

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. Bergbmans 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.