. Inclusion is limited to subjects 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 65cm H_2O". There is no evidence to support any particular cutoff, and the clinical significance of a particular value is questionable given the wide variation in techniques for leak point pressure measurement. Most investigators recognize that abdominal leak point pressures fall along a continuum and that no cut point defines ISD.

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 subjects taking medications affecting the bladder. Many such patients could potentially benefit from therapy, are evaluable, 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.

7. Durability with devices and bulking agents has been limited and often reported sooner than 12 months. Although many investigators recommend 2 year outcome data for pragmatic reasons, the need for indefinite multiyear follow-up is recognized.

8. With devices, patient convenience is often forgotten. Although this outcome may be captured in health related quality of life instruments it represents a significant issue especially for intraurethral or intravaginal devices.

An 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 participants in the study including the reasons for subject withdrawal.

Recommendations for Research in Incontinence Devices

- Safety and serious side effects of incontinence devices must be completely defined with adequate follow-up, especially for use of implantable devices and biologic materials, so that risks can be weighed against efficacy. At a minimum, this requires more use of large scale, prospective, multicenter prospective cohort studies when RCTs are not feasible. **HIGH**
- Physiologic testing such as urodynamics should be considered, in addition to survey instruments, to substantiate the proposed biologic effect of devices especially when a placebo or sham is not available. **MEDIUM**
- Clear, updated guidelines for each of the categories of new devices should be developed that protect patient safety while promoting research in a practical manner. **LOW**
- Patients deserve complete information about implantable devices when considering surgical therapy. New devices may be introduced into the market with no or minimal track record of safety and efficacy for the proposed use. In such cases it may be best to have separate surgical consents for the operation and use of the new device. **LOW**

3. Pharmacotherapy trials

Drug trials are necessary so that new drugs can be clinically and scientifically evaluated for quality, efficacy and safety [60, 106-110]. Since the 1960's administrative bodies such as the Food and Drug Administration have required that new pharmaceuticals undergo controlled investigations to establish efficacy. The specific stages and of study design have been discussed in detail in section IIB. Many large RCTs have been conducted in recent years and incontinence research has generally benefited from the attention to the field and emphasis on valid outcomes. As the financial backing of the pharmaceutical industry has been largely responsible for this research, new conflicts of interest and problems have arisen due to the changing economics. 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" [111]. Several of these issues are discussed in Section V3 below.

Although many RCTs have been published in recent years on pharmacotherapy for urinary incontinence a great deal more remains to be learned. The trials have almost all been limited to 8-12 weeks of treatment giving very little information about long term safety and efficacy of drug therapy. Studies have typically been performed in isolation, i.e. drug vs. placebo, as opposed to a real life scenario where drug therapy is combined with behavioral and pelvic floor therapy. There is less than adequate information about special patient groups—men, children, neurogenic patients, and especially the frail elderly. Because incontinence creates such an impact on the older population good studies to define the utility and safety of drug therapy are greatly needed in this group.

An 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. For drugs in that same class that are already available, an active control agent should be used

whenever possible. In addition, comparator trials of behavioral therapy versus drug or device versus drug are lacking, as are combination studies. Combination studies are especially relevant since such strategies are common in clinical practice despite the absence of data. Researchers must carefully consider these issues in designing a relevant, ethical study. General issues related to placebos are discussed in section V1 below.

Recommendations for Pharmacotherapy Trials:

- Every consideration should be given to making sure that the interests of the subjects are kept at the forefront in designing safe, ethical research. In urinary incontinence safe, effective conservative therapy is available for the vast majority of patients. In most trials, comparison should be with "standard therapy" rather than placebo/no treatment. This approach respects practical patient management where placebo is not an option. **HIGH**
- As effective drug therapy is available for most forms of incontinence comparator arms are recommended for most trials. Claims of superiority of drugs even within the same class are unfounded in the absence of randomized comparator trials. **HIGH**
- Very little is known about the safety, efficacy and tolerability of drug therapy beyond 12 week trials. A concerted effort is needed to create this type of information base. **MEDIUM**

4. Surgical trials

Standards for surgical trials are detailed in recommendations from the CONSORT Organization (http://www.consort.org), ICS, SUFU, and the AUA [12-17]. 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 (and females undergoing surgery for repair of pelvic organ prolapse). Unique research issues for surgical research using observational studies and randomized controlled trials will be presented and insights from other surgical specialties will be discussed.

a) Observational studies

Observational studies are important major sources of descriptive data to understand the patterns of use for surgical procedures and of factors that influence these patterns. A few observational studies that included representative samples with well-conducted data collections have been reported and will serve as examples for future research.

Cross-sectional studies of surgical procedures by type can provide estimates of prevalence, variation by age, race, and region as well as morbidity and mortality. Using the US National Hospital Discharge Survey, it has been determined that incontinence operations have increased in frequency over time, are the third most common surgery in women, that there are large regional and racial differences in the rates of incontinence surgery, and morbidity and mortality are low [156, 157]. This type of information raises important health policy questions regarding physician practices, patient preferences for incontinence treatment, and differential access to and the utilization of care. In another large US national cross-sectional study, participants reported satisfaction with surgery even if they had not achieved complete continence, demonstrating the importance of patient reported outcomes [158].

The largest prospective cohort study published to date included all women undergoing the three most common operations for stress incontinence at 18 representative hospitals in the United Kingdom [159]. A variety of measures of incontinence, symptom severity, symptom impact and complications were used, and participants were followed for 1 year. Overall, 87% of the women reported some improvement in incontinence one year after surgery, but only 28% reported complete continence [21, 160, 161]. This prospective cohort demonstrated that it is possible to collect standard data on multiple outcomes of surgery for stress incontinence to provide women with better information on the likely outcomes and effectiveness of the procedure in community practice. Lessons learned from this study will drive improvements in surgical research in incontinence as discussed below.

b) The importance of surgical randomized Controlled trials

Observational studies can also provide important

information for designing and selecting potential randomized clinical trials. The randomized controlled trial is the accepted "gold standard" for research of treatment effects. However, case series are far more common in the surgical literature, especially for new "innovative" surgical procedures. This is true despite the fact that case series cannot account for selection bias on the part of both the patient and surgeon, non-reporting bias of failures or loss to follow-up, lack of long-term follow-up, and provide the lowest level of evidence for treatment effects. In all surgical specialties, there has been growing concern regarding the limited number of randomized controlled trials for surgical procedures, poor methodological standards in those that have been performed, and a perception that surgeons are reluctant to rigorously test new surgical interventions [162-165]. A number of reasons for the paucity of surgical trials have been suggested including the lack of a regulatory board similar to the Food & Drug Administration responsible for the development of new medications [166]. Surgeons can therefore perform new procedures with little or no limitations from hospital or ethics committees and without any substantive trials [163]. In the United Kingdom, this role is filled by the National Institute of Clinical Excellence (NICE www.nice.org.uk) interventional procedures advisory committee, which analyzes all new procedures and delivers recommendations regarding their value.

The importance of surgical randomized controlled trials was recently demonstrated in a trial of arthroscopic surgery for osteoarthritis of the knee [167], one of the most widely used orthopedic operations. Numerous uncontrolled case series had reported substantial pain relief after arthroscopic surgery. The trial provided strong evidence that the surgical procedure was no better than the placebo procedure [167].

There have been few methodologically rigorous randomized trials of surgical procedures for incontinence [168]. In a recent systematic review of surgical procedures for incontinence, of 943 studies identified there were only 11 randomized controlled trials [21]. Overall, the randomized trials were considered of poor quality because they had few participants and were underpowered to detect small differences between groups, lacked blinding of the participants and/or individuals assessing the outcomes, and had short follow-up. It is particularly important that operations for incontinence undergo methodologically rigorous randomized trials because surgery is elective and non-emergent, the effect difference between two techniques will be at best modest, and patients as well as surgeons need accurate data to make informed choices using risk and benefit data to compare operations [169, 170]. It has also been suggested that the first anti-incontinence procedure provides the highest success rate and subsequent procedures have a far higher failure and complication rate [21].

The Urinary Incontinence Treatment Network is one example of a multi-center consortium created to conduct randomized controlled clinical trials enrolling patients with urinary incontinence. The UITN, established by the U.S. National Institutes of Health in 2000, consists of 9 recruiting centers and a data coordinating center. The clinical expertise includes a mixture of urology and urogynecology. The first clinical trial undertaken by the Network compares standardized forms of the Burch colposuspension (Tanagho modification) and the autologous rectus fascia sling procedures for overall treatment success and stress urinary incontinence success at 24-month post-operatively (the trial is known as the Stress Incontinence Surgical Treatment Efficacy Trial or SISTEr) [171]. Recruitment for this trial, with a target sample size of 650, was completed in June 2004. A second trial, the Behavior Enhances Drug Reduction of Incontinence, BE-DRI, was initiated in the summer of 2004. The purpose of this trial is to test whether the addition of behavior treatment to tolterodine therapy will increase the number of patients who can discontinue tolterodine therapy and sustain a significant reduction of incontinence. Participants are community-dwelling women with pure or urge predominant incontinence.

c) Surgical trial methods

"*Surgeons should realize that using the right tools for clinical research is comparable to selecting and using the right instruments for an operation.*" [172]

To ensure surgical trials are relevant and credible, detailed information about the study design is essential [162]. Reports of randomized trials should follow the current CONSORT flow diagram [30]. In order to understand how surgical results can be generalized to the population at large it is critically important that researchers carefully record the number of patients with incontinence who were not offered enrollment in the trial and those who refused to participate as well as the reasons for each. All participants should undergo a comprehensive baseline evaluation as discussed in the general recommendations and baseline comparability of the intervention groups demonstrated using descriptive statistics. The

randomization technique must be clearly described to confirm random allocation and that none of the study team has influenced the assignment resulting in selection bias.

Differential drop out after randomization can introduce bias. In the largest and most methodologically sound randomized controlled trial to date comparing the tension-free vaginal tape (TVT) and colposuspension (referred to as the UK TVT RCT), a large number of women withdrew from the colposuspension arm after randomization [173, 174]. The loss of participants after randomization introduced bias in favor of the TVT because the drop outs had less severe incontinence resulting in the colposuspension group having more severe incontinence. It has been suggested that participants were only willing to continue if they were randomized to the "new and better" TVT procedure [174, 175]. Accounting for subjects "lost to follow-up" must also be detailed as per the CONSORT flow diagram. In the UK TVT RCT, drop out after surgery was similar for both procedures. In contrast, the UITN study discussed above had no drop outs with 650 participants randomized in the operating room at the time of surgery [171].

The surgical procedure should be described in such detail that it could easily be reproduced in another study. Standardization of the procedure may vary depending on the research question [176]. Trials where the surgical technique is tightly controlled (i.e. small number of highly skilled surgeons) are analogous to medical trials where only compliant patients are randomized, reflecting efficacy of the procedure in an ideal setting. If the surgical procedure is less controlled, it may be more generalizable to a mixture of skill level among surgeons in the community, and so reflect effectiveness of the procedure in usual practice [169].

Masking of participants as to their assigned intervention and those assessing the outcome is particularly important for surgical trials for incontinence because there may be enthusiasm by the patient or surgeon for a new procedure, many outcomes are based on the patient's own assessments such as symptom and quality of life scores, and the intervention is primarily for improvement of symptoms [177]. In the previously mentioned UK TVT RCT, neither the participants nor the staff collecting the post-operative assessments were blinded and this may have resulted in a biased assessment of the outcome.

d) Outcomes of UI:

Surgical outcomes are discussed in detail in Chapters 6, 15, 16, and 17 as well as in the specific patient groups discussed above and will not be repeated here. The key issues are that validated outcome measures should be decided on in advance and data collected prospectively as well as throughout the study.

e) Development & assessment of new surgical frocedures

Surgical research presents unique challenges to efforts at optimizing patient care. It is important to create a pathway for real advances while simultaneously protecting patient safety. It would be desirable to have RCTs of all operations for incontinence; while one may argue that resources are inadequate it is also very costly to introduce ineffective or unsafe procedures without proper research. When new procedures are substantially different from prior operations there should be a broad based preliminary exploration leading to a comparative trial if warranted. At the same time, many minor modifications of surgical procedures are inappropriate for randomized trials and if required, surgical progress would be slowed [178].

It has been argued that the first patient in whom a procedure is performed should be randomized [163, 179]. Alternatively, it has been suggested that case series for new procedures are allowed until the procedure finds its intended use and to avoid doing studies while those performing the procedures are on the "learning curve". Typically, new surgical procedures for incontinence have been reported as case series [77, 180]. Not only do surgical case series provide the lowest level of evidence for treatment effects, case series may be "harmful". An accumulation of "positive" case series may present a premature certainty about benefits of a procedure and make it even more difficult to perform randomized trials [162, 169]. Influential members of the surgical community may endorse a new procedure and if the procedure is considered better it may be difficult to get surgeons and patients to randomize or a trial may appear to be unethical with a "proven" procedure [162, 163, 181].

For new surgical procedures, important issues of adequate informed consent and conflicts associated with incentives for developing, starting and using new procedures have been raised. Informed consent for a new procedure must include:

- acknowledgement that the procedure is new and has not been shown to be more effective than a traditional approach

- discussion of potential complications, especially any integrally related to the procedure or device
- disclosure that information on complications are limited, and
- disclosure that the long-term benefits are unclear [180].

Incentives for adopting new procedures prior to sufficient evidence can arise from self-interest by attracting patients to one's practice, industry marketing, and patient desire for "cutting edge" techniques. Industry sponsorship or a surgeon's financial interest must be disclosed.

It has been recently suggested that innovations in maternal-fetal surgery be conducted in centers of excellence, evaluated as research, and that randomized controlled trials are necessary before procedures become available outside the research setting or are integrated into clinical practice [182]. This recommendation is based on the premise that evidence is critical to ensuring that promising therapies are in fact safe and efficacious. [182]. Unfortunately, clinical trials and systematic analysis of outcomes have not preceded integration of new surgical therapies for incontinence into clinical practice [21, 168, 180, 183].

Fortunately, organizations and treatment networks have been established to address many issues related to surgical interventions. Examples include the UK National Institute of Clinical Excellence (NICE www.nice.org.uk), the Australian Safety and Efficacy Register of New Interventional Procedures- Surgical (ASERNIP-S www.surgeons.org/asernip-s), and the US treatment networks: Urinary Incontinence Treatment Network (UITN http://www. niddk. nih.gov/patient/uitn/uitn.htm) for the NIDDK and the Pelvic Floor Dysfunction Network (PFDN). The NICE and ASERNIP-S provide systematic reviews of new operations, assessment of effectiveness, and recommendations that the technique has sufficient data for widespread use, or that the techniques appear unsafe, or that further audit/research are required before its widespread usage. The UITN and PFDN were established to provide the infrastructure for multicenter large randomized controlled trials for incontinence and prolapse.

The PFDN is a multicenter network in the United States, supported by the National Institute of Child Health and Human Development, one of the institutes of the NIH. Started in 2001, the network has seven clinical sites and a data coordinating center, with the primary goal of performing clinical trials related to the prevention, evaluation, and treatment of pelvic floor disorders in women, including pelvic organ prolapse, and urinary and fecal incontinence as well as other abnormalities of the lower urinary and gastrointestinal tracts. The network has several large studies ongoing, including a surgical trial that will enroll 480 women to test whether the addition of Burch colposuspension prevents postoperative stress incontinence when continent women with advanced prolapse undergo abdominal sacrocolpopexy [184]; a cohort study of 900 primiparous women after their first birth, determining the prevalence and incidence of fecal and urinary incontinence among women who did and did not have an anal sphincter laceration at vaginal delivery, compared to women delivered by cesarean without labor; and comparing symptoms, physical examination, and imaging (pelvic magnetic resonance imaging and endoanal ultrasound) in 255 women after vaginal delivery with and without anal sphincter laceration and in women after cesarean delivery without labor. In addition, the network has completed several small studies so far, including studies comparing different types of catheters used in urodynamic testing, validating questionnaires on fecal incontinence and quality of life, and comparing modifications of the standardized system of staging prolapse.

To make ethical and evidence-based progress in surgical knowledge for incontinence, a new paradigm to balance surgical innovation and research is essential. Cooperative collaboration of investigators and possibly industry, a preliminary phase to develop new procedures and training, prospectively collected comprehensive data, and ongoing assessment as to the need for randomized controlled trials has been suggested [178]. As surgeons, we have the opportunity to improve our understanding of surgical interventions and improve patient care.

Recommendations for Surgical Trials:

- Safety and serious side effects of new operations must be completely defined with adequate follow-up so that risks can be weighed against efficacy. At a minimum, this requires more use of large scale, independent, prospective, multicenter cohort studies when RCTs are not practical. **HIGH**
- Valid informed research consent is required in all trials of surgical interventions, which is separate from the consent to surgery. **HIGH**

- We recommend ongoing research into the usefulness of pre- and post-operative urodynamics in surgical trials. One of the primary research goals should be to collect data to determine the predictive value of urodynamic testing prior to intervention for stress urinary incontinence. Other important areas include the utility and performance of urodynamics for continent women undergoing pelvic organ prolapse repair. **HIGH**
- Reports of successful treatment should be limited to subjects with a minimum (not mean) of one year follow-up and should include a patient perspective measure. Specific assumptions about subjects lost to follow-up should be stated; last observation carried forward may not be the most appropriate method of handling this data as most patients lost to follow-up should be considered to have failed treatment. **HIGH**

5. Summary of specific types of trials

It is appreciated that both device and surgical research trials present significant challenges to trial design. Although decades of experience have refined the conduct of drug trials the absence of comparable robust device and surgical trials derives from these design challenges. **Table 1** lists adherence to Design criteria between Behavioral, Drug, Device and Surgical Trials as specified in General Recommendations for Clinical Research in Incontinence. The net result of failure to adhere to these principles may be that the efficacies of the latter two approaches for urinary incontinence are over-estimated. By analogy to intervention trials for pain, the more invasive the treatment, the greater the placebo effect and desire by the patient to self report better outcomes. For urinary incontinence, like pain, this is especially relevant in that even purported "objective" outcomes of diary and pad tests rely on patient report and can be circumvented.

Table 1. Adherence to Design criteria between Drug, Device and Surgical Trials

Design Issue	Behavioral	Drugs	Devices	Surgery	Comments
Randomization	Sometimes	Often	Rare	Rare	Biases may not be eliminated in device or surgery if patient aware of treatment
Parallel	Sometimes	Often	Rare	Rare	
Crossover	Rare	Sometimes	Never	Never	Unable to un-do procedure
Placebo/sham	Rare	Often	Sometimes	Never	Considered unethical for sham surgery
Run-in period	Often	Often	Sometimes	Never	Including the 5-10% of subjects who would drop out for failure to meet inclusion criteria will favor device, surgery
Single institution	Often	Sometimes	Often	Usually	Reporting a surgeon's results could influence
Multi-institution	Sometimes	Often	Sometimes	Rare	
Single Blinded	Never	Sometimes	Never	Never	
Double Blinded	Never	Often	Never	Never	
Intention to treat	Rare	Often	Never	Never	If assume those lost to follow-up are failures, results when not included biased in favor of treatment
Per protocol	Rare	Rare	Often	Often	Efficacy at last follow-up fails to allow durability assessment
Follow-up	Months	Weeks	Months	Months	Durability questioned with behavioral and procedures
Equivalence trial	Rare	Rare	Rare	Rare	Devices/surgery enrolled after behavioral/drug therapies

V. ETHICAL ISSUES IN RESEARCH

1. The placebo in clinical trials of urinary incontinence

A placebo is defined as any treatment or aspect of treatment that does not possess a "specific" action on a patient's symptoms or disease. The placebo as an intervention is designed to simulate a medical treatment yet not be a specific therapy for the condition as judged by the investigator. Thus preclinical data demonstrating a potential mechanism of action is necessary to justify labeling one arm as "active". The use of a placebo control (or sham) is a critical aspect of predefined criteria for internal validity (quality assessment) for clinical trials developed by the US Preventive Services Task Force and National Health Service Centre for Reviews and Dissemination (UK) and for assessing clinical effectiveness based on Levels of Evidence (Levels Ia. Ib, IIa) used by the US Agency for Healthcare Research and Quality, UK National Health Service and other agencies.

The ability of RCTs to produce unbiased estimates of an intervention's innate (pharmacologic or physiologic) effect rests on at least two assumptions [185]. First, that disease severity and psychological effects are equally distributed across treatment groups. Second, that other effects fail to interact with the action of the intervention. Blinding both investigator and participant seeks to reduce preferences; preferences bias outcome assessments and enhance cognitive mechanisms that may interact with either control (placebo) or intervention (surgery, drug, device, behavioral modification). Urinary incontinence trials have the potential to be greatly influenced by the "placebo effect" because of the complex cognitive influences on bladder and outlet function and use of semi-objective outcome measures (questionnaires, diaries and pads). In this regard, incontinence trials resemble pain and psychiatric trials more that those with biochemical or physiological outcomes such as prostate specific antigen or bone density. To the skeptic, diaries and pads are outcomes that can be mis-reported and tampered with by the patient before arriving at the investigative site. Urodynamics, while more objective, fails to universally correlate with symptoms, is expensive and invasive as an outcome. Moreover, cognitive influences on urodynamics that require active participation are likely.

Controversy has surrounded the placebo effect since its description in the 1950's by H.K.Beecher [186]. An initial review of 15 studies found that symptoms were satisfactorily relieved in 35% of subjects leading to the promulgation of the estimate of placebo effect being about a third. The placebo response has been broadly defined as a change in the patient's health or bodily state that is attributable to the symbolic impact of medical treatment or setting [187]. In essence, a placebo response would be predicted whenever a conscious patient engages in any medical encounter. In a meta-analysis of 130 trials for 40 conditions no effect of placebo was seen for objective outcomes [188]. For more subjective outcomes a positive effect was noted. In contrast, for pain, psychiatric and incontinence trials the placebo arm varies from 8 to 78%. Placebos have time-effect curves and peak, cumulative and carryover effects similar to an active drug. In meta-analyses of placebo response [189], non specific effects are substantial and can be either synergistic or antagonistic toward pharmacologic or physiologic mechanisms. Therefore the additive model (drug effect minus placebo = specific effect) of trial design is too simplistic. If the informed consent process is misunderstood the subjects may believe that they are getting care directed at their problem. This therapeutic misconception could enhance the placebo effect and make it more difficult to detect a difference from active drug.

Two theories concerning the placebo effect – expectancy and conditioning – are relevant to incontinence trials. In examining the contributions of suggestion, desire and expectation on intrarectal drug therapy for irritable bowel syndrome it was found that both placebo and lidocaine had very large effects on visceral pain scores [190]. The authors found that a significant improvement with lidocaine over placebo when no suggestions for pain relief were given. But adding the suggestion of pain relief increased magnitude of placebo analgesia to nearly that of lidocaine. Patient-provider relationship also has a profound effect in addition to expectations. In a study evaluating acupuncture, it was found that excessive expectations had a negative association with outcome (Goal Attainment Score). This negative association is contrary to previous observations for placebo effect and was attributed to greater disappointment and these participants being less likely to perceive benefit than those using acupuncture as a last resort with lower expectations. Also contrary to previous reports, a strong relationship with the investigator had a negative association with outcome. This finding was attributed to the subject being more passive with respect to participation necessary in acupuncture. This latter finding may have implications for incontinence interventions requiring active parti-

cipation of subjects. These observations again highlight the notion that placebo effects can both attenuate and augment a proposed active therapy.

Expectancy, conditioning, patient-provider relations, suggestion and desire theories fail to explain how cognitive functioning influences symptoms. Recent findings from neuroanatomy and neurochemistry have shed insight into placebo mechanisms. Endorphin, catecholamine, cortisol and dopaminergic pathways have been implicated in placebo responses for pain and depression. The ability of naloxone to abolish the placebo effect induced by expectancy suggests an active opiate mechanism [191]. Thus in trials of centrally acting drugs for incontinence, differences over placebo may be minimized depending on the trial conditions. Involvement of descending prefrontal cortex and brainstem pathways acting on the dorsal horn of the spinal cord have been implicated based on differences in visceral versus somatic placebo effects. This observation may be relevant to neuromodulation trials in which PET scanning correlates a positive response of sacral stimulation with activation of these sites [192]. Because these same neurotransmitters and pathways influence micturition it is tempting to speculate that the large placebo effect in clinical trials for incontinence derives from similar mechanisms evoked in pain and psychiatric studies.

These findings have implications for trial design and interpretation for incontinence. For example, it is possible that the benefit of an intervention over placebo may be minimal yet the therapy effective. Alternatively, placebo effects can be exploited for therapeutic benefit. Also, the choice of placebo in behavioral or device trials is crucial in order not to overestimate the specificity of action of an intervention. For pelvic muscle training designing a sham maneuver that fails to contract musculature involved in urine storage yet activates muscles in the vicinity (e.g. adducting thighs while crossing ankles) is challenging. As for devices or surgery, sham surgery raises ethical issues. Because perception of a stimulus itself may have placebo effect it can be argued that in electrical or magnetic studies employing implantation with stimulus still overestimate efficacy. The issue of placebo controls versus an active or no control remains an important issue in incontinence research and argues for scrutiny of multiple controlled RCTs at various sites in different populations before universal adoption of a new therapy to better assess efficacy for an intervention.

2. Payment for recruiting in clinical research

Especially in the US, proceeds from clinical trials (primarily pharmaceutical but also surgical and device 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. It is typical and preferable that researchers do not receive money directly from industry sponsors but rather from a contracted research organization acting as an independent third party. Most quality peer-review scientific journals require a declaration of conflict of interest. There are many potential relationships between physicians and industry; it is preferable that the nature of the relationship and its financial magnitude if any, be fully defined rather than categorized (i.e. "consultant") so that the reader can appropriately assess the actual and potential conflict of interests.

3. Authorship, sponsorship and conflict of interest

In investigator-initiated, government-funded research, there has always been a lead investigator who is ultimately responsible for all aspects of the work. Recently, this paradigm has not been used in 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 are not 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; rarely are investigators involved in the analysis. This presents a real problem with favoritism and inevitably dilutes the force, impact, and responsibility of authorship. Standards of authorship defined by many journals must be followed. Academic leaders must establish standards for interactions between investigators and industry. All data, not just summary data, should be presented to all investigators. Prior to initiation of a pharmaceutical company

trial a publication committee should be established with a chair, representatives of principal investigators at participating sites, and a limited number of key representative of the sponsoring company. If the principal investigators are not adequately trained in statistical methods, independent statisticians should also be included on the Publications Committee.

RECOMMENDATIONS FOR ETHICAL RESEARCH:

- The committee members are particularly concerned about the perceived lack of input of principal investigators in the planning and reporting of clinical trials sponsored by pharmaceutical companies. Therefore, we recommend:
- Continuity in clinical direction from design through authorship is mandatory. Investigators should be involved in the planning stage and a publications committee should be named at the beginning of the clinical trial. The Uniform Requirements for Manuscripts Submitted to Biomedical Journals, from the International Committee of Medical Journal Editors should be followed. Authorship requires:
  1) Substantial contributions to conception and design or acquisition of data or analysis and interpretation of data,
  2) Drafting the article or revising it critically for important intellectual content,
  3) Final approval of the version to be published

  Authors should provide a description of what each contributed and editors should publish that information. **HIGH**
- Authors should have access to all raw data from clinical trials, not simply selected tables. **HIGH**
- Clinical trial results should be published regardless of outcome. In order to promote public awareness and ultimate publication we strongly recommend that all trials be registered in the public domain. The funder should have the right to review manuscripts for a limited period of time prior to publication but the manuscript is the intellectual property of its authors, not the funder. **HIGH**
- All authors should be able to accept responsibility for the published work and all potential conflicts of interest should be fully disclosed. **HIGH**
- All RCTs should be registered and journal editors should decline to publish studies that were not registered. **MEDIUM**

VI. CONCLUSIONS

The goal of this Consultation has been to examine and classify data in order to determine the level of evidence that supports our care of incontinent patients. The goal of this Committee has been to provide a roadmap for the investigators who will produce the high quality research for the next Consultation. Ultimately, good research is credible. Credibility creates impact and generates strong recommendations. Credible research draws others to follow and expand on the work while simultaneously guiding clinical care of patients. Unfortunately, it is clear that much of the published work in incontinence has not been of high quality and thus has not effectively changed patient care. However this can be remedied in the future for, in most cases, the failure has been due to preventable deficiencies in planning and data collection.

The Committee has emphasized that 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. If investigators will work together in true multidisciplinary teams, following the methodology presented here, then the Fourth International Consultation on Incontinence will truly be a landmark event.

REFERENCES

1. Abrams, P., et al., The standardization of terminology of lower urinary tract function. Scand J Urol Nephrol suppl, 1988. 144: p. 5-19.
2. Bates, P., et al., 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, Eur Urol 2:274-276, Scand J Urol Nephrol 11:193-196, Urol Int 32:81-87, 1976.
3. Bates, P., et al., 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, Br J Urol 49:207-210, Eur Urol 3:168-170, Scand J Urol Nephrol 11:197-199., 1977.
4. Bates, P., et al., 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., et al., Fourth report on the standardization of termino-

logy of lower urinary tract function. Terminology related to neuromuscular dysfunction of lower urinary tract. Br J Urol 52:333-335, Urology 17:618-620, Scand J Urol Nephrol 15:169-171, Acta Urol Jpn 27:1568-1571, 1981.

6. Abrams, P., et al., 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, Scand J Urol Nephrol 20:161-164, 1986.
7. Rowan, D., et al., Urodynamic Equipment: technical aspects. J Med Eng Technol, 1987: p. 11, 2, 57-64.
8. Andersen, J., et al., Lower Urinary Tract Rehabilitation Techniques: Seventh Report on the Standardization of Terminology of Lower Urinary Tract Function. Int Urogynecol J, 1992. 3: p. 75-80.
9. Bump, R., et al., The Standardization of Terminology of Female Pelvic Organ Prolapse and Pelvic Floor Dysfuntion. Am J Obstet Gynecol, 1996. 175: p. 10-17.
10. Griffiths, D., et al., Standardization of Terminology of Lower Urinary Tract Function: Pressure-Flow Studies of Voiding, Urethral Resistance, and Urethral Obstruction. Neurourol Urodyn, 1997. 16: p. 1-18.
11. Mattiasson, A., et al., 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, 1998. 17: p. 249-253.
12. Lose, G., et al., Outcome measures in adult women with symptoms of lower urinary tract dysfunction. Neurourol Urodyn, 1998. 17: p. 255-262.
13. Nordling, J., et al., Outcome measures for research in treatment of adult males with symptoms of lower urinary tract dysfunction. Neurourol Urodyn, 1998. 17: p. 263-271.
14. Fonda, D., et al., Outcome measures for research of lower urinary tract dysfunction in frail older people. Neurourol Urodyn, 1998. 17: p. 273-281.
15. Blaivas, J., et al., Standards of Efficacy of Treatment Outcomes in Urinary Incontinence: Recommendations of the Urodynamic Society. Neurourol Urodyn, 1997. 16: p. 145-147.
16. Blaivas, J., et al., Definition and Classification of Urinary Incontinence: Recommendations of the Urodynamic Society. Neurourol Urodyn, 1997. 16: p. 149-151.
17. Blaivas, J., Outcome measures for urinary incontinence. Urology, 1998. 51 (2A Suppl)(Feb): p. 11-19.
18. Spilker, B., Guide to Clinical Studies and Developing Protocols. 1984, New York: Raven Press.
19. Pocock, S., Clinical Trials: a Practical Approach. 1983, Chichester: Wiley.
20. Altman, D., Practical Statistics for Medical Research. 1991, London: Chapman & Hall.
21. Black, N., Why we need observational studies to evaluate effectiveness of health care. Br Med J, 1996. 312: p. 1215-1218.
22. Armitage, P. and G. Berry, Statistical Methods in Medical Research. 3rd ed. 1994, Oxford: Blackwell Science.
23. McAlister, F., et al., Analysis and reporting of factorial trials: a systematic review. JAMA, 2003. 289(19): p. 2545-1553.
24. Montgomery, A., T. Peters, and P. Little, Design, analysis and presentation of factorial randomised controlled trials. BMC Med Res Methodol, 2003. 3(1): p. 26.
25. Donner, A. and N. Klar, Design and Analysis of Cluster Randomization Trials in Health Research. 2000, London: Arnold.
26. Senn, S., Statistical Issues in Drug Development. 1997, Chirchester: Wiley.
27. International Conference on Harmonisation, w.s., Statistical Considerations in the Design of Clinical Trials. May 29, 2001, http://www.ifpma.org/pdfifpma/e9.pdf.
28. Armitage, P., G. Berry, and J. Matthews, Statistical Methods in Medical Research. 4th ed. 2002, Oxford: Blackwell Science.
29. Simon, R. and P. Thall, Phase II trials., in Ecyclopedia of Biostatistics., P. Armitage and T. Colton, Editors. 1998, Wiley: Chichester. p. 3370-3376.
30. CONSORT, w.s., Statistical Considerations in the Design of Clinical Tials. May 29, 2001, http://www.consort-statement.org.
31. Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. JAMA, 1994. 242: p. 1926-1931.
32. Altman, D., Better reporting of randomised controlled trials: the CONSORT statement. BMJ, 1996. 313: p. 570-571.
33. Altman, D., et al., The revised CONSORT statement for reporting randomised trials: explanation and elaboration. Annals of Internal Medicine, 2001. 134: p. 663-694.
34. Egger, M., P. Juni, and C. Bartlett, Value of flow diagrams in reports of randomized controlled trials. JAMA, 2001. 285(15): p. 1996-1999.
35. Moher, D., A. Jones, and L. Lepage, Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA, 2001. 285: p. 1992-1995.
36. Rennie, D., CONSORT revised - improving the reporting of randomized trials. [Editorial] JAMA, 2001. 285: p. 2006-2007.
37. Elbourne, D. and M. Campbell, Extending the CONSORT statement to cluster randomised trials: for discussion. Statistics in Medicine, 2001. 20: p. 489-496.
38. STARD, The STARD Initiative—Towards Complete and Accurate Reporting of Studies on Diagnostic Accuracy. 2004.
39. Peters, T. and J. Eachus, Achieving equal probability for selection under various random sampling strategies. Paediatric and Perinatal Epidemiology, 1995. 9: p. 219-224.
40. Moher, D., et al., The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. JAMA, 2001. 285: p. 1987-1991.
41. Donovan, J., et al., Improving design and conduct of randomised trials by embedding them in qualitative research: ProtecT (prostate testing for cancer and treatment) study. BMJ, 2002. 325: p. 766-770.
42. Featherstone, K. and J. Donovan, Random allocation or allocation at random? Patients' perspectives of participation in a randomised controlled trial. BMJ, 2000. 317: p. 1177-1180.
43. Grant, A., et al., Issues in data monitoring and interim analysis of trials. Health Technology Assessment. in press.
44. Sydes, M., et al., Reported use of data monitoring committees in the main published reports of randomised controlled trials: a cross-sectional study. Clinical Trials, 2004. 1: p. 48-59.
45. Sydes, M., et al., Systematic review of data monitoring committees in randomised controlled trials: a cross-sectional study. Clinical Trials, 2004. 1: p. 60-79.
46. Schulz, K., I. Chalmers, and D. Altman, The Landscape and Lexicon of Blinding in Ranodmized Trials. Annals of Internal Medicine, 2002. 136(3): p. 254-259.
47. Campbell, M., S. Julious, and D. Altman, Estimating sample size for binary, ordered catgorical, and continuous outcomes in two group comparisons. Br Med J, 1995. 311: p. 1145-1148.
48. Matthews, J., An Introduction to Randomized Controlled Clinical Trials. 2000, London: Arnold.
49. Machin, D., et al., Sample Size Tables for Clinical Studies. 2nd ed. 1997, Oxford: Blackwell Science.
50. Sterne, J. and G. Davey Smith, Sifting the evidence - what's wrong with significance tests? BMJ, 2001. 322: p. 226-231.

51. Schwartz, D. and J. Lellouch, Explanatory and pragmatic attitudes in therapeutic trials. Journal of Chronic Diseases, 1967. 20: p. 637-648.
52. Newell, D., Intention-to-treat analysis: implications for quantitative and qualitative research. Int J Epidemiology, 1992. 21: p. 837-841.
53. Peters, T., H. Wildschut, and C. Weiner, Epidemiologic considerations in screening. In: Wildschut, HIJ; Weiner, CP; Peters, TJ; eds. When to Screen in Obstetrics and Gynecology. 1996, London: WB Saunders.
54. Hollis, S. and M. Campbell, What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ, 1999. 319: p. 670-674.
55. Lewis, J. and D. Machin, Intention to treat - who should use ITT? [Editorial] British Journal of Cancer, 1993. 68: p. 647-650.
56. Richards, S., et al., 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: p. 91-98.
57. Bland, M., An Introduction to Medical Statistics. 2nd ed. 1995, Oxford: Oxford University Press.
58. Sterne, J., Commentary: Null points—has interpretation of significance tests improved? Int J Epidemiology, 2003. 32(5): p. 693-694.
59. Kirkwood, B. and J. Sterne, Essential Medical Statistics. 2nd ed. 2003, Oxford: Blackwell Science.
60. Collett, D., Modelling Binary Data. 1991, London: Chapman & Hall.
61. Collett, D., Modelling Survival Data in Medical Research. 1994, London: Chapman & Hall.
62. Matthews, J., et al., Analysis of serial measurements in medical research. BMJ, 1990. 300: p. 230-235.
63. Zar, J., Biostatistical Analysis. 2nd ed. 1984, New Jersey: Prentice-Hall.
64. Brookes, S., et al., Subgroup analyses in randomised controlled trials: quantifying the risks of false positives and negatives. Health Technology Assesment, 2001. 5: p. 33.
65. Brookes, S., et al., Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol, in press.
66. Brookes, S., et al., Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol, 2004. 57(3): p. 229-236.
67. Moher, D., A. Jones, and L. Lepage, Use of the CONSORT statement: revised recommendations for improving the quality of reports of parallel-group trials. Lancet, 2001. 357(9263): p. 1191-1194.
68. Moher, D., et al., Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. Lancet, 1999. 354: p. 1896-1900.
69. Peters, T., et al., Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. Int J Epidemiology, 2003. 32(5): p. 840-846.
70. Berry, H., et al., Expectation and patient preferences-does it matter? JR Soc Med, 1980. 73: p. 34-38.
71. Weber, A., et al., The standardization of terminology for researchers in female pelvic floor disorders. Int Urogynecol J Pelvic Floor Dysfunct, 2001. 12(3): p. 178-186.
72. Ryhammer, A., et al., 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): p. 800-803.
73. Ryhammer, A., J. Djurhuus, and S. Laurberg, Pad testing in incontinent women: a review. Int Urogynecol J Pelvic Floor Dysfunct, 1999. 10(2): p. 111-115.
74. Sandvik, H., et al., A severity index for epidemiological surveys of female urinary incontinence: comparison with 48 hour pad weighing tests. Neurourol Urodyn, 2000. 19(2): p. 127-145.
75. Groutz, A., et al., Noninvasive outcome measures of urinary incontinence and lower urinary tract symptoms: a multicenter study of micturition diary and pad tests. Journal of Urology, 2000. 164(3 Pt 1): p. 698-701.
76. Brown, J.S., et al., Measurement characteristics of a voiding diary for use by men and women with overactive bladder. Urology, 2003. 61(4): p. 802-9.
77. Nygaard, I. and R. Holcomb, Reproducibility of the seven-day voiding diary in women with stress urinary incontinence. Int Urogynecol J Pelvic Floor Dysfunct, 2000. 11(1): p. 15-17.
78. Shaw, C., et al., Validity and reliability of a questionaire to measure the impact of lower urinary tract symptoms on quality of life: the Leicester Impact Scale. Neurourol Urodyn, 2004. 23(3): p. 229-236.
79. Kobelt, Economic considerations and outcome measurement in urge incontinence. Urology, 1997. 50(6A Suppl): p. 100-107; discussion 108-110.
80. Kielhorn, A. and J.-M. Graf von der Schulenburg, The Health Economics Handbook. 2nd ed. 2000, Chester: Adis International.
81. Fusco, F., et al., Videourodynamic studies in men with lower urinary tract symptoms: A comparison of community based versus referral urological practices. J Urol, 2001. 166: p. 910.
82. Grady, D., et al., Postmenopausal hormones and incontinence: the Heart and Estrogen/Progestin Replacement Study. Obstet Gynecol, 2001. 97(1): p. 116-120.
83. Jackson, S., et al., The effect of oestrogen supplementation on post-menopausal urinary stress incontinence: a double-blind pacebo-controlled trial. Br J Obstet Gynaecol, 1999. 106(7): p. 71-718.
84. Grodstein reference, it should be :Grodstein, F., K. Lifford, et al. (2004). "Postmenopausal hormone therapy and risk of developing urinary incontinence." Obstet Gynecol 103(2): 254-60.
85. Shumaker, S., et al., Health-related quality of life measurements for women with urinary incontinence: The urogenital distress inventory and the incontinence impact questionaire. Quality of Life Research, 1994. 3: p. 291-306.
86. Erickson and S. Raz.
87. Groutz, A., J. Blaivas, and J. Rosenthal, A Simplified urinary incontinence score for the evaluation of treatment outcomes. Neurourol Urodyn, 2000. 19(2): p. 127-135.
88. Rogers, R., et al., A new instrument to measure sexual function in women with urinary incontinence or pelvic organ prolapse. Am J Obstet Gynecol, 2001. 184(4): p. 552-558.
89. Yalcin, I., et al., Validation of a clinical algorithm to diagnose stress urinary incontinence for large studies. J Urol, 2004. 171(6 Pt 1): p. 2321-2325.
90. Buchner, D. and E. Wagner, Preventing frail health. Clin Geriatr Med, 1992. 8: p. 1-17.
91. Fultz, N.H., et al., Prevalence and severity of urinary incontinence in older African American and Caucasian women. J Geront, 1999. 54(6): p. M299-303.
92. Thom, D., Variation in estimates of urinary incontinence prevalence in the community: effects of differences in definition, population characteristics, and study type. J Am Geriatr Soc, 1998. 46(4): p. 473-80.
93. Wetle, T., et al., Difficulty with holding urine among older persons in a geographically defined community: prevalence and correlates. J Am Geriatr Soc, 1995. 43(4): p. 349-55.

94. Thom, D.H. and J.S. Brown, Reproductive and hormonal risk factors for urinary incontinence in later life: a review of the clinical and epidemiologic literature. J Am Geriatr Soc, 1998. 46(11): p. 1411-7.
95. Foldspang, A., et al., Parity as a correlate of adult female urinary incontinence prevalence. J Epidemiol Community Health, 1992. 46(6): p. 595-600.
96. Brown, J., et al., Urinary incontinence in older women: who is at risk? Obstet Gynecol, 1996. 87(1): p. 715-21.
97. Brown, J., et al., Prevalence of urinary incontinence and associated risk factors in postmenopausal women. Heart & Estrogen/Progestin Replacement Study (HERS) Research Group. Obstet Gynecol, 1999. 94(1): p. 66-70.
98. Diokno, A.C., et al., Medical correlates of urinary incontinence in the elderly. Urology, 1990. 36(2): p. 129-38.
99. Brown, J.S., et al., Hysterectomy and urinary incontinence: a systematic review. Lancet, 2000. 356(9229): p. 535-9.
100. Marshall, H.J. and D.G. Beevers, Alpha-adrenoceptor blocking drugs and female urinary incontinence: prevalence and reversibility. Br J Clin Pharmacol, 1996. 42: p. 507-509.
101. Diokno, A., M. Brown, and A. Herzog, Relationship between use of diuretics and coninence status in the elderly. Urology, 1991. 38: p. 39-42.
102. Fantl, J., et al., Diuretics and urinary incontinence in community-dwelling women. Neurourol Urodyn, 1990: p. 25-34.
103. Menefee, S.A., R. Chesson, and L.L. Wall, Stress urinary incontinence due to prescription medications: alpha-blockers and angiotension converting enzyme inhibitors. Obstet Gynecol, 1998. 91: p. 853-853.
104. Ouslander, J.G., et al., Urinary incontinence in nursing homes: incidence, remission and associated factors. Journal of the American Geriatrics Society, 1993. 41(10): p. 1083-9.
105. Katz, S., et al., The index of ADL: A standardized measurement of biological and psychological function. JAMA, 1963. 185: p. 914-919.
106. Mahoney, F. and D. Barthel, Funcional Evaluation: The Barthel Index. Maryland State Med J, 1965. 14: p. 61-65.
107. Folstein, M., S. Folstein, and P. McHugh, Mini Mental State: A practical method for grading the congnitive state of patients for the clinician. J Psych Res, 1975. 12: p. 189-198.
108. Inouye, S.K., et al., Clarifying confusion: the confusion assessment method. A new method for detection of delirium. Ann Intern Med, 1990. 113(12): p. 941-8.
109. Buschke, H. and P.A. Fuld, Evaluating storage, retention, and retrieval in disordered memory and learning. Neurology, 1974. 24(11): p. 1019-25.
110. Wechsler, D., Wechsler Adult Intelligence Scale-Revised. 1988, New York: Psychological Corp.
111. Reitan, R., Validity of the Trail making Test as an indicator of organic brain damage. Perceptual & Motor Skills, 1958. 8: p. 271-276.
112. Homma, Y., et al., Voiding and incontinence frequencies: Variability of diary data and required diary length. Neurourol Urodyn, 2002. 21(3): p. 204-9.
113. Locher, J.L., et al., Reliability assessment of the bladder diary for urinary incontinence in older women. J Gerontol A Biol Sci Med Sci, 2001. 56(1): p. M32-5.
114. Colling, J., et al. Continence Program for Care Dependent Elderly (Final Report). in NIH-NCNR-NR 01554. 1995. Bethesda, MD.
115. Siltberg, H., A. Victor, and G. Larsson, Pad weighing tests: the best way to quantify urine loss in patients with incontinence. Acta Obstet Gynecol Scand Suppl, 1997. 166(1): p. 28-32.
116. Griffiths, D.J., P.N. McCracken, and G.M. Harrison, Incontinence in the elderly: objective demonstration and quantitative assessment. Br J Urol, 1991. 67(5): p. 467-71.
117. Fonda, D., et al., Outcome measures for research of lower urinary tract dysfunction in frail older people. Neurourol Urodyn, 1998. 17(3): p. 273-81.
118. Holtedahl, K., et al., Usefulness of urodynamic examination in female urinary incontinence - Lessons from a population-based, randomized, controlled study of conservative treatment. Scandinavian Journal of Urology & Nephrology, 2000. 34(3): p. 169-174.
119. Goode, P.S., et al., Measurement of postvoid residual urine with portable transabdominal bladder ultrasound scanner and urethral catheterization. Int Urogynecol J Pelvic Floor Dysfunct, 2000. 11(5): p. 296-300.
120. Alnaif, B. and H.P. Drutz, The accuracy of portable abdominal ultrasound equipment in measuring postvoid residual volume. Int Urogynecol J Pelvic Floor Dysfunct, 1999. 10(4): p. 215-8.
121. Bent, A.E., D.E. Nahhas, and M.T. McLennan, Portable ultrasound determination of urinary residual volume. Int Urogynecol J Pelvic Floor Dysfunct, 1997. 8(4): p. 200-2.
122. Grosshans, C., Y. Passadori, and B. Peter, Urinary retention in the elderly: a study of 100 hospitalized patients. J Am Geriatr Soc, 1993. 41(6): p. 633-8.
123. Ouslander, J.G., et al., Use of a portable ultrasound device to measure post-void residual volume among incontinent nursing home residents. J Am Geriatr Soc, 1994. 42(11): p. 1189-92.
124. Health, N.I.o., NIH POLICY AND GUIDELINES ON THE INCLUSION OF CHILDREN AS PARTICIPANTS IN RESEARCH INVOLVING HUMAN SUBJECTS. March 6, 1998, http://grants.nih.gov/grants/guide/notice-files/not98-024.htmlNational.
125. Farhat, W., et al., The dysfunctional voiding scoring system: quantitative standardization of dysfunctional voiding symptoms in children. J Urol, 2000. 164: p. 1011-1015.
126. Sureshkumar, P., et al., A reproducible pediatric daytime urinary incontinence questionaire. J Urol, 2001. 165(569-573).
127. Campbell, M.F., et al., eds. Campbell's Urology. Vol. 8. 2004, WB Saunders Company: Philadelphia.
128. McGuire, E., Myelodysplasia references.
129. Lacima, G. and M. Pera, Combined fecal and urinary incontinence: an update. Curr Opin Obstet Gynecol, 2003. 15(5): p. 405-410.
130. Baxter, N., D. Rothenberger, and A. Lowry, Measuring fecal incontinence. Dis Colon Rectum, 2003. 46(12): p. 1591-1605.
131. Jorge, J. and S. Wexner, Etiology and management of fecal incontinence. Dis Colon Rectum, 1993. 36(1): p. 77-97.
132. Rockwood, T., et al., Patient and surgeon ranking of the severity of symptoms associated with fecal incontinence: the fecal incontinence severity index. Dis Colon Rectum, 1999. 42(12): p. 1525-1532.
133. Vaizey, C., et al., Prospective comparison of fecal incontinence grading systems. Gut., 1999. 44(1): p. 77-80.
134. Rockwood, T., et al., Fecal Incontinence Quality of Life Scale: quality of life instrument for patients with fecal incontinence. Dis Colon Rectum, 2000. 43(1): p. 9-16; discussion 16-17.
135. Bugg, G., E. Kiff, and G. Hosker, A new condition-specific health-related quality of life questionnaire for the assessment of women with anal incontinence. BJOG, 2001. 108: p. 1057-1067.
136. Curhan, G., et al., Epidemiology of interstitial cystitis: a population based study. J Urol, 1999. 161(2): p. 549-552.
137. Gillenwater, J. and A. Wein, Summary of the National Institute of Arthritis, Diabetes, Digestive and Kidney Diseases Workshop

on Interstitial Cystitis, National Institutes of Health, Bethesda, Maryland, August 28-29, 1987. J Urol, 1988. 140(1): p. 203-206.

138. Hanno, P., et al., The diagnosis of interstitial cystitis revisited: lessons learned from the National Institutes of Health Interstitial Cystitis Database study. J Urol, 1999. 161(2): p. 553-557.
139. Propert, K., et al., Pitfalls in the disign of clinical trials for interstitial cystitis. Urology, 2002. 60(5): p. 742-748.
140. Abrams, P., et al., The standardization of terminology of lower urinary tract function: report from the Standardisation Sub-committee of the International Continence Society. Neurourol Urodyn, 2002. 21(2): p. 167-178.
141. Payne, C., A. Terai, and K. Komatsu, Research criteria versus clinical criteria for interstitial cystitis. Int J Urol, 2003. 10 Suppl: p. S7-S10.
142. Sant, G., et al., A pilot clinical trial of oral pentosan polysulfate and oral hydroxyzine in patients with interstiatial cystitis. J Urol, 2003. 170(3): p. 810-815.
143. Mayer, R., et al., submitted for publication, 2004, in press.
144. Ellerkmann, R., et al., Correlation of symptoms with location and severity of pelvic organ prolapse. Am J Obstet Gynecol, 2001. 185(6): p. 1332-1337; discussion 1337-1338.
145. Mouritsen, L. and J. Larsen, Symptoms, bother and POPQ in women referred with pelvic organ prolapse. Int Urogynecol J Pelvic Floor Dysfunct, 2003. 14(2): p. 122-127.
146. Barber, M., et al., Psychometric evaluation of two comprehensive condition-specific quality of life instruments for women with pelvic floor disorders. in press.
147. Samuelsson, E., et al., Signs of genital prolapse in a Sweedish population of women 20 to 59 years of age and possible related factors. Am J Obstet Gynecol, 1999. 180(2 Pt 1): p. 299-305.
148. Bland, D., et al., Use of the Pelvic Organ Prolapse staging system of the International Continence Society, American Urogynecologic Society, and Society of Gynecologic Surgeons in perimenopausal women. Am J Obstet Gynecol, 1999. 181(6): p. 1324-1327; discussion 1327-1328.
149. Swift, S., The distribution of pelvic organ support in a population of female subjects seen for routine gynecologic health care. Am J Obstet Gynecol, 2000. 183(2): p. 277-285.
150. Risk factors for genital prolapse in non-hysterectomized women around menopause. Results from a large cross-sectional study in menopausal clinics in Italy. Progetto Menopausa Italia Study Group. Eur J Obstet Gynecol Reprod Biol, 2000. 93(2): p. 135-140.
151. Lose, G., et al., Efficacy of desmopressin (Minirin) in the treatment of nocturia: a double-blind placebo-controlled study in women. Am J Obstet Gynecol, 2003. 189(4): p. 1106-1113.
152. van Kerrebroeck, P., et al., The stadardization of terminology in nocturia: report from the standardization subcommittee of the International Continence Society. BJU Int., 2002. 90 Suppl 3: p. 11-15.
153. Goode, P., et al., Effect of behavioral training with or without pelvic floor electrical stimulation on stress incontinence in women: a randomized controlled trial. JAMA, 2003. 290(3): p. 345-352.
154. Burgio, K., et al., Behavioral training with and without biofeedback in the treatment of urge incontinence in older women: a randomized controlled trial. JAMA, 2002. 288(18): p. 2293-2299.
155. Health, U.S.F.a.D.A.-C.f.D.a.R., Guidance Documents and Reports. Draft guidance for preclinical and clinical investigations of urethral bulking agents used in the treatment of urinary incontinence. November 29, 1995, http: //www.fda.gov/cdrh/ode/ oderp850.html.
156. Waetjen, L., et al., Stress urinary incontinence surgery in the United States. Obstet Gynecol, 2003. 101(4): p. 671-676.
157. Korn, A. and L. Learman, Operations for stress urinary incontinence in the United States, 1988-1992. Urology, 1996. 48(4): p. 609-612.
158. Diokno, A., et al., Prevalence and outcomes of continence surgery in community dwelling women. J Urol, 2003. 170(2 Pt 1): p. 507-511.
159. Hutchings, A. and N. Black, Surgery for stress incontinence: a non-randomized trial of colposuspension, needle suspension and anterior colporrhaphy. Eur Urol, 2001. 39(4): p. 375-382.
160. Black, N., et al., Impact of surgery for stress incontinence on morbidity: cohort study. BMJ, 1997. 315(7121): p. 1493-1498.
161. Black, N., et al., Impact of surgery for stress incontinence on the social lives of women. Br J Obstet Gynaecol, 1998. 105(6): p. 605-612.
162. Hall, J., et al., Methodologic standards in surgical trials. Surgery, 1996. 119(4): p. 466-472.
163. McLeod, R., et al., Randomized controlled trials in surgery: Issues and problems. Surgery, 1996. 119(5): p. 483-486.
164. Solomon, M., et al., Randomized controlled trials in surgery. Surgery, 1994. 115(6): p. 707-712.
165. Solomon, M. and R. McLeod, Should we be performing more randomized controlled trials evaluating surgical operations? Surgery, 1995. 118(3): p. 459-467.
166. Grimes, D., Technology follies. The uncritical acceptance of medical innovation. JAMA, 1993. 269(23): p. 3030-3033.
167. Moseley, J., et al., A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med, 2002. 347(2): p. 81-88.
168. Leach, G., et al., Female Stress Urinary Incontinence Clinical Guidelines Panel summary report on surgical management of female stress urinary incontinence. The American Urological Association. J Urol, 1997. 158(3 Pt 1): p. 875-880.
169. Baum, M., Reflections on randomized controlled trials in surgery. Lancet, 1999. 353 Suppl 1: p. S16-S18.
170. Lustig, A. and P. Scardino, Elective Patients. Surgical Ethics, ed. L. McCullough, J. Jones, and B. Brody. 1998, Houston, TX: Oxford University Press. 133-151.
171. Tennstedt, S., et. al., Design of teh SISTEr (Stress Incontinence Surgical Treatment Efficacy Trial) Study: A Randomized Surgical Trial Comparing the Burch Colposuspension and the Autologous Rectal Fascial Sling. Am J Obstet Gynecol, in press.
172. Madhok, R. and H. Handoll, Randomised trials in surgery. Integrated approach is needed. BMJ, 2002(325(7365)): p. 658.
173. Ward, K., P. Hilton, and U.K.a.I.T.-f.V.T.T. Group., Prospective multicentre randomised trial of tension-free vaginal tape and colosuspension as primary treatment for stress incontinence. BMJ, 2002. 325(7355): p. 67.
174. Hilton, P., Trials of surgery for stress incontinence—thoughts on the 'Humpty Dumpty principle'. BJOG, 2002. 109(10): p. 1081-1088.
175. Maddern, G., P. Middleton, and A. Grant, Urinary stress incontinence. BMJ, 2002. 325(7368): p. 789-790.
176. McLeod, R., Issues in surgical randomized controlled trials. World J Surg, 1999. 23(12): p. 1210-1214.
177. Johnson, A. and J. Dixon, Removing bias in surgical trials. BMJ, 1997. 314(7085): p. 916-917.
178. McCulloch, P., et al., Randomised trials in surgery: problems and possible solutions. BMJ, 2002. 324(7351): p. 1448-1451.
179. Chalmers, T., Randomization of the first patient. Med Clin North Am, 1975. 59(4): p. 1035-1038.
180. Gates, E., Ethical considerations in the incorporations of new

technologies into gynecologic practice. Clin Obstet Gynecol, 2000. 43(3): p. 540-550.

181. Frader, J. and D. Caniano, Research and Inovation in Surgery. Surgical Ethics, ed. L. McCullough, J. Jones, and B. Brody. 1998, Houston, TX: Oxford University Press. 216-241.
182. Lyerly, A., et al., Toward the ethical evaluation and use of maternal-fetal surgery. Obstet Gynecol, 2001. 98(4): p. 689-697.
183. Merlin, T., et al., A systmatic review of tension-free urthropexy for stress urinary incontinence: intravaginal slingplasty and the tension-free vaginal tape procedures. BJU Int., 2001. 88(9): p. 871-880.
184. Brubaker, L., et al., A randomized trial of Colpopexy and Urinary Reduction Efforts (CARE): Design and methods. Controlled Clinical Trials, 2003. 24: p. 629-642.
185. Helpern, S., Evaluating preference effects in partially unblended, randomized clinical trials. J Clin Epidemiol, 2003. 56: p. 109-115.
186. Turner, J., et al., The importance of placebo effects in pain treatment and research. JAMA, 1994. 271(20): p. 1609-1614.
187. Howard, B., The placebo response: Recent research and implications for Family Medicine. J Fam Pract, 2000. 49: p. 649-654.
188. Hrobjartsson, A. and P. Gotzsche, Placebo treatment versus no treatment. Cochrane Database of Systematic Reviews., 2003. 1:CD003974.
189. Kienle, G. and H. Kiene, Placebo effect and placebo concept: a critical methodological and conceptual analysis of reports on the magnitude of the placebo effect. Altern Ther Health Med, 1996. 2(6): p. 39-54.
190. Vase, L., et al., The contributions of suggestion, desire, and expectation to placebo effects in irritable bowel syndrome patients. An empirical investigation. Pain, 2003. 105(1-2): p. 17-25.
191. Amanzio, M. and F. Benedetti, Neuroparmacological dissection of placebo analgesia: expectation-activated opioid systems versus conditioning-activated specific subsystems. J Neurosci, 1999. 19(1): p. 484-494.
192. Blok, B., et al., Brain activation and urodynamics during sacral neuromodulation in urge incontinence: A PET study. J Urol, 2002. 167: p. 273.
193. Abrams P, et. al.: The standardisation of terminology of lower urinary tract function: report from the Standardisation Sub-committee of the International Continence Society. Neurourol Urodyn. 2002;21(2):167-78.