

How to Review and Understand a Research Article: Part I

Rebecca G. Stephenson PT,*
Nancy C. Rich, PhD, PT,
FACSM**

*Stephenson Physical Therapy
Medfield, Massachusetts
**Department of Physical
Therapy, University of New
England, Biddeford, Maine

INTRODUCTION

As our profession examines and explores treatment alternatives, the clinician must be able to read the research literature and decide if the outcomes have any relevance for their clinical practice. Frequently the journal articles appear intimidating and a quick scan of the abstract and conclusion will not tell the full story or get to the level of sophistication that will advance our practice.

Our goal is to teach the process of reading and understanding a current paper in the field of women's health. In this article we will explain some terminology found in research articles, define the characteristics of a well-designed study, critique a published article, and hopefully, inspire therapists to scrutinize journals.

This 2-part series will look critically at an article that was published in the *British Medical Journal* Feb. 20, 1999. The complete original article appears here in italics with our comments in regular font. The article is reprinted with permission from BMJ Publishing Group. *British Medical Journal*. London: 1999;318:487-493. Copyright 1999, BMJ Publishing Group.

Research articles are divided into 6 parts: abstract, introduction, methods, results, discussion, and conclusion. For our tutorial, Part I will include the title, abstract, introduction, and methods and Part II, in the next issue will cover the results, discussion, conclusion, and as an extra, a commentary by other researchers with a response back by Dr. Kari Bo.

Definitions follow underlined words so that the reader can understand the research language. Please note that some terms in the article are spelled differently from Standard English, and we have left them as they appear in the original text.

THE TITLE

The title is what hooks the reader to the article and should use terms that will index the article in databases. A title can be descriptive, describing the research or assertive, which summarizes the results in the title or lets the reader draw their own conclusion.¹

Single blind, randomised controlled trial of pelvic floor exercises, electrical stimulation,

vaginal cones, and no treatment in management of genuine stress incontinence in women

Bo, Kari; Talseth, Trygve; Holme, Ingar
Norwegian Centre for Physiotherapy Research and
Norwegian University of Sport and Physical Education,
PO Box 4014, Ullevål Stadion, 0806 Oslo, Norway
Kari Bo, professor

National Hospital of Norway, Oslo
Trygve Talseth, consultant urologist

Norwegian University of Sport and Physical
Education, Oslo
Ingar Holme, professor of biostatistics

Correspondence to: Professor Bo
karib@brage.idrettsbs.no

This is a descriptive title telling us what kind of study it is.

1. Single blinded which means that either the participants or the investigator does not know which intervention they are receiving.
2. This study involves 1 independent variable with 4 levels (the independent variable being "treatment"): (1) pelvic floor exercises, (2) electrical stimulation, (3) vaginal cones, and (4) no treatment. The treatment variable is being manipulated by the researchers to investigate the effect on the dependent variable, which here is genuine stress incontinence.
3. The sample included in the study is drawn from the population of only women. In each of the groups the number of participants included in the sample is indicated with the letter n. In statistical terms, population refers to all the possible members in a defined group of interest. Since it is obviously quite impossible to include all women in the world with a diagnosis of stress incontinence in one investigation, the researchers are going to use a subgroup, or a representative sample, of the population. The results of the data collected from the sample will be used to make generalizations about the entire population.
4. Anyone searching using a data base could use key words from the title and this article should come up. The author Kari Bo is a professor at the Norwegian Centre for Physiotherapy Research and Norwegian University of Sport and Physical Education. Trygve

Talseth is a consultant urologist at the National Hospital of Norway in Oslo. Ingar Home is a professor of biostatistics at the Norwegian University of Sport and Physical Education in Oslo Norway.

THE ABSTRACT

The abstract always comes at the beginning of an article and is often the only part that is available from on-line searches. The abstract may be limited to less than 250 words and summarizes the purpose of the study, procedures, results, and major conclusions. References and statistical information is not given here as this is relevant in the context of the entire study. Often major indexing terms are used in the abstract as readers can search databases by title or abstract.

Abstract

Objective: To compare the effect of pelvic floor exercises, electrical stimulation, vaginal cones, and no treatment for genuine stress incontinence.

Design: Stratified, single blind, randomised controlled trial.

Setting: Multicentre.

Participants: 107 women with clinically and urodynamically proved genuine stress incontinence. Mean (range) age was 49.5 (24-70) years, and mean (range) duration of symptoms 10.8 (1-45) years.

Interventions Pelvic floor exercise ($n = 25$) comprised 8-12 contractions 3 times a day and exercise in groups with skilled physical therapists once a week. The electrical stimulation group ($n = 25$) used vaginal intermittent stimulation with the MS 106 Twin at 50 Hz 30 minutes a day. The vaginal cones group ($n = 27$) used cones for 20 minutes a day. The untreated control group ($n = 30$) was offered the use of a continence guard. Muscle strength was measured by vaginal squeeze pressure once a month.

Main outcome measures: Pad test with standardised bladder volume and self report of severity.

Results: Improvement in muscle strength was significantly greater ($P = 0.03$) after pelvic floor exercises (11.0 cm H₂O (95% confidence interval 7.7 to 14.3) before v 19.2 cm H₂O (15.3 to 23.1) after) than either electrical stimulation (14.8 cm H₂O (10.9 to 18.7) v 18.6 cm H₂O (13.3 to 23.9)) or vaginal cones (11.8 cm H₂O (8.5 to 15.1) v 15.4 cm H₂O (11.1 to 19.7)). Reduction in leakage on pad test was greater in the exercise group (-30.2 g; -43.3 to 16.9) than in the electrical stimulation group (-7.4 g; -20.9 to 6.1) and the vaginal cones group (-14.7 g; -27.6 to -1.8). On completion of the trial one participant in the control group, 14 in the pelvic floor exercise group, three in the electrical stimulation group, and two in the vaginal cones group no longer considered themselves as having a problem.

Conclusion: Training of the pelvic floor muscles is superior to electrical stimulation and vaginal cones in the treatment of genuine stress incontinence.

The authors tell us that the objective of the study was to compare the effect of 4 treatments on genuine stress incontinence (which we knew from the descriptive title). The 107 participants (or sample = n , which we see later how they were divided up) were from multicenter sites. This abstract does cite statistics, and we will review those statistics in our methods section. They let us know that the results were significant

($p = 0.03$) and the confidence interval was 95%. A p value refers to the probability that an error was committed. A p value of 0.03 indicates that the result would have occurred by chance only 3 out of 100 times. (This will be more thoroughly discussed in Part II of this series). Confidence intervals are estimates of the limits of what the population mean might be based on data from a representative sample. A confidence interval of 95% indicates that if the same investigation was performed 100 times on different samples drawn from the same population, the mean of the population would be included within the confidence interval 95 out of those 100 times.

Although many women at the end of the study did improve, the exercise group did the best. This is a study that does capture our interest because we want to know how the groups performed their treatments and how good was the outcome. In other words we would have to read further to find out if the electrical stimulation, vaginal cone, and control group came close to the results of the exercise group.

INTRODUCTION

The introduction lays a foundation for the problem that lead to the study. A review of scholarly literature is not exhaustive but serves to place the study within the context of the literature.

An excellent reference for understanding research is the text, *Foundations of Clinical Research Applications to Practice*, 2nd ed. by Portney and Watkins.² From that reference come the following questions the reader needs to think of when the introduction is read.

1. What is the problem?
2. Is it important?
3. Has the problem been clearly stated?
4. Has the author provided a theoretical context for the study?
5. Are the references appropriate and comprehensive?
6. Is it clear if the study is experimental, correlational, or descriptive?
7. Is the specific purpose clearly stated?
8. Are the hypotheses or guiding questions clearly stated or easily discernable?

Introduction

Urinary incontinence is defined by the International Continence Society as "a condition in which involuntary loss of urine is a social or hygienic problem and is objectively demonstrable."¹ Urinary incontinence is more common in women than in men and affects women of all ages. Prevalence rates in women between 15 and 64 years of age vary from 10% to 30%.² Although only a quarter of all women with this problem seek help,² the approximate annual cost of the condition in the United States has been estimated at \$11.2 billion in the community and \$5.2 billion in nursing homes.² The most common type of urinary incontinence in women is stress incontinence, defined as the involuntary loss of urine during coughing, sneezing, or physical exertion such as sporting activities

or sudden change in position. Genuine stress incontinence is urodynamically proved involuntary loss of urine when the intravesical pressure exceeds that of the urethra with no simultaneous detrusor contraction.¹ Risk factors for genuine stress incontinence are inherently weak connective tissue, vaginal delivery, obesity, strenuous work, and old age.²

Urinary incontinence is a socially embarrassing condition, causing withdrawal from social situations and reduced quality of life.^{3,4} Genuine stress incontinence may lead to withdrawal from regular physical and fitness activities.^{5,6} This withdrawal may be a threat to women's general health and wellbeing as regular moderate physical activity is important in the prevention of osteoporosis, high blood pressure, coronary heart disease, depression, and anxiety.⁷

In 1948 Kegel reported a cure rate of 84% after training of the pelvic floor muscles for women with various types of incontinence.⁸ Surgery soon became the first choice of treatment, however, and not until the 1980s was there renewed interest in physical therapies.⁹ This renewed interest for conservative treatment may be because of higher awareness among women and cost of and morbidity after surgery. Physical therapies to treat genuine stress incontinence include pelvic floor exercises with or without biofeedback, electrical stimulation, and weighted vaginal cones.⁹ Pelvic floor exercise is known to be an effective treatment for genuine stress incontinence,² but randomised controlled trials evaluating electrical stimulation and vaginal cones have given conflicting and inconclusive results, and many of these studies are flawed because of small sample sizes.^{9,10} Though neither electrical stimulation nor vaginal cones have been compared with no treatment, they are commonly used.

We compared the effect of pelvic floor exercises, electrical stimulation, vaginal cones and no treatment in women with genuine stress incontinence.

The authors make a good case for study of genuine stress incontinence as historically only 25% of women with this problem will seek help. The review of the literature tells us that without medical assistance these women will withdraw from society, have a poor quality of life, decrease their regular activity, and as a result have an increased possibility of osteoporosis, high blood pressure, cardiovascular disease, depression, and anxiety. As stated by Bo, Talseth, and Holme in their introduction, billions of dollars a year are spent on this condition as cited in the United States statistics. Pelvic floor exercises have been shown to be an effective treatment for stress incontinence, but the effectiveness of electrical stimulation and vaginal cones is not clear. This experimental study is interesting because it compares exercise with other treatments. The authors are clear in that they are comparing the treatments of pelvic floor exercises, electrical stimulation, and vaginal cones with no treatment at all in genuine stress incontinence. However, they do not give us any hypotheses as to the outcome they expect.

Introduction to Methods

The methods section is often overlooked, or simply skimmed, by readers as it includes a lot of detailed and technical information. However, the truth is that this section is very important to a

clinician who might want to incorporate the investigated variable (eg, intervention/treatment) into patient management. In order to make this judgment, you need to be able to make conclusions about the validity of the study. Domholdt refers to validity as "...the extent to which the conclusions of that research are believable and useful."¹ Two categories of validity that are threats to the believability and usefulness of research results are internal validity and external validity.

Portney and Watkins wrote, "Internal validity is concerned with the question: Did the experimental treatment really cause the observed change in the dependent variable? In other words, are there other (extraneous) factors that may be responsible for that change?"² Cook and Campbell describe several factors which might decrease the ability to conclude that the independent variable (intervention or treatment) caused the change in the dependent variable (eg, incontinence).³ The following are the most common threats to internal validity.

1. **History:** Participation in other activities which may affect the dependent variable (eg, if a participant is placed in an electrical stimulation group but is also concurrently participating in a recreational program that includes exercises which strengthen the pelvic floor muscles).
2. **Maturation:** If the time period of the investigation is long, participants may show increases or decreases in the dependent variable simply due to boredom, strength changes, age-related phenomena, etc.
3. **Attrition:** Individuals drop out of research experiments for many reasons. This may create a bias in favor of, or against, being able to determine the effectiveness of the intervention. For example, if the strength of the pelvic floor muscles was similar with the other group(s) at the start of the study and the people who drop out were the stronger people, then the treatment may either fail because the weaker subjects were too weak to benefit from the exercise, or it may be interpreted as successful because there was a greater chance for them to become stronger.
4. **Testing:** Individuals may demonstrate change in a variable simply because they perform it during the tests of that parameter. For example, it is possible that the simple execution of muscle contractions during the pretest sessions might be enough of a stimulus to increase strength.
5. **Instrumentation:** If the equipment used to measure variables is not carefully calibrated before each measurement, errors in the actual measurement may occur. In addition,

if there is inconsistency in the way that the different investigators take their measurements, this can likewise introduce error into the study. These sources of error can be controlled by ensuring test-retest, intrarater, and inter-rater reliability.

The mechanism by which investigators attempt to control the threats to internal validity is to control as many factors as possible. For example, placing participants in groups by random assignment should result in each group being balanced in terms of their characteristics. Investigators also should attempt to control rater reliability by training and testing the ability of the investigators who will be involved to ensure that they are consistent and standardized in their techniques. All equipment must be calibrated before each measurement session.

Regarding external validity, Portney and Watkins wrote, "External validity refers to the extent to which the results of a study can be generalized to other populations, environmental conditions, or times."² The reader of research may only generalize the conclusions of the research to a person, or a sample of people, who are similar to the participants included in the investigation. That is, the results of an investigation that included only female participants diagnosed with stress incontinence may not be used to make the same conclusions about the effects of a treatment in females diagnosed with urge incontinence.

Critical Reading of Methods

When reading the methods section, you should first look for a description of the type of design employed in the research. Overall, there are 2 main categories of research: experimental and nonexperimental. Portney and Watkins explained that, "Experimental research refers to an investigation where the researcher manipulates and controls one or more variables and observes the resultant variation in other variables.... Nonexperimental research refers to investigations that are generally more descriptive or exploratory in nature and that do not exhibit a strong degree of control over the studied variables."² There are many types of designs within these categories which might be used depending on the nature of the question being asked by the investigator(s), and there are excellent descriptions of these methodologies in other references. Since the paper we are reviewing for this series includes an experimental design, this tutorial will concentrate on that type of design. In terms of experimental research, following is a review of terminology often found in a report.

1. **Cohort:** A group of participants/patients that share specific characteristics.
2. **Control Group:** A group of participants who

are similar to the experimental group and are tested at the same times as the experimental group (typically at the beginning and at the end of the study), but do not receive the intervention. This group provides a way to evaluate the effectiveness of a treatment.

3. **Double-blinded:** Investigators and participants do not know which treatment group each person is assigned into (treatment groups versus a control group).
4. **Experimental Group:** It is the experimental group that will receive what is referred to as the independent variable, or the variable being manipulated or studied (eg, a modality, an exercise). In this study, the variables are electrical stimulation, exercise, and vaginal cones.
5. **N of 1 Randomized Control Trial (RCT):** With this design 2 different interventions (treatments) are randomly implemented for 1 patient at different phases of the study. There are 2 time periods in each phase, a treatment period and either another treatment period or a control or a placebo period. The sequence of the 2 periods is randomly assigned. The patient is monitored to study the effectiveness of the treatments.⁴
6. **Randomized Control (or Clinical) Trials (RCT):** At the very minimum, a randomized control trial includes one experimental group and one control group. Each participant has an equal chance of being assigned to either of the groups included in the investigation. The design also may include more than 1 experimental group, a control group, and/or a placebo group. The assignment is not decided upon by either the participants or the clinicians. Randomized control trials are the most effective method to determine the effectiveness of an intervention. This design is sometimes referred to as a between-subjects design.
7. **Reliability:** This parameter refers to the repeatability or consistency of measurements that are taken during an investigation. Sources of error which may decrease the repeatability of measurements may include equipment, investigators (or raters), and/or the participants. There are several categories of rater reliability:
 - **Intra-rater reliability** is the ability of one rater to be consistent in methods across trials.
 - **Inter-rater reliability** is the ability of two or more raters to be consistent in measuring the same parameter in the same group of participants.
8. **Single Blinded:** Either the participant or the investigator collecting data does not know

which intervention they are receiving, or the participant is not made aware of the specific hypothesis being studied. The study here is singleblinded.

Methods

This study was a multicentre, single blind, randomised controlled trial with stratified design. Participants were women with genuine stress incontinence who were on the surgical waiting list or women with symptoms of stress incontinence recruited by local newspaper articles. Five centres in southeast Norway participated. A standardised assessment at enrollment included a comprehensive urogynaecological history, urodynamic assessment including uroflowmetry and cystometry, bacteriological examination, and pad test with standardised bladder volume. The study was approved by the local ethics committee, and all women gave written consent.

Inclusion criteria were history of stress urinary incontinence and > 4 g of leakage measured by pad test with standardised bladder volume. Exclusion criteria were urinary incontinence other than genuine stress incontinence, involuntary detrusor contractions exceeding 10 cm H₂O on cystometry, abnormal bladder function (residual urine > 50 ml and maximal uroflow < 15 ml/s), previous surgery for genuine stress incontinence, neurological or psychiatric disease, ongoing urinary tract infections, other diseases that could interfere with participation, use of concomitant treatments during the trial, and inability to understand instructions given in Norwegian.

The power calculation of the study was based on the power estimation and results of a previous study designed to detect differences between groups of 1 SD with a power of 80% and α of 5%.¹¹ In the previous study significant differences in the same outcomes after the same training programme were shown in groups of 23 and 29 subjects; therefore 30 participants were recruited for each of the four groups in this study.

Randomisation procedure

The participants were stratified into two groups (≤ 20 g and > 20 g leakage) according to results of the pad test with standardised bladder volume. Randomisation schemes stratified by degree of incontinence were constructed for all sites by using computer generated random numbers. Participants within each stratum were randomised by using opaque sealed envelopes to one of the four study groups: pelvic floor exercises, electrical stimulation, vaginal cones, or untreated control. Information for decoding randomisation was kept locked in the statistician's office. The main investigator (KB) was not involved in any interventions and was blind to group allocation. Physicians evaluating the effect of the treatments were also blind to allocation of treatments.

In the first sentence of the methods section you find the description of the research design. From it you can determine that the data were collected at more than one site and that the participants were randomly assigned to groups. We want to clarify the meaning of random selection of participants versus random assignment of participants. If people are randomly selected to participate in a study, this means that each and every person in the defined population has an equal chance at being selected to be included in the sample to be studied. For example, each woman from a specified geographic location who has been diagnosed with stress urinary incontinence might be placed on a list. Each potential participant could then be selected in a lottery-like fashion. Random assignment is the process

of placing the participants that were chosen as the representative sample of the population into the experimental or control group(s). It is essential that each participant have an equal chance of being placed into any of the groups. The goal of this procedure is to ensure that there is no systematic bias in group assignment. For example, it is hoped that with random assignment that those with the weakest pelvic floor muscles are not over-represented in any one group. In the study by Bo, Talseth, and Holme, the authors indicated that they did not use a pure randomization technique. They implemented a stratification design. This technique is used to attempt to ensure more homogenous groups on the basis of a parameter which could affect the results. The authors describe the procedure in the section that is titled "randomisation procedure." They explained, "The participants were stratified into two groups (≤ 20 g and > 20 g leakage) according to the results of the pad test with standardized bladder volume. The purpose was to ensure that one group did not get biased by possibly having more subjects with less leakage while another group would have more subjects that demonstrated more leakage at the start of the study. In addition, the investigators who collected the data were blinded to which treatment group each participant was assigned.

The next very important thing you would look for in a research report is a detailed description of the subjects included in the investigation. The reason that you need to know this information is so that you can determine if the participants are similar to the patient/client that you are going to treat. As you read a research paper, look for the following information.

1. How many participants were included in the investigation (later in this paper will be a discussion of how to determine whether there were enough subjects)?
2. What were the characteristics of the participants? For example, are they in the same age range of your patient? Did they have the same diagnosis as your patient, and how was this diagnosis determined? How long have they had the condition in question? Additional characteristics which may have some effect on the results of the ability of the participants to respond to any interventions (eg, sex, height, weight, socioeconomic status, other health conditions or comorbidities, surgeries) are important to know as they allow you to consider the similarity of your patient to the population studied. Comorbidities are defined as other disease states occurring at the same time. These could influence how the participant responds to the treat-

ment variable.

3. What were the inclusion and exclusion criteria used to select the population to be studied. Again, these can further assist you in determining if this study included people similar to your patient.
4. How were the participants recruited? This is important to know because if health care practitioners recommend certain patients for treatment, this could set up a bias to recommend only those who have a better chance at responding to the treatment. If it is via newspaper advertisements, or via announcements at a specific site, such as a health club, this can result in a sample of only literate people or people who exercise. From the information provided, once again, you can decide if the participants are similar in nature to your patient.

Bo, Talseth, and Holme did provide a very thorough description of their subject population, although all the information is not found in the methods section. The number of women, age range, diagnosis, and duration of symptoms was placed in the abstract.

They stated that the participants "... were women who were on the surgical waiting list or women with symptoms of stress incontinence recruited by local newspaper articles." Both inclusion and exclusion criteria were very specific.

Please note the last sentence of the first paragraph in the methods section. "The study was approved by the local ethics committee, and all women gave written consent." We want to place a strong emphasis on this point. In the *Guide to Physical Therapist Practice* there is a Guide for Professional Conduct that is included in appendix 3.⁵ Section 4.3 is titled Research and subsection B: 1 emphasizes that informed consent of each subject must be obtained. This pertains to all research, whether it is conducted in a hospital, university, clinic, or private practice setting. Each setting must develop a protocol for obtaining signed informed consent. The text by Portney and Watkins includes an outline of the information that must be given to a potential participant in order to ensure that they are giving informed consent prior to participating in the study.² It is extremely important that each person is given the information in her/his native language.

In the third paragraph, the authors list the power (80%) and the alpha (α) level (.05) that they chose for statistical analysis of the data that was generated in this study. We believe that it is worth the time to understand these concepts because it can determine the amount of credibility you want to place on the conclusions reached by the authors.

As you have already determined, the objective of the research by Bo, Talseth, and Holme was to compare the effectiveness of several treatments used for the condition of genuine stress incontinence. The review of the literature performed by the authors revealed the most common treatments used for stress incontinence and the fact that questions remained unanswered by the available data. Before collecting data, researchers must make a decision regarding the probability that they will be able to reject a null hypothesis. Fraenkel, Wallen, and Sawin explain that, "A null hypothesis is the hypothesis that is actually tested in using a procedure for checking on statistical significance."⁶ A null hypothesis is a statement that there was no difference at the end of the study between the groups that received different treatments. Unfortunately, even the most intricately designed experiments can lead to wrong conclusions because of the difficulty in controlling all extraneous variables that can influence the outcomes (eg, characteristics of participants, equipment error, and investigator error). There are 4 possibilities of conclusions about data to include: (1) Conclusion that the null hypothesis is true when it is true. That is, the difference was indeed due to chance. (2) Rejection of the null hypothesis when it is false. That is, the difference was indeed due to the intervention. (3) Rejection of the null hypothesis when it was true. That is, the researchers conclude that the difference was due to the intervention when it was indeed due to chance. This is referred to as a type I error. (4) Conclusion that the null hypothesis is true when it is false. That is, the researchers conclude that the difference was due to chance when it was indeed due to the intervention. This is referred to as a type II error.

As you may guess, most researchers usually undertake an investigation hoping to be able to conclude that one treatment is more effective than another (or no treatment). That is, they would like to be able to reject the null hypothesis. The power of a statistical test is the probability that the null hypothesis will only be rejected when the treatment used on a group of subjects resulted in improvement that was less likely due to chance alone. That is, the treatment was the cause of the improvement. Fraenkel, Wallen, and Sawin outlined what we believe is a very clear analogy regarding what different levels of power mean to research. They wrote, "The power of a statistical test is in some ways like the power of a telescope. Astronomers looking at the planets Mars and Venus with a low-power telescope probably can see that they look like spheres, but it is unlikely that they can see much by way of differences in terrain—such as mountains, valleys, and canyons. With an extremely high-

power telescope, however, they can see such differences. When the purpose of a statistical test is to check on differences, power is the likelihood that the test will correctly yield a conclusion that there *are* differences when, in fact, differences actually exist."⁶ Therefore, the power of a test is really the sensitivity of a test to detect true differences. When researchers state that the power of their data was 80%, this means that there is a probability of 80% that they will reject the null hypothesis when it is indeed false and therefore that statistically significant differences exist between the means of the groups.

Investigators are usually hopeful that at the end of their data collection, they can state that there was a significant difference between the means of the groups. Regarding the concept of significant difference Bailey wrote, "Significance testing is based on the laws of probability. It answers the question: What is the probability that this change occurred because of events in the research study, and what is the probability that this change would have occurred anyway, by chance? The tests that are used to make this determination result in a level of probability, and it is the researcher who decides whether or not this level is significant."⁷

Another variable determined by the authors is known as alpha (α) level. Before beginning an investigation, the researchers must decide how willing they are to draw an incorrect conclusion from their data. That is, what probability are they willing to accept that they might conclude that the null hypothesis is false when it is indeed true? Again, if this is done, then researchers have concluded that their treatment is more effective than chance when it was not. It is extremely common that this level is set to be 5%, or .05. Domholdt explains, "If a difference in means is significant at the .05 level, this means that 5% of differences of this magnitude would have been the result of chance fluctuations caused by sampling errors. That is, 95% of the time the difference would represent a true difference and 5% or the time the difference would represent sampling error."¹ In other words, there is a 5% chance that the investigators have committed a type I error.

Bo, Talseth, and Holme used the power of 80% and an alpha level of .05 in a prior study to determine sample size. Regarding sample size, Portney and Watkins comment, "the influence of sample size on power of a test is critical. The larger the sample, the greater the statistical power. Smaller samples are less likely to be good representations of population characteristics, and, therefore, true differences between groups are less likely to be recognized."² It follows then that larger samples are more likely

to be representative to the population as a whole, and therefore, the conclusions can be generalized to the population with greater trust. When designing any research, one of the first things to be determined is how many subjects will be required in order to make correct conclusions. Anyone who has been involved in research understands the necessity of limiting the number of subjects required to complete an investigation. Each subject requires a significant amount of time, effort, and perhaps cost to take through the research process. Domholdt outlines the parameters that are required to calculate sample size. These include the desired power, the alpha level that will be used for the analysis, an estimate of the between-groups difference that would be considered relevant to the practicing clinician, and an estimate of the variability that is expected within the groups included in the research design. She continues to explain, "A researcher can obtain these values from previous research or from a pilot study."¹ Indeed, as you read the paper by Bo, Talseth, and Holme, you can see that they used data from their previous research to calculate the sample size for this project. The variability is determined by the standard deviation of the data. We think it is helpful to understand that both a higher power level and a lower significance level will require more subjects.⁸ It has been estimated that for experimental research, having 30 participants per group is usually the minimum required in order to apply the results to the population.⁹

The next section of the paper by Bo and Talseth contains descriptions of all the interventions that were included as independent variables. These included: pelvic floor exercises, electrical stimulation, and vaginal cones. As anyone who has used the above interventions knows, there are many ways that each of the 3 treatments can be implemented. For example, the number of pelvic floor exercises and the duration of each contraction, the frequency and pulse width of the electrical stimulation waveform, the intensity of electrical stimulation, the weight of the vaginal cone, and the duration of daily exercise with the cone inserted. One of the major goals of the methods section is to be so precise in the description of how each intervention was implemented that any person could accurately replicate the protocols. Portney and Watkins wrote, "Once a problem has been formulated and variables of interest identified, a researcher must define those variables in precise terms that explain how they will be used in the study."² These definitions are referred to as operational definitions.

Interventions

Participants were taught about the anatomy of the pelvic floor and lower urinary tract, physiology, and continence

mechanisms by the local project physical therapist. All were taught to contract the pelvic floor muscles correctly, and this was assessed by vaginal palpation.

Participants in the three treatment groups were told that the three treatments were expected to be equally effective and were discouraged from using other treatments during the 6 month trial period. All patients in the three intervention groups met the physical therapist once a month for motivation, monitoring of pelvic floor muscle strength, and adjustment of treatment if necessary. The untreated control group had no contact during the intervention period but were offered instruction on the use of the continence guard (Coloplast AS).¹²

Pelvic floor muscle training-The protocol has been published previously¹¹ and followed recommendations for general training to increase strength of skeletal muscle.¹³ Participants were asked to conduct 8-12 high intensity (close to maximum) contractions three times a day at home with additional training in groups once a week for 45 minutes with a physical therapist. Group training was performed in lying, standing, kneeling, and sitting positions with legs apart to emphasise specific strength training of the pelvic floor muscles and relaxation of other pelvic muscles. Participants aimed at holding each muscle contraction for 6-8 seconds, three or four fast contractions were then added. The rest period was about 6 seconds. A total of 8-12 contractions were completed in each position with maximal contraction effort encouraged. Body awareness, breathing, relaxation exercises, and strength training for the abdominal, back, and thigh muscles were performed to music between positions. The participants were encouraged to use their preferred position and perform equally intensive contractions at home. An audiotape with verbal guidance for 12 maximum contractions was available for home training, and a training diary was kept.

Electrical stimulation-An MS 106 Twin (Vitacon AS, Trondheim, Norway) was used according to the manufacturer's recommended protocol for 30 minutes of intermittent vaginal electrical stimulation per day. Selected parameters included biphasic intermittent current, frequency 50 Hz, pulse width 0.2 milliseconds, and current intensity between 0-120 mA with individually adapted on-off (duty) cycles on the basis of each woman's ability to hold a voluntary contraction. On time ranged from 0.5 seconds to 10 seconds, and off time from 0 seconds to 30 seconds. If ability to hold the contraction improved the duty cycle was progressed each month. All patients were encouraged to tolerate as high an intensity as possible to get a contraction. Treatment adherence was electronically monitored and recorded. At every monthly visit the physical therapist observed the patients receiving electrical stimulation from their home stimulators in the clinic.

Vaginal cones-Mabella cones (Vitacon AS, Trondheim, Norway) were used for 20 minutes a day according to the manufacturer's recommendations. Patients progressed through three cone weights-20, 40, and 70 g-according to their ability to hold the cones. Adherence was noted in a training diary.

Adverse effects and tolerance to treatment

Adverse effects and treatment tolerance were monitored with a training diary and during monthly clinic visits.

Main outcome measures

Pad test with standardised bladder volume-After the bladder was emptied by catheter it was refilled with 200 ml saline. Women wore preweighed pads and ran on the spot for 30 seconds followed by 30 seconds of jumping with legs in subsequent adduction and abduction (jumping jacks) at a preset metronome rate of 132 beats per minute. After the test the pad was reweighed.

Subjective assessment-Women recorded how they perceived the condition before and after treatment on a 5 point scale (unproblematic, minimal problem, moderate problem, problematic, very problematic).¹

Secondary outcome measures

Three day leakage episodes-The number of episodes of involuntary leakage in 3 days was recorded in a home voiding diary before and after the intervention period. Mean number of episodes was calculated.

Twenty four hour pad test-Twenty four hour pad weights were conducted by patients at home before trial entry and after the last clinic visit. Women chose a typical day that mirrored their average level of activity.

Leakage index-Patients indicated on a 5 point scale (5 always, 4 often, 3 sometimes, 2 seldom, 1 never) the frequency of urinary leakage during sneezing, coughing, laughing, walking, walking downhill, running, jumping, and lifting. The mean was calculated as an index of leakage frequency before and after treatment.¹⁴

Social activity index-Perceived problems in participating in nine different social situations were recorded on a 10 cm visual analogue scale (0 impossible to participate, 10 no problem taking part). As an overall index of quality of life the mean was calculated before and after treatment.¹⁴ After treatment participants also rated improvement on a 5 point scale (worse, unchanged, improved, almost continent, continent)¹¹ and stated whether they wanted further treatment.

Muscle function and strength

Pelvic floor muscle function was assessed by the physical therapist with vaginal evaluation during contraction. Muscle strength was evaluated by a vaginal balloon catheter (balloon size 6.7 x 1.7 cm) connected to a pressure transducer (Camtech AS 1300, Sandvika, Norway). The middle of the balloon was placed 3.5 cm inside the vaginal introitus.¹⁵ Only contractions with simultaneous observable inward movement of the perineum were considered valid.¹⁶

Resting maximum urethral pressure and maximum urethral closure pressure were measured before and after treatment with a fiberoptic microtransducer. All terminology conforms to International Continence Society standards.¹

Statistical methods

The primary analysis was carried on data from treated participants, with exclusion of data from those without final evaluation on efficacy variables. Additional intention to treat analyses were also done for all randomised patients including those who dropped out. The missing last values were considered as equal to baseline values. Results are given as mean values with 95% confidence intervals. As several variables were not normally distributed, however, the Kruskal-Wallis analysis of variance was chosen as the global test of differences between groups on visual analogue scales and other interval scaled variables. Pair-wise comparisons were made with the Mann-Whitney U test to compare each group with the control and one intervention group with another. Cochran-Mantel-Haenszel tests or [chi squared] tests were used if data were nominal or categorical. P values < 0.05 were considered significant.

As you read through the description of the interventions, you should decide if enough information was given by the authors for you to understand what was done. In the first paragraph the authors inform the readers that every subject received education regarding the physiology and anatomy of pelvic floor, urinary tract, and continence. In addition they stated, "All were taught to contract the pelvic floor muscles correctly, and this was assessed by vaginal palpation" (in the muscle function and strength section). In our opinion, the information in the first paragraph is not adequate for future attempts at replicating this research. We are left with the following questions:

1. Were the participants taught the pelvic floor contractions with only verbal instructions, or was electromyographic activity (muscle activity) used to monitor the contractions?
2. What were the specific instructions used to teach the contractions?
3. What was assessed by the vaginal palpation?
4. What was the procedure for palpation?
5. Which grading scale was used for the palpation?
6. What steps were taken to ensure high inter-rater reliability between the researchers at multiple sites?

Another problem area is in the section describing the pelvic floor muscle training. The authors document that "Body awareness, breathing, relaxation exercises, and strength training for the abdominal, back, and thigh muscles were performed to music between position." Without precise operational definitions of each parameter, exercise, or technique, readers cannot replicate these methods. The fact that participants also performed exercises other than pelvic floor muscle contractions is confounding variables which may have affected the results. How are readers to know if the pelvic floor muscle exercises by themselves would have been successful? Also, there was no description of the other exercises or an outline of the protocols employed (number of repetitions, etc.).

A very important section is the description of the outcome measures employed. These are the tools researchers use to analyze the effectiveness of each intervention. In this study we were informed that the pad test with standardized bladder volume and a subjective assessment by each participant were the primary outcome measures. The authors report the secondary measures to be: (1) a 3-day leakage episode account, (2) 24-hour pad test, (3) leakage index, and (4) social activity index. In addition, pelvic floor muscle strength was assessed via a vaginal balloon catheter connected to a pressure transducer, and resting maximum urethral pressure and maximum urethral closure were assessed with a fiberoptic microtransducer.

In order for readers to make a judgment about the believability of the data, researchers need to inform them of the established reliability and validity of the outcome measures employed. While there are references for the established reproducibility, or reliability, of the outcome measures (Bo, Talseth, Holme references^{4,14,27,28}), we believe it would be more helpful to readers if the reliability values were included.

The final section of the paper contains a description of the statistical methods which were used to analyze the data. These will be discussed in part II of this tutorial.

In summary, we have read and analyzed the first 3 sections of this study on genuine stress incontinence: the abstract, introduction, and methods. This lays the groundwork for us to look at Part 2: the results, discussion, and conclusion. Are there any questions you have as a reader? Would this paper be relevant for your practice in women's health care?

ACKNOWLEDGEMENT

We want to express our gratitude to Mary P. Watkins for her assistance in reviewing this paper and offering helpful critique, advice, and suggestions.

REFERENCES

1. Domholdt E. *Physical Therapy Research: Principles and Applications*. Philadelphia, Pa: W.B. Saunders Co.; 2000.
2. Portney LC, Watkins MP. *Foundations of Clinical Research: Applications to Practice* 2nd ed. Upper Saddle River, NJ: Prentice-Hall Inc.; 2000.
3. Cook TD, Campbell DT. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston, Mass: Houghton Mifflin; 1979.
4. Sackett DT, Straus AE, Richardson WS, et al. *Evidence-Based Medicine: How to Practice and Teach EBM*. London, UK: Churchill Livingstone; 2000.
5. American Physical Therapy Association. *Guide to physical therapist practice*. 2nd ed. *Phys Ther*. 2001;9:744.
6. Fraenkel J, Wallen N, Sawin ET. *Visual Statistics*. Boston, Mass: Allyn and Bacon; 1999.
7. Bailey DM. *Research for the Health Professional: A Practical Guide*. Philadelphia, Pa: FA. Davis Co.; 1997.
8. Peipert JE, Glennon M. Observational studies. *Clin Obstet Gynecol*. 1998;41:235-244.
9. Kraemer HC, Thieman S. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, Calif: Sage Publications; 1987:27.

REFERENCES from Bo article

1. Abrams P, Blaiwas JG, Stanton SL, Andersen JT. The standardisation of terminology of the lower urinary tract function. *Scand J Urol Nephrol Suppl* 1988;114:5-19.
2. Fand J, Newman D, Colling J, DeLancey JOL, Keeys C, Loughery R, et al. *Urinary incontinence in adults: acute and chronic management*. 2nd update. Rockville, Maryland: Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, 1996; (Clinical Practice Guideline, 96-0682.)
3. Norton P, MacDonald LD, Sedgwick PM, Stanton SL. Distress and delay associated with urinary incontinence, frequency, and urgency in women. *BMJ* 1988;297:1187-9.

4. Hunskaar S, Vinsnes A. The quality of life in women with urinary incontinence as measured by the sickness impact profile. *J Am Geriatr Soc* 1991;39:378-82.
5. Bo K, Hagen R, Kvarstein B, Larsen S. Female stress urinary incontinence and participation in different sport and social activities. *Scand J Sports Sci* 1989;11:117-21.
6. Nygaard I, DeLancey JOL, Arnsdorf L, Murphy E. Exercise and incontinence. *Obstet Gynecol* 1990;75:848-51.
7. Bouchard C, Shephard R, Stebbens T, eds. *Physical activity, fitness, and health. International proceedings and consensus statement*. Champaign: Human Kinetics Publishers, 1994.
8. Kegel AH. Progressive resistance exercise in the functional restoration of the perineal muscles. *Am J Obstet Gynecol* 1948;56:238-49.
9. Bergbomans LC, Hendricks HJ, Bo K, Hay-Smith EJ, de Bie RA, van Waalwijk van Doorn ES. Conservative treatment of stress urinary incontinence in women. A systematic review of randomized clinical trials. *Br J Urol* 1998;82:181-91.
10. Bo K. Effect of electrical stimulation on stress urinary incontinence. Clinical outcome and practical recommendations based on randomized controlled trials. *Acta Obstet Gynecol* 1998;77(suppl 168):3-11.
11. Bo K, Hagen RH, Kvarstein B, Jorgensen J, Larsen S. Pelvic floor muscle exercise for the treatment of female stress urinary incontinence. III: Effects of two different degrees of pelvic floor muscle exercise. *Neurourol Urodyn* 1990;9:489-502.
12. Thyssen H, Lose G. Long-term efficacy and safety of a disposable vaginal device (continence guard) in the treatment of female stress incontinence. *Int Urogynecol J* 1997;8:130-3.
13. American College of Sports Medicine. *Position stand. The recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness in healthy adults*. *Med Sci Sports Exerc* 1990;22:265-74.
14. Bo K. Reproducibility of instruments designed to measure subjective evaluation of female stress urinary incontinence. *Scand J Urol Nephrol* 1994;28:97-100.
15. Bo K. Pressure measurements during pelvic floor muscle contractions: the effect of different positions of the vaginal measuring device. *Neurourol Urodyn* 1992;11:107-13.
16. Bo K, Kvarstein B, Hagen R, Larsen S. Pelvic floor muscle exercise for the treatment of female stress urinary incontinence. II: Validity of vaginal pressure measurements of pelvic floor muscle strength and the necessity of supplementary methods for control of correct contraction. *Neurourol Urodyn* 1990;9:479-87.
17. Lagro-Janssen TLM, Debruyne FMJ, Smits AJA, Van Weel C. Controlled trial of pelvic exercises in the treatment of urinary stress incontinence in general practice. *Br J Gen Pract* 1991;41:445-9.
18. Henalla S, Hutchins C, Robinson P, MacVicar J. Non-operative methods in the treatment of female genuine stress incontinence of urine. *J Obstet Gynaecol* 1989;92:22-5.
19. Bo K. Pelvic floor muscle exercise for the treatment of stress urinary incontinence. An exercise physiology perspective. *Int Urogynecol J* 1995;6:282-91.
20. Sand PK, Richardson DR, Staskin SE, Swift SE, Appell RA, Whitmore KE, et al. Pelvic floor stimulation in the treatment of genuine stress incontinence: a multi-center placebo controlled trial. *Am J Obstet Gynecol* 1995;173:72-9.
21. Luber K, Wolde-Tsadik G. Efficacy of functional electrical stimulation in treating genuine stress incontinence: a randomized clinical trial. *Neurourol Urodyn* 1997;16:543-51.
22. Brubaker L, Benson JT, Bent A, Clark A, Shott S. Transvaginal electrical stimulation for female urinary incontinence. *Am J Obstet Gynecol* 1997;177:536-40.
23. Dudley GA, Harris RT, Komi PV, eds. *Use of electrical stimulation in strength and power training. In: Strength and power in sport*. Oxford: Blackwell Scientific Publications, 1992:329-37.
24. Bouchard C, Shephard RJ, Stebbens T, ed. *Physical activity, fitness and health. Consensus statement. In: Physical activity, fitness, and health: status and determinants. Adjuvants to physical activity*. Champaign: Human Kinetics Publishers, 1993:33-40.
25. Bo K, Talseth T. Change in urethral pressure during voluntary pelvic floor muscle contraction and vaginal electrical stimulation. *Int Urogynecol J* 1997;8:3-7.
26. Bo K. Vaginal weight cones. Theoretical framework, effect on pelvic floor muscles strength and female stress urinary incontinence. *Acta Obstet Gynecol Scand* 1995;74:87-92.
27. Blaiwas J, Appell R, Fantl J, Leach G, McGuire E, Resnick N, et al. Standards of efficacy for evaluation of treatment outcomes in urinary incontinence: recommendations of the urodynamic society. *Neurourol Urodyn* 1997;16:147.
28. Lose G, Rosenkilde P, Gammelgaard J, Schroeder T. Pad weighing test performed with standardized bladder volume. *Urol* 1988;32:78-80.
29. DeLancey J. Stress urinary incontinence: where are we now, where should we go? *Am J Obstet Gynecol* 1996;175:311-9.
30. Bo K, Talseth T. Long term effect of pelvic floor muscle exercise five years after cessation of organized training. *Obstet Gynecol* 1996;87:261-5.

How to Review and Understand a Research Article: Part II

Nancy Rich, PhD, PT, FACSM,*
Rebecca G. Stephenson PT**

In Part I of this tutorial we presented guidance regarding the review of the abstract, introduction, and methods sections of a research article. We continue in Part II to provide instruction for reviewing the statistical analyses, results, discussion, and conclusion sections. The format of this article will be similar to that of part I in that we will continue with our review of the published article by Bo, Talseth and Holme¹ which was reprinted in the March issue of the *Section on Women's Health Journal* (Bo K, Talseth T, Holme I. Single blind randomized controlled trial of pelvic floor exercises, electrical stimulation, vaginal cones, and no treatment in management of genuine stress incontinence in women. *BMJ*. 1999;318:487-93.). The relevant sections of the article will follow our comments. In addition, the article we chose to review in Parts I and II of this tutorial is unique in that a group of researchers published a brief critique of their perceived flaws of the study. Bo, Talseth, and Holme had the opportunity to respond to that critique. We have included our review of those 2 commentaries.

STATISTICAL METHODS

In this brief section, Bo et al¹ outline all of the statistical analyses that were performed on the data. The authors make a particular note of the fact that for the primary analyses they used only the data obtained from the subjects that completed the entire investigation. However, they also completed analyses on all the women originally enrolled in the study, including subjects who dropped out before final data could be collected. They explained that because they had no final values for the subjects who dropped out of the study, they filled in the missing data values with the original baseline values of each subject for the analyses. This is one of the mechanisms that have been designed to deal with subjects who withdraw from a study prior to the collection of final data. In a later section, the authors account for the reasons that each subject dropped out. Since some subjects who drop out do so because of deleterious effects of the treatment, it is easy to understand that if outcome results are based only on the subjects who completed the study, then the conclusions may be biased in favor of those who did not experi-

ence any bad side-effects. A further interpretation is that only those seeing benefits from the treatment complete the study. That is, those that were not benefiting discontinued participation. Both of the above situations could result in a treatment being interpreted as more efficacious than it would be in the population. Alternatively, it could also occur that subjects who achieved their goal prior to the final data collection discontinue their participation. In that scenario the analysis might be biased against finding that a treatment was effective. Riegelman wrote, "When statistically significant differences between groups remain after assuming the worst case for those lost to follow-up, the reader can be quite confident that loss to follow-up does not explain the observed differences."²

Lilienfeld and Lilienfeld³ wrote, "In these attempts to avoid bias, the investigator usually should assume the most conservative outcome for those patients who have withdrawn from the study or have been lost to follow-up. A broad estimate of the effect of these groups on the overall findings can be made by calculating the 2 extremes of a range, one based on assuming the most conservative outcome and the other based on assuming the best possible outcome." Again, the authors completed one analysis without the data from the subjects who dropped out and another analysis in which they used the original values as the final values. The latter method would signify a result of no difference for those particular subjects in the analysis. Therefore, in the intention-to-treat analyses, those that completed the study would have to show favorable results to overcome the no difference from the people who dropped out.

Bo et al¹ next inform readers that they chose to report their data in the form of mean values with 95% confidence intervals. The mean is the most basic statistic calculated and reported. It is simply the arithmetic average calculated by summing all the data points from each subject for a specific measurement (eg, age, weight, duration of incontinence) and dividing by the total number of subjects. Many researchers choose to report the means and the standard deviations for their data. While the means inform the readers of a representative score of the distribution of scores, the value does not give the reader any information regarding how much

*Department of Physical Therapy,
University of New England,
Biddeford, Maine

**Stephenson Physical Therapy,
Medfield, Massachusetts

the subjects varied from one another, or the spread (or dispersion) of each individual score from the mean. A measure of variability (eg, the standard deviation) is useful in that it informs the reader how similar or dissimilar the subjects were with respect to specific variables. An example of how this might be relevant is that if the pelvic floor muscle strength values were not very variable before a treatment but were more variable after the treatment, one could conclude that the treatment was less effective for some of the subjects compared to other subjects. One main point to remember when looking at data is that if the variability within groups is large at the beginning of an investigation, it will be more difficult to determine if the groups are statistically different from one another at the conclusion of the study.

We believe it is helpful to understand that the significance of any difference between groups is determined from mean and variability values. Portney and Watkins⁴ wrote, "Thus, we would want the separation between the group means to be large and the variability within groups to be small." It is very important to look at initial values for all groups to see that they do not differ significantly. Readers should look for a statement from the authors that no significant differences existed between the means of the groups for the baseline measurements. Bo et al¹ included the statement, "At baseline there were no significant differences between the groups in any of the background characteristics such as age, body mass index, duration of symptoms, pelvic floor muscle strength, urodynamic assessment, or degree of leakage (table 1)." We believe it would have been clearer if the researchers had included the result of their analysis on each baseline measurement in Table 1. They could have included one more column in which they placed "NS" beside each measurement that was determined to demonstrate no significant differences between the groups.

It is possible to get an idea of variability of the sample studied by subtracting each score from the mean (the deviation from the mean). It is easy to see that the larger the deviation score is, the more variable are the subjects in a sample. Those data sets with smaller deviations from the mean are more homogenous (less variable) and those with larger deviations are termed heterogeneous (more variable). However, to calculate the deviations from the mean, one would have some values that are less than the mean and some values that are more than the mean. If one subtracts each score from the average or mean value, the summary of pluses and minuses will always equal zero. This results in an inability to calculate the average deviation score. To solve

that problem, it makes sense to square the deviations so that each value will have a plus sign. If all the squared deviations are added, this produces a number termed the sum of squares. It still is not possible to compare these numbers with other groups because with a large number of subjects this number would automatically be greater. So, the final step is to divide the sum of squares by the number of subjects in the group. This calculation is referred to as the mean of the squared deviations (mean square), or the variance. However, because variance is based on squared deviation units, early statisticians looked for a method to transform the results back to the form of the original data (unsquared units). The problem was solved by calculating the square root of the variance. This final calculation is referred to as the standard deviation. The standard deviations reflect the dispersion or variability of the values for a particular measurement. That is, a larger standard deviation means that the individual scores were farther apart.

In Table 1 we can see that the age in years of the control group was 51.7 (mean). The number in parentheses (8.8) is the standard deviation. Often the standard deviations are presented as ± 8.8 , meaning that the average of the deviations on each side of the mean is 8.8. In order to look further into the spread of the scores, one can actually calculate the numbers. For example, one standard deviation below the mean can be calculated as 42.9 years (51.7 minus 8.8) and one standard deviation above the mean is 60.5 (51.7 plus 8.8). Table 1 is the only table that presents the data in terms of means and standard deviations. In Tables 2 and 3 the authors present the data in means and the 95% confidence intervals and 99% of the scores are within 3 standard deviations of the mean.

It is important to remember that one of the basic premises of many studies is that investigators want to be able to collect data from a sample of the total population and then infer that similar results would be obtained if data were collected on each member of the defined population. Confidence intervals (CI) are calculated to give readers the range (or lower and upper confidence limits) of values in which they can be reasonably confident that the mean of the population (versus the sample) will fall. When authors produce a 95% confidence interval, measurement theory holds that if the study was repeated using 100 different samples, the confidence intervals around each of the sample means would be expected to include the population mean on 95 occasions.⁴⁶ Sim and Reid⁷ produced a very clearly written tutorial on the topic of confidence intervals in 1999. They wrote that "For a given level of confidence, the

narrower the CI, the greater the precision of the sample mean as an estimate of the population mean."

The authors continued to outline 3 factors that contribute to the width of the confidence interval. They stated that a confidence interval will be narrow if the variability of the sample is small, if the sample size is large, and if a lower level of confidence is chosen to be calculated for the sample. If one chooses to calculate the limits of a 99% confidence interval, the interval will be wider than the limits of the 95% CI. Researchers can attempt to reduce variability by employing reliable measurements and by selecting a homogenous sample. The latter is best obtained by having careful inclusion and exclusion criteria for subject selection. What is the benefit of having a narrow confidence interval? Sim and Reid⁷ explained, "If a 95% CI is narrow, this means that only a small range of possible values has to be included in order to be 95% confident that the CI contains the parameter." It is important to remember that, by definition, if the CI is 95%, there is still a 5% chance that the mean of the population would not be found in the calculated interval. Some researchers choose to calculate the wider interval of 99%. This indicates that the interpreter can be 99% confident that the population mean is included in the interval. For the same data, a 99% CI will be wider than a 95% CI. In other words, precision is sacrificed in order to have a greater chance at being accurate.

In Table 2, Bo et al¹ present the data for the difference between the baseline values for subjects and the values after 6 months of treatment (or no treatment for the control group). They also present the 95% confidence intervals in parentheses. For example, the difference from pretest to post-test for the control group was a mean increase of 0.3 episodes of leakage in 3 days. The exercise group demonstrated a mean decrease of 1.2 episodes of leakage in 3 days. For the control group, the authors reported a confidence interval of -0.5 to 1.1 . An appropriate way to interpret this data is to think that if 100 other control groups were measured using the exact same protocol and measurement techniques, the mean change in episodes of leakage would probably be between -0.5 (a decrease in leakage episodes) to 1.1 (an increase in leakage episodes) for 95 of the samples. For the exercise group, the subjects demonstrated a decrease of 1.2 episodes of leakage over 3 days and if 100 other samples were studied it can be inferred that 95 of the samples would demonstrate a mean decrease of 2.0 to 0.4 episodes of leakage over 3 days.

A clinically relevant use of confidence intervals is to provide clinicians with information that

can assist in the decision of whether or not to incorporate a treatment into a plan of care. Every clinician should be aware of the concept of statistical significance versus clinical significance. Lang and Secic⁸ wrote, "Statistical significance essentially reflects the influence of chance on the outcome; clinical importance reflects the biological value of the outcome." They continued, "In general, small differences between large groups can be statistically significant but clinically meaningless." Alternatively, statistical tests might also show that treatments did not result in statistically significant improvements. This could be due to the fact that the change was not greater than that which was due to chance alone. Another possible reason is that the researchers simply did not have enough subjects in the study. Portney and Watkins⁴ wrote, "Smaller samples are less likely to be good representations of population characteristics, and therefore true differences between groups are less likely to be recognized." For example, looking at the data from the research by Bo et al,¹ do you believe that a reduction of 0.4 to 2.0 leakage episodes over 3 days is a great enough decrease that this will be meaningful to the person receiving the treatment? What about the change after using the vaginal cones? How do you feel about recommending training with the vaginal cones when the data from this study show that the mean change could be a decrease of 1.2 episodes to an increase of 2.8 episodes? We believe that this is an extremely important concept to remember. While properly executed statistical tests are important for the purpose of rejecting or not rejecting the null hypothesis(es), we believe that a clinician must pay particular attention to the amount of change that occurs as the result of a treatment. A clinician must determine if the expected result of a treatment is going to be meaningful to the patient/client. This, in our opinion, is best determined by very simply looking at the means, standard deviations and confidence intervals of the data set, versus just the result of a statistical test.

Still within the statistical methods section of the article, the authors list the statistical procedures they used to analyze the data. These include the Kruskal-Wallis analysis of variance, the Mann-Whitney U test, and the Cochran-Mantel-Haenszel tests. While we believe that it is crucial for clinicians to understand how to interpret the results of each statistic, we believe that it is the responsibility of the editorial board of each journal to evaluate the appropriateness of the analysis procedure.

The authors state a point that we believe deserves our explanation. They wrote, "As several variables were not normally distributed, however, the Kruskal-Wallis analysis of variance

was chosen as the global test of differences between groups on visual analogue scales and other interval scaled variables." What does it mean when a variable is said to be normally distributed? This concept is usually introduced early in a statistics course and is a very important consideration when deciding which statistical methods will be employed. When each score for a data set is placed on a frequency histogram, if the variable is normally distributed, that graph would take the shape of a smooth and symmetrical bell with the greatest number of scores around the mean. Those values that occur less frequently would be to the left or right of the mean depending on whether they were less than or greater than the mean. In a normal distribution, 68% of the scores would be found within one standard deviation below the mean and one standard deviation above the mean, and 95% of the scores would be found within two standard deviations above and below the mean. In addition, the mean, median and mode would all have the same value.

Huck⁹ instructed that there are 4 underlying assumptions about the sample and the population that must be met in order to make correct decisions regarding the statistical analysis techniques that will be used to evaluate the results of data analysis. He wrote, "If one or more of these assumptions are violated, then the probability statements attached to the statistical results may be invalid." The assumptions include: (1) that the subjects are a randomly selected sample of the population, (2) that each subject's scores are not influenced by what happens to other subjects in the study (independent samples), (3) that the variable(s) being studied is (are) normally distributed in the population, (4) the variance is the same in each group included in the study (homogeneity of variance). There are specific mathematical methods to ensure that the assumptions of normality of distribution and homogeneity of variance. When data are proven to have a normal distribution, parametric tests are employed to analyze the data. Portney and Watkins⁴ wrote that clinical research often involves studying variables that have not been studied enough to support the assumptions of population normality and homogeneity of variance. They stated that, "In all likelihood, most pathological conditions are represented by skewed distributions rather than symmetrical ones. In addition, small clinical samples and samples of convenience cannot automatically be considered representative of larger normal distributions."

Bo et al¹ inform the readers that because several variables were not normally distributed, they analyzed their data with the nonparametric test called the Kruskal-Wallis Analysis of Variance (ANOVA). They stated that they used the Kruskal-

Wallis ANOVA for the visual analogue scale data and the "...other interval scaled variables." Interval data include measurements on a continuous scale that have equal distances (intervals) between any 2 units of measurement (versus data which might require only yes-no answers, blood types or categories such as better-same-worse). Examples from this study include the pad test with standardized bladder volume and the 24-hour pad test. In order to perform the Kruskal-Wallis ANOVA all of the data points from each group are combined into a single group. Each data point is then ranked in terms of performance from the smallest to the largest. Next, the individual groups are re-established with their ranked data points. The sum of each group's ranks is next placed into a formula which results in a numerical value, the H statistic. The reporting of this H statistic is common in research articles, and we believe that it is a good practice to place this statistic in the report. Bo et al¹ only listed the p value in Table 2.

The next step of the analysis is to obtain (from textbook tables or a computer program) the probability value (the p value) associated with the specific H statistic. The p value indicates the probability that the results obtained were due to chance or sampling error. Hicks¹⁰ wrote, "... the smaller the p value, the smaller the possibility that random error or chance factors can account for your results." It follows then that the smaller the probability of your results occurring by chance, the greater is the probability that the result is due to treatment. However, Bailar and Mosteller¹¹ caution, "Although small P values may make chance an unlikely explanation for differences between 2 groups, they do not necessarily imply that the differences are due to the therapy or exposure. Instead, the differences may result from other characteristics of the groups. The tendency of groups within a study to respond differently because of such noncomparability is called a nonrandom error or a nonsampling error or bias." Probability values may be written in text as a percentage (1%, 5%) or as a decimal (0.01, 0.05). A value of 5% or 0.05 indicates that there is a 5% chance that the results obtained were due to chance or random error. It should be noted that some authors will use a small p to indicate probability and others will use a capital P. This is dictated by the writing style manual chosen by the journal in which the article is placed. That is, some journals ask authors to adhere to the format and reference style of the *American Medical Association (AMA) Manual of Style*,¹² whereas other journals require authors to write in the style outlined in the *Publication Manual of the American Psychological Association*.¹³

In part I of this series we outlined that the purpose of most research is to test a proposed hypothesis. Riegelman² wrote, "Statistical significance testing or hypothesis testing assumes that only 2 types of relationships exist. Either differences between groups within a study population exist or they do not exist. When we conduct statistical significance tests on study data, we assume at the beginning that no such differences exist in the population." The statement that there are no differences between 2 or more groups as a result of treatments administered, other than from chance or random error, is an example of a null hypothesis. Many projects are undertaken with the hope of being able to reject the null hypothesis which would allow for the conclusion is that it is more likely the case that the treatment was the cause of the significant differences observed between the groups. Lang and Secic⁸ wrote that a small p value is stronger evidence against the null hypothesis. That is, if the p value is equal to, or smaller than, the level of significance selected prior to the collection of data (the alpha level) then the results are determined to be significant. This means that the null hypothesis can be rejected. Written another way, Phillips¹⁴ stated that the probability calculated is "...the probability that we are rejecting a true null hypothesis..." So, when the p value associated with the H value calculated in the Kruskal-Wallis analysis is smaller than the predetermined level of significance, the null hypothesis can be rejected.

Portney and Watkins⁴ describe the premise behind that Kruskal-Wallis test as "If the null hypothesis is true, we would expect an equal distribution of ranks..." It is important to understand that the Kruskal-Wallis ANOVA is designed only to indicate whether there are differences that exist between the groups. The analysis does not tell which group had the better result. Huck⁹ wrote, "The post hoc procedure used most frequently following a statistically significant H test is the Mann-Whitney U test." This was, indeed, the next analysis the authors performed. They used the Mann-Whitney U Test (a nonparametric test) to compare each group with the control group and also each of the pelvic floor muscle exercise, electrical stimulation, and vaginal cone groups with each other. This procedure involves combining the 2 groups being analyzed and ranking the scores from smallest to largest. The sum of the ranks for each group is then calculated. Portney and Watkins⁴ explain, "Under the null hypothesis, we would expect the groups to be equally distributed with regard to high and low ranks, and the sum of the ranks would be fairly equal for both groups." They continue, "The test will determine if the differ-

ence between the sums of ranks is sufficiently large to be considered significant." For the significance test, the Mann-Whitney statistic is calculated and checked against a table called the Critical Values for the Mann Whitney U test. This results in the recording of the probability (p) value from which the researcher can determine if significant differences exist between the 2 groups being compared.

The final statistical procedure that the authors used was the Cochran-Mantel-Haenszel tests or X^2 tests (also known as the Mantel-Haenszel chi square statistic). Chi square tests are used with nominal or categorical data. It is helpful to understand that nominal or categorical data include those responses that can be named or placed into unordered categories. Examples include political parties, blood types, clinical diagnoses, and pass/fail indications. The nominal data from this investigation includes the subjective assessment by each subject of her perceived condition before and after treatment. That is, each subject was asked to rate the outcome of her treatment given the options of continent, almost continent, improved, unchanged, or worse. The Mantel-Haenszel chi square statistic allowed the authors to see if there was an association between the responses of the subjects and the treatments they received. Similar to the previous statistical tests, chi square calculations result in a value that is then looked up in tables called the critical values of X^2 . Once again, the table will give the probability value. If the value is <0.05 , it is appropriate to conclude that there is an association between the two variables.

In the last sentence, Bo, Talseth, and Holme informed their readers that they decided the p values less than 0.05 were to be considered significant. Therefore, if any p value derived from the statistical tests are less than 0.05, then the authors concluded that the differences are statistically significant.

Statistical methods

The primary analysis was carried on data from treated participants, with exclusion of data from those without final evaluation on efficacy variables. Additional intention to treat analyses were also done for all randomised patients including those who dropped out. The missing last values were considered as equal to baseline values. Results are given as mean values with 95% confidence intervals. As several variables were not normally distributed, however, the Kruskal-Wallis analysis of variance was chosen as the global test of differences between groups on visual analogue scales and other interval scaled variables. Pairwise comparisons were made with the Mann-Whitney U test to compare each group with the control and one intervention group with another. Cochran-Mantel-Haenszel tests or [chi squared] tests were used if data were nominal or categorical. P values < 0.05 were considered significant.

RESULTS

In this section, the first thing that Bo et al did was to account for the subjects who did not complete the study. We know that 130 subjects

began the study and 107 completed the study. Therefore, the authors must account for the loss of 23 subjects. Trisha Greenhalgh¹⁵ published a tutorial regarding the assessment of the quality of the methods section of a journal article. She wrote, "Simply ignoring everyone who has withdrawn from a clinical trial will bias the results, usually in favor of the intervention."

It is important to be given the reasons for subject dropout in order to evaluate whether they discontinued because of a worsening condition, or perceived ineffectiveness of the treatment etc. The reasons may suggest that a clinician might be wary about using a particular treatment with a patient/client. Greenhalgh continued by saying, "It is, therefore, standard practice to analyze the results of comparative studies on an intention to treat basis." We were informed earlier that the authors did perform the intention-to-treat analysis. The authors did a very nice job, both in the text and in Figure 1, at describing the reasons why only 107 women completed the study.

The authors report that at the outset of the study, there were no significant differences between the groups for all of the variables listed in Table 1. This is essential to being able to draw any meaningful conclusions from the data regarding treatment effectiveness. Each group must start out as similar as possible. Looking at the data presented in Table 1, however, reveals that the standard deviations for the control group (116.1 g) and the vaginal cone group (158.3 g) for the 24-hour pad test are quite large compared to the other 2 groups (15.2 g, 15.5 g). The large standard deviations raises some level of concern regarding the usefulness of this as a measurement tool because of the high variability. In part I we emphasized that researchers must be very concerned about the reliability of their measurement techniques. That is, they must be concerned that their measurements are accurate, consistent, and repeatable. Bruton, Conway, and Holgate¹⁶ wrote, "It is very rare to find any clinical measurement that is perfectly reliable, as all instruments and observers or measurers (raters) are fallible to some extent and all humans respond with some inconsistency. Thus any observed score (X) can be thought of as a function of 2 components, ie, a true score (T) and an error component (E)..." The authors continue to inform that there are several sources of error that may affect a measurement. These include, variable performance of the subject, errors in technique by the person taking the measurement, and uncalibrated equipment. These, and other, sources of error may result in an increase in variability among the scores obtained from subjects within a sample. Therefore, researchers must use outcome measures that have high

reliability coefficients. Portney and Watkins⁴ advised that “As a general guideline, coefficients below .50 represent poor reliability, coefficients from .50 to .75 suggest moderate reliability, and values above .75 indicate good reliability. For most clinical measurements, reliability should exceed .90 to ensure valid interpretations of findings...” Reliability coefficients of new measurements are established by specific types of investigations and analyses. The high variability for the 24-hour pad test is a cause for concern for the usefulness and validity of using the tool as a measure of continence. It is impossible to know if the high variability is due to true response variability in humans, or error variance. The large variability demonstrated in the groups means that it will require even larger differences (before treatment versus after treatment) to declare a statistically significant difference.

In the next paragraph the authors list the compliance each group demonstrated regarding performing the assigned regimen of treatment. They listed the mean and the standard error (SE) in the parentheses. Regarding the standard error, the punch line is that when samples are large, the smaller the SE, the stronger is the sample mean as an estimate of the population mean. However, it is also possible that the SE is small simply due to chance. This is less likely with a large sample size. We can determine from the author’s presentation that the subjects assigned to the pelvic floor muscle-training group were the most compliant in performing the assigned treatment. It is obvious that noncompliance, whether it be in research or a clinic, is an issue that must be considered when designing an intervention protocol. In addition, if one group was more noncompliant than the other groups, readers might question whether that affected the outcome of that group. Indeed, the authors report that the increased amount of compliance for the pelvic floor muscle exercise group was statistically significantly greater than for the other 2 groups. However, there was no difference between the electrical stimulation group and the vaginal cones group.

There is a considerable amount of information provided in the section titled, “Changes after treatment.” The authors present their data in a combination of tables and figures, or in the text. Tables are often used to present numerical or descriptive data and the resulting values from statistical analysis (means, standard deviations, confidence intervals ratios, p values etc.). Regarding figures, the phrase that comes to mind is, “A picture is worth a thousand words.” Figure 2 from this study is a great example of that phrase. It is immediately obvious that the pelvic

floor muscle exercise group had the largest change for muscle strength.

The authors chose to first highlight the change in muscle strength for each group after a 6-week training period. Again, it is easy to see that the exercise group gained the most strength. Although the control group did show some improvement, it was not as large a change. The data from the remaining 2 groups did result in showing that there was also an increase in muscle strength. It is important to note that one cannot determine from the figure if the differences of change within or between groups were statistically significant. In the text, the authors present more detailed information than can be obtained from the figure. For example, the pelvic floor muscle strength in the exercise group improved from a baseline of 11.0 cm H₂O to 19.2 cm H₂O at the end of the study. The vertical line extending up from the bar depicts the upper limit of the 95% confidence interval. The lower limit is found in the text. It is important to note that these extensions from the bars are also typically used by other authors to signify standard deviations or standard errors. It would have made the figure presentation easier to interpret if the authors had identified what was depicted in the table. For example, the figure title could have been “Change in strength of pelvic floor muscles in control group and treatment groups (means and 95% confidence intervals).” In the text the authors inform the readers that the increase in strength of the control group was not statistically significant. They report that the pre-and post-test changes were statistically significant for the 3 treatment groups. However, when the change of strength of each of the 3 groups was compared to the change demonstrated by the control group, only the pelvic floor muscle exercise group data were statistically significant. This means that even though the control groups showed some improvement, the exercise group demonstrated an even greater proportional improvement. Additionally, when statistical tests were completed to compare the results of the electrical stimulation and the vaginal cones groups, they did not differ significantly.

Table 2 presents the results of the Kruskal-Wallis analysis and the means of 5 outcome measures for each group. In this case, the Kruskal-Wallis analysis was used to determine if there were any significant differences between the 4 groups in the amount of change from baseline to final measurements. The last entry on the right is the p value. Again, remember that the authors decided that to be termed a statistically significant result the p value had to be less than 0.05. Readers can see that only 4 of the outcome measures produced p values of less than 0.05.

The analysis for the 24-hour pad test resulted in a p value of 0.684, which is greater than 0.05. So, the null hypothesis could not be rejected for the 24-hour pad test. Note the larger mean and confidence intervals for the vaginal cone group. However, looking back at the baseline data from Table 1, readers can see that that group began with a larger mean and standard deviation. Therefore, there would have had to be a very large change in order to result in statistically significant improvement.

Continuing with the information from Table 2, we are informed that the calculated p values for the between group comparisons were 0.01 for the episodes of leakage in 3 days test, 0.038 for the stress pad test, 0.001 for the leakage index, and 0.001 for the social activity index. Since each of these values is less than 0.05 it was concluded that statistically significant differences did exist between the 4 groups. The values for the 24-hour test was 0.684 which is not less than 0.05 so the null hypothesis could not be rejected.

In Table 3, the authors present the baseline to 6-month differences between each of the 3 treatment groups versus the control group for the 5 outcome measurements. For example, the mean number of episodes of leakage in 3 days from Table 2 was +0.3 (an increase in leakage of 0.3 episodes in 3 days) for the control group and the mean for the exercise group was -1.2 (a decrease in number of episodes). Therefore, the authors calculated that the total number of leakage episodes differed by -1.5 (a difference of -1.2 added to the difference of 0.3 that they theorized would have occurred if they had been in the control group). This calculation allows the reader to see the benefit of each treatment over the control, or nontreatment, group. It can be seen that the exercise group had a greater decrease in episodes of leakage in 3 days as compared to the exercise or vaginal cone group, a greater decrease on the stress pad test, a greater decrease in the number of times they experienced leakage with specific activities and more improvement in participation in social events that the vaginal cone group (tied with the electrical stimulation group). The authors chose to place the p values associated with the comparisons shown in Table 3 in the text. These comparisons were made using the Mann-Whitney tests. For example, the difference between the control group versus the exercise group (-17.5) was statistically significant with a p value of 0.02. The authors continued with the comparisons between the treatment groups using the data from Table 2. While all of the information is presented accurately in the text, it is in our opinion, more difficult to follow. We find it easier to follow if asterisks are placed in the table to

highlight those comparisons for which the groups were significantly different. However, the reader is informed that the improvement in the stress pad test for the exercise group (-30.2) was a statistically significant greater improvement than the improvement in the stress pad test of the electrical stimulation group (-7.4) as measured by the pad test with standardized bladder volume (p value of 0.02) and the leakage index (p value of < 0.01), and also compared with the vaginal cone group for the pad test (p value < 0.01), episodes of leakage over 3 days (p value 0.03), and for the leakage index (p value < 0.01). We are informed that there were no statistically significant differences between the electrical stimulation group and the vaginal cones group. The last sentence is the only time we see a statement regarding the measurement of maximum urethral pressure or maximum closure pressure. We are told that no changes were recorded for either parameter. This would indicate that these 2 variables were not good indicators of continence for the subjects included in this study.

Table 4 contains data from the subject's subjective labeling of improvement. It becomes obvious through observations of the data in the table that the women in the exercise group did rate themselves as being continent, or at least improved, more than the electrical stimulation or vaginal cones group. Indeed, we are informed in the text that this difference was statistically significant.

The authors next reported on the results of the intention-to-treat analysis. We believe it would have been helpful if the authors chose to present differences in the raw data that support their overall findings of the electrical stimulation group being "weaker when compared with the control group," and "improvement on the social activity index..." for the exercise group.

In the final paragraph before the discussion section, the authors listed the adverse side-effects that were reported by the participants. These are relevant to clinicians as all patient/clients must be informed of the possibility of side-effects, and what they might be, before they decide what their treatment protocol will be so that they can give informed consent to treatment.

Results

One hundred and twenty two patients were randomised. Three women could not complete the study (asthma, change of work, and death in the family), and two were excluded because they used other treatments during the trial. Ten women dropped out with motivation problems or adverse effects: two from pelvic floor muscle training (one motivation problem, one because of travel time to the training group), seven from electrical stimulation (two because of pain, one for bleeding, and four through lack of motivation), and one from vaginal cones (vaginal bleeding). This left 107 participants: 30 in the control group, 25 in the pelvic floor exercise group, 25 in the elec-

trical stimulation group, and 27 in the vaginal cones group (Figure 1).

Figure 1. Trial profile of 130 women recruited to study of treatment of stress incontinence

At baseline there were no significant differences between the groups in any of the background characteristics such as age, body mass index, duration of symptoms, pelvic floor muscle strength, urodynamic assessment, or degree of leakage (Table 1).

Table 1. Background and outcome variables before treatment. Values are means (SD) unless stated otherwise

Compliance with treatment

Mean (SE) adherence with treatment was 93% (1.5%) for pelvic floor muscle training, 75% (2.8%) for electrical stimulation, and 78% (4.4%) for vaginal cones. Adherence with pelvic floor muscle training was significantly greater than with electrical stimulation or vaginal cones ($P < 0.001$ and < 0.002 , respectively). The difference between the electrical stimulation and cone groups was not significant.

Changes after treatment

(Figure 2) shows details of the change in strength of the pelvic floor muscles. There was no significant change in the control group, but significant improvement was seen after treatment in the other groups. Only in the pelvic floor exercise group, however, was the improvement significant when it was compared with the control group ($P < 0.01$). The change in the strength of pelvic floor muscle was significantly greater ($P=0.03$) in the pelvic floor exercise group (11.0 cm H₂O (95% confidence interval 7.7 to 14.3) before test v 19.2 cm H₂O (15.3 to 23.1) after test) compared with electrical stimulation (14.8 cm H₂O (10.9 to 18.7) v 18.6 cm H₂O (13.3 to 23.9)) and vaginal cones (11.8 cm H₂O (8.5 to 15.1) v 15.4 cm H₂O (11.1 to 19.7)). There was no difference in changes of strength between the electrical stimulation and vaginal cones groups ($P=0.90$). Intention to treat analyses did not change the results.

Figure 2. Change in strength of pelvic floor muscles in control group and treatment groups

Analysis with Kruskal-Wallis test showed significant differences between groups in changes in all outcome variables except the 24 hour pad test (Table 2). Table 3 shows the difference between active and control treatment in changes from baseline to 6 months with 95% confidence intervals for efficacy variables.

Table 2. Mean change (95% confidence interval) in measures of stress incontinence from baseline to 6 months

Table 3. Differences (95% confidence intervals) between active and control treatment in change in stress incontinence measured by efficacy variables from baseline to 6 months

There were also significant differences in change between the pelvic floor exercise group and the control group according to the results of the pad test with standardised bladder volume ($P=0.02$), episodes of leakage in 3 days ($P < 0.01$), social activity index ($P < 0.01$), and leakage index ($P < 0.01$). The difference between electrical stimulation and control was significant for episodes of leakage in 3 days ($P=0.02$), social activity index ($P < 0.01$), and leakage index ($P < 0.04$) (Table 3). The difference between the vaginal cones group and the control group was significant for social activity index ($P=0.04$) and leakage index ($P=0.02$) (Table 3). The pelvic floor exercise group improved significantly more than the electrical stimulation group measured by pad test with standardised bladder volume (reduction in urine leaked 30.2 g v 7.4 g; difference 22.8 (3.8 to 41.8); $P=0.02$) and leakage index

(0.9 v 0.2 lower; difference -0.7 (-0.4 to -1.0); $P < 0.01$) and significantly more than the vaginal cones group in pad test (reduction in urine leaked 30.2 g v 14.7 g; difference -15.5 (-34.1 to 3.1); $P < 0.01$), episodes of leakage over 3 days (1.2 fewer v 0.8 more; difference -2.0 (-4.0 to 0.1); $P=0.03$), and leakage index (0.9 v 0.3 lower; difference -0.6 (-0.9 to -0.3); $P < 0.01$). There were no significant differences between the electrical stimulation and vaginal cones groups in any outcome variable. There were no significant changes in maximum urethral pressure or maximum closure pressure for any group.

Objective cure (<or=to2 g leakage on the pad test with standardised bladder volume) was achieved by two women in the control group, 11 in the pelvic floor exercise group, seven in the electrical stimulation group, and four in the vaginal cones group ($P = 0.02$).

Subjective cure (number of women stating that the condition was "unproblematic" after the treatment) was reported by one woman in the control group, 14 in the pelvic floor exercise group, three in the electrical stimulation group, and two in the vaginal cones group. When corrected for baseline values the change in the pelvic floor exercise group was significantly greater than the change in the other groups ($P < 0.001$).

(Table 4) shows subjective improvement after intervention. Significantly more women in the exercise group reported being continent or almost continent ($P < 0.001$) than in the other groups. Fourteen of the 30 participants in the control group chose to use the continence guard. Four felt completely dry when wearing the guard, and five felt somewhat better. Three participants in the control group and two in the electrical stimulation group considered themselves worse after treatment.

Table 4. Subjective assessment of improvement in stress incontinence according to treatment

Twenty eight women in the control group, four in the pelvic floor exercise group, 19 in the electrical stimulation group, and 23 in the vaginal cones group wanted further treatment, apart from the one they had been randomised to, after the trial period. The difference between groups was significant in favour of the pelvic floor exercises ($P < 0.001$).

The results according to the intention to treat analysis showed virtually the same results as the treatment analyses. The only group that came out somewhat weaker when compared with the control group was the electrical stimulation group. Only the variable of leakage in 3 days showed nominally significant differences in change from baseline compared with the control group ($P = 0.047$). Improvement on the social activity index also became significant in favour of exercises compared with electrical stimulation.

Adverse effects and treatment tolerance

There were no side effects reported for pelvic floor exercises. In the electrical stimulation group two participants reported smarting (one tenderness and bleeding, one discomfort), and eight women reported motivation problems and difficulties in using the stimulator. Of those participants who used vaginal cones, one reported abdominal pain, two vaginitis, and one bleeding, and 14 reported motivation problems and trouble in using the cones.

DISCUSSION

After reading the method and results section, the reader has enough information to judge the validity of the study. Critical questions guide the reader's ability to synthesize the study. From *Foundations of Clinical Research Application to Practice*, 2nd ed by Portney and Watkins⁴ these questions assist the reader in evaluating the whole study.

- How does the author interpret the results?
- Does the author consider alternative explanations for the obtained findings?
- Are these discussions supported by the literature?
- Does the author identify limitations to the study?
- If results are not significant, does the author consider the possibility of a Type II error?
- Regardless of the statistical outcome, are the results clinically important?
- Does the author present suggestions for further study?
- Do the stated conclusions flow logically from the obtained results?

The author have clearly stated that pelvic floor muscle training was a more effective treatment for genuine stress incontinence than no treatment, electrical stimulation treatment, or treatment with vaginal cones. Bo et al¹ has made the argument clearly through the results section. The ultimate purpose of any study is to produce results that can be applied to the population of interest.⁸

In order to generalize this study to the larger population we need to know if this is an explanatory or pragmatic study. An explanatory study or an efficacy study is controlled by the need to understand a disease or therapeutic process. These studies are conducted under conditions that allow for tight control over patient selection, treatment and follow-up. Conversely, pragmatic studies, also called effectiveness studies, are dominated by the need to make clinical decisions.

Pragmatism comes from the American movement in philosophy founded by CS Peirce and William James. Their doctrine stated that the meaning of concepts is to be "sought in their practical bearings, and that the function of thought is to guide action and that truth is preeminently to be tested by the practical consequences of belief."¹⁷

These pragmatic studies are conducted under conditions that are reflective of the circumstances which best replicate contemporary theory and medical care. The author states clearly that this was a pragmatic study reflecting current practice. The reader is likely to read this study to see if this research would change their own treatment parameters.

Bo follows an important style in the discussion section which is to compare and contrast other studies of related work and offer support for existing clinical theory. Additionally, she suggests an alternative theory as to the success of voluntary contractions by weaving in a previous study that she did in 1997 which measured urethral pressure with a pelvic floor muscle contraction.

Readers should be encouraged to refer to the reference section to understand the context of the author's comments. For completeness here we referred to the 1997 article mentioned by Bo.¹⁸ It is not necessary to always go back to the article of reference unless the reader wishes to trace the theory back to its origin. In this case here there was a commentary on the study and the measurement techniques used were not specific to pelvic floor muscle training. The issue of measurement tools is critical to our collective understanding of the outcome of a study. If an outcome measurement has not been developed, then using an existing one and reporting its effectiveness is a first step in developing measurement tools that can be created and used for specific treatments. The researcher must look for a technique that already has established reliability. If there is none, then the investigator needs to develop a tool and test its reliability before using it as an outcome measure. In order to make clinical judgements based on outcome measures, the outcome measures must be reliable and be able to be repeated with confidence by different researchers so that accuracy is ensured.

Bo poses that current imaging techniques are limiting and future ones may improve the practitioner's ability to correctly diagnose. Other limitations to the study appear in the clinical commentary section. Bo responds to the comments there. Refer to that section later in this paper.

The study does flow logically from the stated conclusions. Missing from the discussion is the author's suggestions on a direction that future studies might take. The need for future studies are not stated in this article but one can be sure that this active researcher (Bo) is busy on other relevant studies as evidenced by the volume of her previous work. The authors leave the reader to wonder about what the impact of this research could have on clinical practice. A complete discussion that would highlight the ramifications on clinical practice is missing.

Discussion

To our knowledge, this is the first study to compare three of the most commonly used conservative treatments with no treatment for genuine stress incontinence. We have shown that pelvic floor muscle training was a more effective treatment for genuine stress incontinence than no treatment, electrical stimulation, or vaginal cones. Compared with women in the control group only women in the pelvic floor exercise group increased pelvic floor muscle strength and reduced urinary leakage significantly when it was measured by pad test with standardised bladder volume. In addition, significantly more women in the pelvic floor exercise group stated that after the intervention the condition was no longer a problem.

Pragmatic Study

This was a pragmatic study reflecting current practice. The intention was to give the optimal treatment in each group on the basis of current theory and recommendations. Because the exercise group met once a week for

training the women had more attention than those in the two other treatment groups. Electrical stimulation and vaginal cones, however, are advertised as treatments that patients can undertake at home after introduction by health staff. In an attempt to give equal individual attention and motivation all participants met once a month for individual follow up by a skilled physical therapist. On the other hand, both the electrical stimulation group and cone group spent more time per day with the treatment than the exercise group (30, 20, and less than 10 minutes, respectively). From the present study we cannot conclude which part of the three treatment packages caused the results. A decision to exclude the control group from monthly visits to measure strength of the pelvic floor muscles was taken to prevent this acting as a stimulus for training—that is, the "avis effect." The electrical stimulation and vaginal cones groups were not protected from this effect either, although they were specifically asked not to undertake pelvic floor exercises during the trial.

The use of this structured programme of pelvic floor exercise has previously been reported to be more effective than exercise carried out just at home. [11] Our results confirm that such a programme is more effective than no treatment for genuine stress incontinence, as have other well designed randomised controlled studies. [17,18]

The finding that pelvic floor exercises are more effective than electrical stimulation confirms the results of Henalla et al, who found pelvic floor exercise was more effective than electrical stimulation or oestrogen therapy in the treatment of genuine stress incontinence, [18] but contrasts with other studies that did not find any differences in outcome between the two interventions. [9,10] These studies, however, were of small samples, and non-significant results may be due to type II error. The effectiveness of the exercise regimens used can also be questioned. [19] Interestingly, two well designed, randomised controlled trials that compared electrical stimulation with sham electrical stimulation (placebo) have shown conflicting results, [20,21] and Brubaker et al found a 49% cure rate after electrical stimulation for urge incontinence but no effect for genuine stress incontinence. [22]

That regular exercises seem to be more effective than electrical stimulation is not surprising from a physiological perspective. Several consensus statements have concluded that electrically stimulated muscle contractions in humans are less effective than voluntary contractions for strengthening. [23,24] In addition, Bo and Talseth showed that voluntary contraction of pelvic floor muscle was twice as effective as an electrically stimulated contraction at increasing urethral pressure. [25] By attributing baseline values to all participants who dropped out in an intention to treat analysis the effect of electrical stimulation diminished further.

The theoretical basis of vaginal cones has been questioned. They may produce prolonged isometric contractions of the pelvic floor muscles, and in other muscles this may cause injuries due to overuse. [26] Our results showed pelvic floor exercises to be superior to vaginal cones in increasing muscle strength and reducing urinary leakage.

We found no differences in effect between electrical stimulation and vaginal cones, although both were more effective than no treatment measured by some secondary outcome measures. Many women, however, found electrical stimulation and vaginal cones difficult to use, and adverse effects were reported with both methods. Adverse effects have also been reported by other research groups. [20]

OUTCOME

In the first sentence of this section, Bo et al¹ wrote, "Lack of reproducible and valid tests to measure urinary leakage makes the choice of

outcome measures difficult.” We strongly believe that researchers and clinicians must understand the issues related to reliability (reproducibility) and validity (meaningfulness) if research and clinical techniques are to be credible. Huck wrote that, “the basic idea of reliability is summed up by the word consistency” and that “...the core essence of validity is captured nicely by the word accuracy.”¹⁹ While a review of both concepts is beyond the purpose of this tutorial, we urge all readers to take the time to read and become familiar with these topics. For now you can ask yourselves the following questions.

1. Do you think that incontinence is measured the same way in different clinics?
2. Do you think that you and your colleagues would be consistent in your measurement of pelvic floor muscle strength?
3. Have you been measuring incontinence in the same way Bo, Talseth, and Holme did?
4. Have you found in the literature that researchers from different institutions are measuring incontinence using the same measurement tools?
5. Are the tests and measures you use each day reliable?
6. Do you think that urinary leakage measured after performing jumping jacks is an accurate way to record urinary leakage for women with the goal of remaining continent in their daily lifestyles? Is it a meaningful measurement?

The authors continued to cite 2 sources that have published recommendations for recommended measures of incontinence.^{19,20} We obtained these references and found only definitions of terms and lists of techniques which might be included in the evaluation of a patient/client. There is no description of how a technique should be performed. Indeed, the authors continued by stating, “...there is no agreement about the most appropriate measures to date.” Consequently there are no valid tests to measure urinary leakage. The pad test referred to measures the volume of urine in a sanitary pad that is worn by the patient during rest and activity. This has been shown to be reproducible with a standardized bladder volume so is used here as the primary outcome measure.²⁰

Currently there are no good outcome measures. As clinicians we need good reliable tools and measures to determine outcomes and effectiveness of treatment. Members of the Section on Women’s Health are in the process of working with a documentation company to standardize evaluation, exam, and follow-up documentation so that outcome measures can be developed.

For this study the authors stated that they did include tests which have undergone analysis of reproducibility. We went to the original sources to get the actual reliability data. In the study by Bo et al,¹⁹ they did incorporate a newly designed pad test for the purpose of measuring urine loss. Basically the bladder was filled with saline and the amount was determined individually according to each woman’s cystometric capacity. Each woman was then required to run on the spot for 30 seconds, performing jumping jacks for 30 seconds, perform supine sit-ups for 30 seconds, and then stand up. Unfortunately the only comments from the authors regarding the reliability of this technique is that “This 90 sec pad test has been tested for reproducibility and compared to the International Continence Society (I.C.S.) 1 hour’s test...” They provide a reference for this but the reference is to a paper that was presented at the ICS conference in 1988. However, Lose et al²⁰ performed a reliability test on a pad weighing test that involved filling the bladder to 50% of the cystometric capacity, making it a standardized test, and found the reliability coefficient to be 0.97.

The second reference from Bo, Talseth, and Holme is to a study performed by Bo.²¹ In this study her goal was to establish the reliabilities of the leakage test and the social activity index. She reported the respective correlation coefficients to be 0.92 and 0.94. Her conclusion was that “both the leakage index and the social activity index proved to be reproducible methods to quantify how women experience SUI [stress urinary incontinence].”

We want to make a very strong point regarding the design and use of a new measurement tool by both researchers and clinicians. While we do not want to diminish the value of creativity and innovation, it is absolutely crucial that the first research completed must be to establish the reliability of any new tests and techniques. If a test is not consistent or repeatable, why should anyone trust any of the data generated using that measurement tool? If a test is not reliable, it cannot be judged to be valid.

The authors make a very appropriate comment regarding the fact that the standardized pad test required activity that may not be encountered in women’s daily lifestyles. The utility and meaningfulness of an outcome measure should perhaps be based upon the satisfaction obtained by patient’s real daily activities versus those contrived in a laboratory or clinical setting.

The authors do pose that 56% may not be a satisfactory cure rate although it is clinically significant in the strict sense of the definition. Because these women do not go on to surgery then they suggest that is acceptable. A final

important consideration posed by the authors is the problem of being able to determine which treatments are effective for specific diagnoses. Their example is that perhaps some women diagnosed with stress urinary incontinence may actually have ligament, fascial, or nerve damage. By unknowingly including women with these diagnoses in a study, the true effectiveness of a treatment protocol may be masked.

Outcome

Lack of reproducible and valid tests to measure urinary leakage makes the choice of outcome measures difficult. The Urodynamic Society and the standardisation committee of the International Continence Society have recommended use of measures for urinary leakage and self report to evaluate treatment effect, [27] although there is no agreement about the most appropriate measures to date. Because of the need for inclusion of randomised controlled trials in future meta-analyses we used a range of outcome measures used in clinical practice and research that have been previously tested for reproducibility. [11,14] A pad test with standardised bladder volume was chosen as the primary outcome measure because it has been shown to be more reproducible than pad tests with no standardisation. [28] The pad test used in our study, however, entailed movements likely to cause leakage. Some women who leak urine with this test may consider themselves subjectively cured. Few women may include such rigorous physical activity as part of their everyday life. Therefore an outcome that assesses how problematic incontinence is during daily life may be the most appropriate measure of cure.

It can be questioned as to whether 56% is a satisfactory cure rate. As these women do not then need surgery we suggest that this is highly acceptable. Although all the women who participated in this study had genuine stress incontinence, there can be several causes for this condition. DeLancey considers that if the cause is rupture of ligaments or fascias or severe peripheral nerve damage training may not be effective. [29] Future imaging techniques may improve our ability to give a more specific diagnosis and thus improve the results of conservative treatments.

So far there are no long term results available from this study and as participants were offered other treatment options after cessation of the trial a follow up study will be difficult. A 5 year follow up of a previous clinical study that used the same pelvic floor exercise programme has been published. [30] Five years after cessation of the organised training three of 23 women had been treated surgically; of the 20 remaining, 14 were still satisfied with the results of pelvic floor exercises and did not want other treatment, 15 had no leakage on urodynamic cough test, and 14 were still doing pelvic floor exercises once or more a week. [30] No long term follow up data after cessation of treatment have so far been published for electrical stimulation and vaginal cones.

CONCLUSION

The conclusion should be limited to the conclusions that are supported by the results of the study that can be based on fact and logic not conjecture or speculation.⁸ Here the logical flow of thought from the discussion and outcome leads to the concluding statement. The conclusions should be fully based on the authors interpretation of the findings and results. The authors conclude that exercise should be offered as a first choice for treatment of urinary stress incontinence.

The reader decides if this study is relevant to their practice and if what is learned here will alter the way that they practice. If you as a practitioner have your patients contract their pelvic floor muscles during electrical stimulation and/or cone use, then you may question whether or not this study is pertinent to your practice.

The remainder of the article explains that Dr Kari Bo was the main investigator and what the other authors' roles were. Funding was from 4 sources: the Norwegian Fund for Postgraduate Studies in Physiotherapy, Norwegian Research Council, Coloplast AS (continence guards), and Vitacon AS (electrical stimulators and cones). The authors report that there were no competing interests. So we can see that there was no conflict for the investigators, they did not work for the funding sources, or design any of the equipment that was used. This assures us that the investigation was not unduly biased.

Key messages

- Training to increase the strength of pelvic floor muscles was superior to electrical stimulation and vaginal cones in treatment of genuine stress incontinence
- Adverse effects were reported with use of electrical stimulation and vaginal cones but not with exercises
- Patients' tolerance for electrical stimulation and vaginal cones was low
- Pelvic floor exercise should be first choice of treatment for genuine stress incontinence

Conclusion

Pelvic floor exercises are more effective than electrical stimulation, vaginal cones, and no treatment for women with genuine stress incontinence. As such exercise seems to be safe and effective it should be offered as first choice of treatment for genuine stress incontinence.

In addition to the authors the Norwegian Pelvic Floor Study Research Group comprised Ruth Dyresen, Tom Engebretsen, Hanne Borg Finckenbagen, Magne Halvorsen, Anke Helgar, Kjersti Dybvig Jensen, Marit Nicolaisen, Anne Sophie MacLeod, Ann Brit Sangvik, Latis Sadek, Hjalmar Schiotz, Trine Lise Urnes, and Bjorg Wandas (project secretary). E Jean Hay-Smith has given valuable help with the English revision of the manuscript.

Contributors: KB was the main investigator. She initiated and planned the study, supervised the physicians and physical therapists, administered the whole trial, and wrote the manuscript. TT was in the planning group with KB from the beginning of the study. He was responsible for the planning and administration of the inclusion and exclusion criteria and was head of all urodynamic investigations. He supervised the other physicians in the inclusion and exclusion procedures, urodynamic assessment, and pad testing. In addition, he was the physician in one of the five centres and assessed the patients himself. He has revised the manuscript thoroughly several times, specifically the results from urodynamic investigations. IH advised on the study design and was responsible for stratification and randomisation procedures and planned and supervised all the statistical analysis. He carried out the more advanced statistical analyses, thoroughly revised the manuscript several times, and wrote details on statistical analyses. All participants in the research group contributed throughout the trial period with inclusion, exclusion, assessments, and treatments. KB is the guarantor of the study.

If the vaginal cones were great it may have brought a conflict as they funded the article.

Funding: Norwegian Fund for Postgraduate studies in Physiotherapy and Norwegian Research Council. Coloplast AS provided the continence guards and Vitacon AS provided the electrical stimulators and cones. They also gave financial support to seminars for the research group.

Competing interests: None declared. (Accepted 31 December 1998)

COMMENTARY

A commentary is an opportunity for an expert in the field of the research to comment on the research article and offer a critique. Customarily the author responds to the commentary and that is the format that we see used here.

The Department of Urology at King's College in London offers a commentary on some possible flaws in the study. The major points taken from this commentary are that this study may not accurately reflect current practice in the instruction of voluntary muscle contraction with the use of electrical stimulation and cones. Voluntary contraction along with the use of electrical stimulation from previous studies contradicts the results that Bo et al found.^{22,23} Additionally, the time that the exercise group spent with the therapist was more than the other groups and that may have put undo influence over the results.

Bo, et al respond that they were testing the effect of the electrical stimulation and the cones purely without the added confounding factor of voluntary muscle contraction. They offer reasonable support to the argument as to style of instruction as being consistent with the manner in which patients are trained in Scandinavian countries.

In conclusion, it is up to the reader to evaluate for herself if the research article has influence on the manner in which she practices. If only the abstract, findings, and conclusion were read then it is easy to see how the complete detail of the article could be lost. If the discriminating reader knows what is being studied, how it is researched, the tests that are used and the interpretation of the results, then they will have an understanding of how to best to treat patients. Hopefully, more clinicians will assist researchers for collaboration of future studies.

ACKNOWLEDGEMENT

We express our gratitude to Chris A. McGibbon, PhD at the Massachusetts General Hospital Institute of Health Professions and Michael Fillyaw PT, MS at the University of New England for their assistance in reviewing this paper and offering helpful critique and suggestions.

Conservative management of genuine stress incontinence in women: Study's flaws may be misleading. BMJ: Knullar: BMJ, Volume 319(7203). July 17, 1999.190-191.

Knullar, Vik; Salvatore, Stefano; Bidmead, John; Anders, Kate; Cardozo, Linda

EDITOR-Bo et al's study comparing the various methods used in the conservative management of genuine stress incontinence has several flaws, which may mislead readers. [1] The study has been described as pragmatic, reflecting current practice. This view is undermined by the instructions to the women in the vaginal cone and electrical stimulation groups not to perform pelvic floor exercises while using their treatments; this does not reflect current clinical practice. In a prospective randomised study comparing the efficacy of pelvic floor exercises in combination with vaginal cones and pelvic floor exercises alone the combination of the two treatments was significantly more efficacious than either alone. [2]

In a prospective randomised study comparing the efficacy of pelvic floor exercises in combination with vaginal cones, vaginal cones alone, and vaginal electrical stimulation alone, again the combination of two techniques produced greater improvement in urinary incontinence. [3]

We are also concerned about the differing numbers of visits to a therapist for each group. The pelvic floor exercise group had weekly visits whereas the other groups were seen monthly. This would introduce bias; Wyman et al proposed that the specific conservative treatments are not as important as having frequent contact with the patients, with education and counselling. [4] Thus the pelvic floor exercise group should have a better response to treatment owing to the increased time they had with a therapist.

Vik Knullar Subspecialty trainee in urogynaecology

Stefano Salvatore Research fellow

John Bidmead Research fellow

Kate Anders Urogynaecology nurse specialist

Linda Cardozo Professor of urogynaecology

Department of Urogynaecology, King's College Hospital, London SE5 9RS

REFERENCES

1. Bo K, Talseth T, Holme I. Single blind, randomised controlled trial of pelvic floor exercises, electrical stimulation, vaginal cones, and no treatment in management of genuine stress incontinence in women. *BMJ* 1999;318:487-93. (20 February.)
2. Haken J, Benness C, Cardozo L, Cutner A. A randomised trial of vaginal cones and pelvic floor exercises in the management of genuine stress incontinence. *Neurourol Urodyn* 1991;10:393-4.
3. Wise BG, Haken J, Cardozo L, Wise BG, Plevnik S. A comparative study of vaginal cone therapy, cones and Kegel exercises and maximal electrical stimulation in the treatment of female genuine stress incontinence. *Neurourol Urodyn* 1993;12:436-7.
4. Wyman JF, Fantl JA, McClish DK, Bump RC. Comparative efficacy of behavioral interventions in the management of female urinary incontinence. *Am J Obstet Gynecol* 1998;179:999-1006.

Conservative management of genuine stress incontinence in women: Study's flaws may be misleading: Authors' reply BMJ; Volume 319(7203). July 17, 1999.191 Bo, Kari; Holme, Ingar; Talseth, Trygve

Norwegian Centre for Physiotherapy Research and Norwegian University of Sport and Physical Education, PO Box 4014, Ullevål Stadion, 0806 Oslo, Norway Consultant urologist (Talseth) National Hospital of Norway, Oslo

EDITOR-We tried to give the best possible treatment for all groups within a pragmatic setting. Thus the electrical

stimulation and vaginal cones groups had 30 and 20 minutes' training daily, respectively, whereas the exercise group had less than 8-10 minutes' training. This should favour the electrical stimulation and vaginal cones groups, and it is strange that Khullar et al do not mention this as a flaw. Another flaw that works against the exercise group is that both the vaginal cones and electrical stimulation groups had individual treatment with direct proprioception to the pelvic floor, while the exercise group was taught without proprioception.

If the women are contracting simultaneously with the current, how can we then conclude that it is the electrical stimulation and not the voluntary contraction that gives the effect? The point of this study was to evaluate the effect of electrical stimulation and cones. An interesting hypothesis is whether contraction simultaneously with electrical stimulation gives better results than contraction without. This should be investigated in a future study. Other studies have shown no significant additional effect of adding electrical stimulation to exercise. [1]

Strong motivation and instruction are important factors in increasing muscle strength and part of strength training regimens. One of the benefits advocated by manufacturers of vaginal cones and electrical stimulators is that these methods can be used at home without the therapist, thus being cheaper. These methods have been used in this way in the Scandinavian countries for years. In our study all groups had the same monthly visits, for motivation, individual follow up, and contact with the therapist. The exercise group had weekly contacts in groups in addition. This may have enhanced their improvement. This is the way we teach pelvic floor muscle exercise, and it is difficult to understand how this group contact could give such huge differences in a provocation test at the office of a blinded investigator.

Khullar et al give references to their own work presented as two abstracts. As far as we can see their results are similar to our findings. In the first study exercise and vaginal cones did not give significantly different results. However, the drop out rate in the vaginal cone group was 25%, and no intention to treat analysis was performed. In the second study, adding pelvic floor muscle exercise to treatment with vaginal cones was more effective than treatment with vaginal cones alone.

Kari Bo Exercise scientist karib@brage.idrettsbs.no

Ingar Holme Professor Norwegian Centre for Physiotherapy Research and Norwegian University of Sport and Physical Education, PO Box 4014, Ullevål Stadion, 0806 Oslo, Norway

Trygve Talseth Consultant urologist National Hospital of Norway, Oslo

REFERENCES

1. Knight S, Laycock J, Naylor D. Evaluation of neuromuscular electrical stimulation in the treatment of genuine stress incontinence. *Physiotherapy* 1998; 84:61-71. [CINAHL Link] Accession Number: 00002591-199907170-00054

REFERENCES

1. Bo K, Talseth T, Holme I. Single blind, randomized controlled trial of pelvic floor exercises, electrical stimulation, vaginal cones, and no treatment in management of genuine stress incontinence in women. *BMJ*. 1999;318:487-493.

2. Riegelman RK. *Studying a Study and Testing a Test: How to Read the Medical Evidence*. 4th ed. Philadelphia, Pa: Lippincott Williams & Wilkins, 2000.
3. Lilienfeld AM, Lillienfeld DE. *Foundations of Epidemiology*. 2nd ed. New York, NY: Oxford University Press; 1980.
4. Portney LC, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Upper Saddle River, NJ: Prentice-Hall Inc.; 2000.
5. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ*. 1986;292:746-750.
6. Bulpitt CJ. Confidence intervals. *Lancet*. 1987;1:494-497.
7. Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys Ther*. 1999;79:186-195.
8. Lang TA, Secic M. *How to Report Statistics in Medicine*. Philadelphia, Pa: American College of Physicians; 1997.
9. Huck SW. *Reading Statistics and Research*. 3rd ed. New York, NY: Addison Wesley Longman Inc; 2000.
10. Hicks CM. *Research for Physiotherapists: Project Design and Analysis*. 2nd ed. New York, NY: Churchill Livingstone; 1995.
11. Bailar JC, Mosteller F. *Medical Uses of Statistics*. 2nd ed. Boston, Mass: Massachusetts Medical Society; 1992.
12. American Medical Association. *Manual of Style*. 8th ed. Baltimore, Md: William & Wilkins; 1989.
13. American Psychological Association. *Publication Manual*. 4th ed. Washington, DC: American Psychological Association; 1994.
14. Phillips JL. *How To Think About Statistics*. 6th ed. New York, NY: W.H. Freeman and Company; 2000.
15. Greenhalgh T. How to read a paper: assessing the methodological quality of published papers. *BMJ*. 1997;315:305-308.
16. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? *Physiotherapy*. 2000;86:94-99.
17. Webster's Seventh New Collegiate Dictionary. Springfield, Mass: G. and C. Merriam Company; 1967.
18. Bo K, Talseth T. Change in urethral pressure during voluntary pelvic floor muscle contraction and vaginal electrical stimulation. *Int Urogynecol J*. 1997;8:3-7.
19. Bo K, Hagen RH, Kvarstein B, Jorgensen J, Larsen S. Pelvic floor muscle exercise for the treatment of female stress urinary incontinence. III: Effects of two different degrees of pelvic floor muscle exercise. *Neurourol Urodyn*. 1990;9:489-502.

20. Lose G, Rosenkilde P, Gammelgaard J, Schroeder T. Pad weighing test performed with standardized bladder volume. *Urol*. 1988;32:78-80.
21. Bo K. Reproducibility of instruments designed to measure subjective evaluation of female stress urinary incontinence. *Scand J Urol Nephrol*. 1994;28:97-100.
22. Haken J, Benness C, Cardozo L, Cutner A. A randomised trial of vaginal cones and pelvic floor exercises in the management of genuine stress incontinence. *Neurourol Urodyn*. 1991;10:393-394.
23. Wise BG, Haken J, Cardozo L, Wise BG, Plevnik S. A comparative study of vaginal cone therapy, cones and Kegel exercises and maximal electrical stimulation in the treatment of female genuine stress incontinence. *Neurourol Urodyn*. 1993;12:436-437.

References from Bo article

1. Abrams P, Blaivas JG, Stanton SL, Andersen JT. The standardisation of terminology of the lower urinary tract function. *Scand J Urol Nephrol Suppl* 1988;114:5-19.
2. Fand J, Newman D, Colling J, DeLancy JOL, Keelys C, Loughery R, et al. Urinary incontinence in adults: acute and chronic management. 2nd update: Rockville, Maryland: Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, 1996; (Clinical Practice Guideline, 96-0682.)
3. Norton P, MacDonald LD, Sedgwick PM, Stanton SL. Distress and delay associated with urinary incontinence, frequency, and urgency in women. *BMJ* 1988;297:1187-9.
4. Hunskaar S, Vinsnes A. The quality of life in women with urinary incontinence as measured by the sickness impact profile. *J Am Geriatr Soc* 1991;39:378-82.
5. Bo K, Hagen R, Kvarstein B, Larsen S. Female stress urinary incontinence and participation in different sport and social activities. *Scand J Sports Sci* 1989;11:117-21.
6. Nygaard I, DeLancey JOL, Arnsdorf L, Murphy E. Exercise and incontinence. *Obstet Gynecol* 1990;75:848-51.
7. Bouchard C, Shephard R, Stephens T, eds. *Physical activity, fitness, and health. International proceedings and consensus statement*. Champaign: Human Kinetics Publishers, 1994.
8. Kegel AH. Progressive resistance exercise in the functional restoration of the perineal muscles. *Am J Obstet Gynecol* 1948;56:238-49.
9. Bergbman LC, Hendricks HJ, Bo K, Hay-Smith EJ, de Bie RA, van Waalwijk van Doorn ES. Conservative treatment of stress urinary incontinence in women. A systematic review of randomized clinical trials. *Br J Urol* 1998;82:181-91.
10. Bo K. Effect of electrical stimulation on stress urinary incontinence. Clinical outcome and practical recommendations based on randomized controlled trials. *Acta Obstet Gynecol* 1998;77(suppl 168):3-11.
11. Bo K, Hagen RH, Kvarstein B, Jorgensen J, Larsen S. Pelvic floor muscle exercise for the treatment of female stress urinary incontinence. III: Effects of two different degrees of pelvic floor muscle exercise. *Neurourol Urodyn* 1990;9:489-502.

12. Thysen H, Lose G. Long-term efficacy and safety of a disposable vaginal device (continence guard) in the treatment of female stress incontinence. *Int Urogynecol J* 1997;81:30-3.
13. American College of Sports Medicine. Position stand. The recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness in healthy adults. *Med Sci Sports Exerc* 1990;22:265-74.
14. Bo K. Reproducibility of instruments designed to measure subjective evaluation of female stress urinary incontinence. *Scand J Urol Nephrol* 1994;28:97-100.
15. Bo K. Pressure measurements during pelvic floor muscle contractions: the effect of different positions of the vaginal measuring device. *Neurourol Urodyn* 1992;11:107-13.
16. Bo K, Kvarstein B, Hagen R, Larsen S. Pelvic floor muscle exercise for the treatment of female stress urinary incontinence. II: Validity of vaginal pressure measurements of pelvic floor muscle strength and the necessity of supplementary methods for control of correct contraction. *Neurourol Urodyn* 1990;9:479-87.
17. Lagro-Janssen TLM, Debruyne FMJ, Smits AJA, Van Weel C. Controlled trial of pelvic exercises in the treatment of urinary stress incontinence in general practice. *Br J Gen Pract* 1991;41:445-9.
18. Henalla S, Hutchins C, Robinson P, MacVicar J. Non-operative methods in the treatment of female genuine stress incontinence of urine. *J Obstet Gynaecol* 1989;92:22-5.
19. Bo K. Pelvic floor muscle exercise for the treatment of stress urinary incontinence. An exercise physiology perspective. *Int Urogynecol J* 1995;6:282-91.
20. Sand PK, Richardson DR, Staskin SE, Swift SE, Appell RA, Whitmore KE, et al. Pelvic floor stimulation in the treatment of genuine stress incontinence: a multicenter placebo controlled trial. *Am J Obstet Gynecol* 1995;173:72-9.
21. Luber K, Wolde-Tsodik G. Efficacy of functional electrical stimulation in treating genuine stress incontinence: a randomized clinical trial. *Neurourol Urodyn* 1997;16:543-51.
22. Brubaker L, Benson JT, Bent A, Clark A, Sbott S. Transvaginal electrical stimulation for female urinary incontinence. *Am J Obstet Gynecol* 1997;177:536-40.
23. Dudley GA, Harris RT, Komi PV, eds. Use of electrical stimulation in strength and power training. In: *Strength and power in sport*. Oxford: Blackwell Scientific Publications, 1992;329-37.
24. Boucharde C, Shephard RJ, Stephens T, ed. Physical activity, fitness and health. Consensus statement. In: *Physical activity, fitness, and health: status and determinants. Adjuvants to physical activity*. Champaign: Human Kinetics Publishers, 1993;33-40.
25. Bo K, Talseth T. Change in urethral pressure during voluntary pelvic floor muscle contraction and vaginal electrical stimulation. *Int Urogynecol J* 1997;8:3-7.
26. Bo K. Vaginal weight cones. Theoretical framework, effect on pelvic floor muscles strength and female stress urinary incontinence. *Acta Obstet Gynecol Scand* 1995;74:87-92.
27. Blaiwas J, Appell R, Fanil J, Leach G, McGuire E, Resnick N, et al. Standards of efficacy for evaluation of treatment outcomes in urinary incontinence: recommendations of the urodynamic society. *Neurourol Urodyn* 1997;16:147.
28. Lose G, Rosenkilde P, Gammelgaard J, Schroeder T. Pad weighing test performed with standardized bladder volume. *Urol* 1988;32:78-80.
29. DeLancey J. Stress urinary incontinence: where are we now, where should we go? *Am J Obstet Gynecol* 1996;175:311-9.
30. Bo K, Talseth T. Long term effect of pelvic floor muscle exercise five years after cessation of organized training. *Obstet Gynecol* 1996;87:261-5.