Nephrol Dial Transplant (2010) 25: 1388–1393
doi: 10.1093/ndt/gfp732
Advance Access publication 12 January 2010

CME Series

NDT
Nephrology Dialysis Transplantation

Sample size calculations: basic principles and common pitfalls

Marlies Noordzij[1], Giovanni Tripepi[2], Friedo W. Dekker[3], Carmine Zoccali[2], Michael W. Tanck[4] and Kitty J. Jager[1]

[1]ERA-EDTA Registry, Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands, [2]CNR–IBIM, Clinical Epidemiology and Pathophysiology of Renal Diseases and Hypertension, Renal and Transplantation Unit, Ospedali Riunti, Reggio Calabria, Italy, [3]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands and [4]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Correspondence and offprint requests to: Marlies Noordzij; E-mail: m.noordzij@amc.uva.nl

Abstract

One of the most common requests that statisticians get from investigators are sample size calculations or sample size justifications. The sample size is the number of patients or other experimental units included in a study, and determining the sample size required to answer the research question is one of the first steps in designing a study.

Although most statistical textbooks describe techniques for sample size calculation, it is often difficult for investigators to decide which method to use. There are many formulas available which can be applied for different types of data and study designs. However, all of these formulas should be used with caution since they are sensitive to errors, and small differences in selected parameters can lead to large differences in the sample size. In this paper, we discuss the basic principles of sample size calculations, the most common pitfalls and the reporting of these calculations.

Keywords: epidemiology; nephrology; power; sample size; statistics

Introduction

The sample size is the number of patients or other experimental units included in a study, and one of the first practical steps in designing a trial is the choice of the sample size needed to answer the research question. Also in the critical appraisal of the results of published trials, evaluating the sample size required to answer the research question is an important step in interpreting the relevance of these results. It is therefore not surprising that one of the most frequent requests that statistical consultants get from investigators are sample size calculations or sample size justifications.

Techniques for sample size calculations are described in most conventional statistical textbooks. However, the wide range of formulas that can be used for specific situations and study designs makes it difficult for most investigators to decide which method to use. Moreover, these calculations are sensitive to errors because small differences in selected parameters can lead to large differences in the sample size.

In this paper, we explain the basic principles of sample size calculations by means of examples from the nephrology literature. In addition, we discuss the most common pitfalls in sample size calculations and comment on how to report these calculations.

Why sample size calculations?

The main aim of a sample size calculation is to determine the number of participants needed to detect a clinically relevant treatment effect. Pre-study calculation of the required sample size is warranted in the majority of quantitative studies. Usually, the number of patients in a study is restricted because of ethical, cost and time considerations. However, if the sample size is too small, one may not be able to detect an important existing effect, whereas samples that are too large may waste time, resources and money. It is therefore important to optimize the sample size. Moreover, calculating the sample size in the design stage of the study is increasingly becoming a requirement when seeking ethical committee approval for a research project.

Components of sample size calculations

In order to calculate the sample size, it is required to have some idea of the results expected in a study. In general, the greater the variability in the outcome variable, the larger the sample size required to assess whether an observed effect is a true effect. On the other hand, the more effective (or harmful!) a tested treatment is, the smaller the sample size needed to detect this positive or negative effect. Calculating the sample size for a trial requires four basic components:

1. The type I error (alpha). Clinical studies are usually performed in a sample from a population rather than in the whole study population. In research, we are testing hypotheses to determine whether (results in) particular samples

Table 1. Overview of errors in clinical research

	Population	
	Difference does not exist	
Difference does not exist		False negative result Type II error (beta)
Difference exists	False positive result Type I error (alpha)	Power (1-beta)

differ from each other. On the one hand, the null hypothesis (H_0) hypothesizes that the groups of subjects (samples) that are being compared are not different, that is they come from the same source population. The alternative hypothesis (H_1), on the other hand, hypothesizes that these groups are different and that therefore they seem to be drawn from different source populations. Sample size calculations are needed to define at what number of subjects it becomes quite unlikely that adding more subjects will change the conclusion.

In the process of hypothesis–testing, two fundamental errors can occur. These errors are called type I and type II errors, and an overview of these errors is presented in Table 1. The type I error (alpha) measures the probability that, given the H_0 that the samples come from the same source population, the differences found are likely to happen. In other words, the alpha represents the chance of a falsely rejecting H_0 and picking up a false-positive effect. The alpha is most commonly fixed at 0.05, which means that the researcher desires a <5% chance of drawing a false-positive conclusion.

2. Power. Instead of a false-positive conclusion, investigators can also draw a false-negative conclusion. In such cases, they conclude that there is no difference between two groups or treatments when in reality there is, or in other words, they falsely accept the H_0 that the compared samples come from the same source population. This is called a type II error (beta). Conventionally, the beta is set at a level of 0.20, meaning that the researcher desires a <20% chance of a false-negative conclusion. For the calculation of the sample size, one needs to know the power of a study. The power reflects the ability to pick up an effect that is present in a population using a test based on a sample from that population (true positive). The power is the complement of beta: 1-beta. So, in case of a beta of 0.20, the power would be 0.80 or 80%, representing the probability of avoiding a false-negative conclusion, or the chance of correctly rejecting a null hypothesis.

3. The smallest effect of interest. The smallest effect of interest is the minimal difference between the studied groups that the investigator wishes to detect and is often referred to as the minimal clinically relevant difference, sometimes abbreviated as MCRD. This should be a difference that the investigator believes to be clinically relevant and biologically plausible. For continuous outcome variables, the minimal clinically relevant difference is a numerical difference. For example, if body weight is the outcome of a trial, an investigator could choose a difference of 5 kg as the minimal clinically relevant difference. In a trial with a binary outcome, for example the effect of a drug on the development of a myocardial infarction (yes/no), an investigator should estimate a relevant difference between the event rates in both treatment groups and could choose, for instance, a difference of 10% between the treatment group and the control group as minimal clinically relevant difference. Even a small change in the expected difference with treatment has a major effect on the estimated sample size, as the sample size is inversely proportional to the square of the difference. For instance, if one would need 1000 subjects to detect an absolute difference of 4.8%, 4000 subjects per treatment group would be required to detect a 2.4% difference.

4. The variability. Finally, the sample size calculation is based on using the population variance of a given outcome variable that is estimated by means of the standard deviation (SD) in case of a continuous outcome. Because the variance is usually an unknown quantity, investigators often use an estimate obtained from a pilot study or use information from a previously performed study. For example, in an echocardiography substudy of the Australian Initiating Dialysis Early And Late (IDEAL) Study, Cooper *et al.* aim to determine whether the timing of dialysis initiation has an effect on left ventricular mass. For their sample size calculation, the investigators used recent data from another laboratory indicating that the mean left ventricular mass in

Table 2. Summary of the components for sample size calculations

Component	Definition
Alpha (type I error)	The probability of falsely rejecting H_0 and detecting a statistically significant difference when the groups in reality are not different, i.e. the chance of a false-positive result.
Beta (type II error)	The probability of falsely accepting H_0 and not detecting a statistically significant difference when a specified difference between the groups in reality exists, i.e. the chance of a false-negative result.
Power (1-beta)	The probability of correctly rejecting H_0 and detecting a statistically significant difference when a specified difference between the groups in reality exists.
Minimal clinically relevant difference	The minimal difference between the groups that the investigator considers biologically plausible and clinically relevant.
Variance	The variability of the outcome measure, expressed as the SD in case of a continuous outcome.

Abbreviations: H_0, null hypothesis; i.e. the compared samples come from the same source population (the compared groups are not different from each other); SD, standard deviation.

renal failure patients in Australia is 140 g/m^2 with an SD of 60 g/m^2.

Sometimes, the minimal clinically relevant difference and the variability are combined and expressed as a multiple of the SD of the observations; the standardized difference. The standardized difference is also referred to as the effect size and can be calculated as:

$$\text{Standardized difference} = \frac{\text{difference between the means in the two treatment groups}}{\text{population standard deviation}}$$

A summary of all components of sample size calculations is presented in Table 2.

How to calculate the sample size for randomized controlled trials

Formulas for sample size calculation differ depending on the type of study design and the studies outcome(s). These calculations are particularly of interest in the design of randomized controlled trials (RCTs). In RCTs, a lot of money is invested, and it is therefore important to be sufficiently sure that enough patients are included in the study arms in order to find as statistically significant a difference that we assume there is in the population. In general, sample size calculations are performed based on the primary outcome of the study.

Based on two examples, we will now demonstrate how to calculate sample size using the simplest formulas for an RCT comparing two groups of equal size. Suppose one wished to study the effect of a new hypertensive drug on (i) systolic blood pressure (SBP) as a continuous outcome and (ii) SBP as a binary outcome, i.e. below or above 140 mmHg (hypertension yes/no). These situations are illustrated in Box 1 and Box 2, respectively [1].

SBP as a continuous outcome

Box 1
Simplest formula for a continuous outcome and equal sample sizes in both groups, assuming: alpha = 0.05 and power = 0.80 (beta = 0.20) [1].

n = the sample size in each of the groups
μ_1 = population mean in treatment Group 1
μ_2 = population mean in treatment Group 2
$\mu_1 - \mu_2$ = the difference the investigator wishes to detect
σ^2 = population variance (SD)
a = conventional multiplier for alpha = 0.05
b = conventional multiplier for power = 0.80

$$n = \frac{2\left[(a+b)^2\sigma^2\right]}{(\mu_1-\mu_2)^2}$$

When the significance level alpha is chosen at 0.05, like in these examples, one should enter the value 1.96 for a in the formula. Similarly, when beta is chosen at 0.20, the value 0.842 should be filled in for b in the formula. These multipliers for conventional values of alpha and beta can be found in Table 3.

Suppose the investigators consider a difference in SBP of 15 mmHg between the treated and the control group ($\mu_1 - \mu_2$) as clinically relevant and specified such an effect to be detected with 80% power (0.80) and a significance level alpha of 0.05. Past experience with similar experiments, with similar measuring methods and similar subjects, suggests that the data will be approximately normally distributed with an SD of 20 mmHg. Now we have all of the specifications needed for determining sample size using the approach as summarized in Box 1. Entering the values in the formula yields: $2 \times [(1.96 + 0.842)^2 \times 20^2] / 15^2 = 27.9$, this means that a sample size of 28 subjects per group is needed to answer the research question.

SBP as a binary outcome

For a study with a binary outcome, calculating the required sample size is slightly more complicated. It should be calculated based on the number of events rather than on the number of people in the study (Box 2). The number of events can be increased either by choosing higher risk patients, by increasing the follow-up time, or by increasing the sample size.

Box 2
Simplest formula for a binary outcome and equal sample sizes in both groups, assuming: alpha = 0.05 and power = 0.80 (beta = 0.20).

n = the sample size in each of the groups
p_1 = proportion of subjects with hypertension in treatment Group 1
q_1 = proportion of subjects without hypertension in treatment Group 1 (= $1 - p_1$)
p_2 = proportion of with hypertension in treatment Group 2
q_2 = proportion of subjects without hypertension in treatment Group 2 (= $1 - p_2$)
x = the difference the investigator wishes to detect
a = conventional multiplier for alpha = 0.05
b = conventional multiplier for power = 0.80

$$n = \frac{[(a+b)^2(p_1q_1 + p_2q_2)]}{x^2}$$

In this case, we suppose that the investigators consider a difference in event rate of 10% (0.10) as clinically relevant. Based on recently published findings from studies with a similar design, they expect that the proportion of subjects with hypertension in the treated group will be ~20% (p_1 =

Table 3. Multipliers for conventional values of alpha and beta

Multipliers for conventional values of alpha	
Alpha	Multiplier
0.05	1.96
0.01	2.58
Multipliers for conventional values of beta	
Beta	Multiplier
0.20	0.842
0.10	1.28
0.05	1.64
0.01	2.33

0.20) and in the control group, ~30% ($p_2 = 0.30$). This automatically means that q_1 and q_2 are 0.80 and 0.70, respectively. Again, the investigators assume a power of 80% (0.80) and an alpha of 0.05, which means that the value 1.96 should be filled in for *a* and the value 0.842 should be filled in for *b*. We can now enter all values in the formula presented in Box 2: $[(1.96+0.842)^2 \times (0.20 \times 0.80 + 0.30 \times 0.70)] / 0.10^2 = 290.5$, this means that a sample size of 291 subjects per group is needed to answer the research question.

Many of the formulas to calculate sample size are not straightforward, and it is recommendable to ask for the help of a statistician in all but the most basic studies. For some simple clinical trials, nomograms or graphs can be used to estimate the sample size required for the study. An example of such a nomogram published by Altman is presented in Figure 1 [2]. However, one should keep in mind that, although these graphical methods work well, they often make assumptions about the type of data and statistical tests to be used.

Other outcome types

In many trials, the outcomes may not be continuous or binary as above, but instead may be survival (e.g. time to event). In these cases, the details of calculation differ, but using the four aforementioned components, persist through calculations with other types of outcomes. However, other assumptions can be necessary.

Other study designs than RCTs

In this paper, we focus on sample size calculations for RCTs, but also for studies with another design such as case-control or cohort studies, sample size calculations are sometimes required. Although the calculation of sample size is based on the same principles for all parallel study designs, the formulas for sample size calculations for other study designs than RCTs often need some adaptations. For example, published formulas for case-control designs provide sample sizes required to determine that

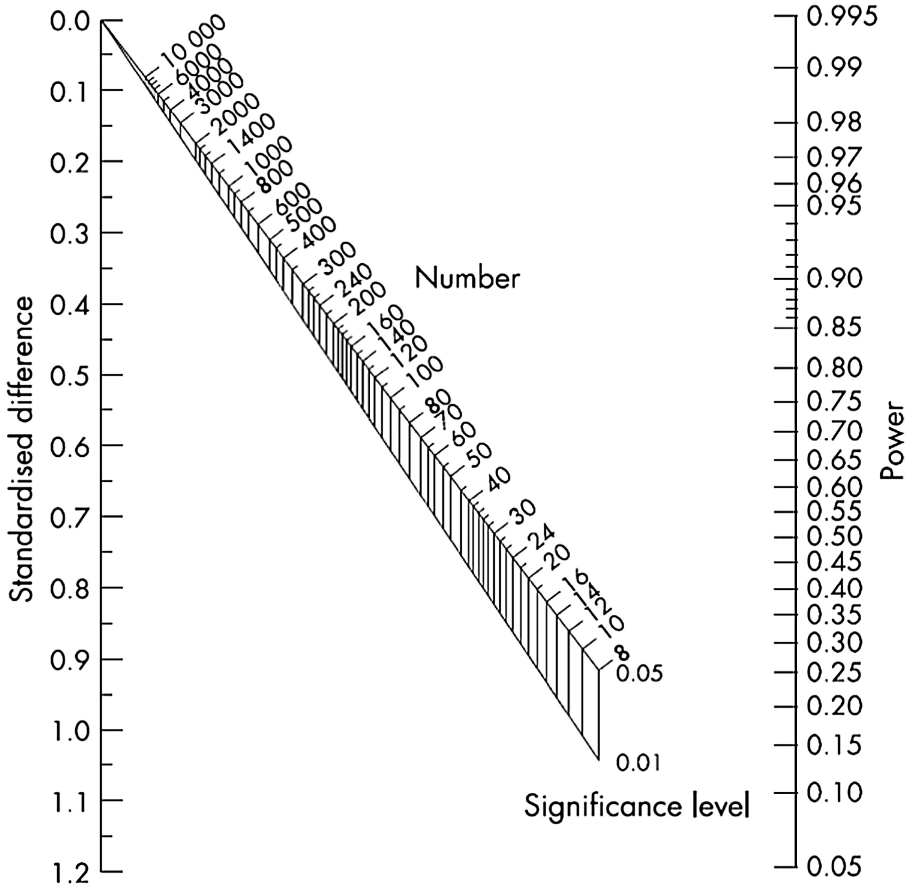

Fig. 1. Nomogram for the calculation of sample size or power (adapted from Altman 1982) [2].

Table 4. Approximate relative sample size for different levels of alpha and power

	Alpha (type I error)		
	0.05	0.01	0.001
Power (1-beta)			
0.80	100	149	218
0.90	134	190	266
0.99	234	306	402

an odds ratio is significantly different from one, after adjustment for potential confounders [3]. Also, sample size calculations for special types of RCTs, like cluster-randomized trials, in which health interventions are allocated randomly to intact clusters, communities, health centres or practices rather than to individual subjects, need an alternative approach [4]. The same holds true for trials with a crossover design, because these studies compare the results of two treatments on the same group of patients. The sample size calculated for a crossover study can also be used for a study that compares the value of a variable after treatment with its value before treatment [5]. Finally, sample size calculations for clinical trials testing the equivalence rather than the superiority between two treatments need another approach. These equivalence or non-inferiority trials usually demand higher sample sizes [6]. The aforementioned alternative calculations are less common and more complicated and will in most cases require statistical assistance.

Common pitfalls

The calculation of the sample size is troubled by a large amount of imprecision, because investigators rarely have good estimates of the parameters necessary for the calculation. Unfortunately, the required sample size is very sensitive to the choice of these parameters.

The effects of selecting alpha and the power. In most cases, the conventional choices of an alpha of 0.05 and a power of 0.80 are adequate. However, dependent on the topic studied, other assumptions can be made. Different assumptions of alpha and the power will directly influence the sample size, as is illustrated by Table 4. A lower alpha and a higher power will both lead to a larger sample size and as a result to higher costs. To be aware of the influence of changes in these parameters, it can be helpful to perform sample size calculations for different values of the parameters (sensitivity analyses). In case of doubt, one should generally choose the largest sample size.

Estimating the difference and SD. Studies often aim to determine parameters like event rates in the treated group and the control group. Needing to estimate these parameters before the start of the study therefore seems strange for many investigators. It is, however, important to realize that the parameters they are estimating in order to calculate the required sample size are not the population parameters as such, but the treatment effects they consider biologically plausible and clinically relevant.

In most studies, investigators estimate the difference of interest and the standard deviation based on results from a pilot study, published data or on their own knowledge and opinion. This means that the calculation of an appropriate sample size partly relies on subjective choices or crude estimates of certain factors which may seem rather artificial to some. Unless the pilot study was large, using information from a pilot study often results in unreliable estimates of the variability and the minimal clinically relevant difference. By definition, pilot studies are underpowered, and the observed difference in a pilot study is therefore an imprecise estimate of the difference in the population. Not accounting for this sampling error will lead to underpowered studies [7]. Also published reports could provide an estimate of the outcome in the control group. Although they often incorporate a lot of differences with the study one aims to perform, such as dissimilar eligibility criteria, endpoints and treatments, some information on the control group usually exists [8]. Finally, another approach is to survey experts in the field to determine what difference would need to be demonstrated for them to adapt a new treatment in terms of costs and risks [9].

After mentioning these pitfalls, it may seem useless to perform a sample size calculation. However, even if based on estimates and assumptions, a sample size calculation is considerably more useful than a completely arbitrary choice.

Post hoc sample calculations. Sometimes, published studies wrongfully report their 'power' instead of 95% confidence intervals (CIs). The power should always be calculated prior to a study to determine the required sample size, since it is the pre-study probability that the study will detect a minimum effect regarded as clinically significant. After the study is conducted, one should not perform any '*post hoc*' power calculations. Once the effect of the study is known, investigators should use the 95% CI to express the amount of uncertainty around the effect estimate.

Reporting of sample size calculations

According to the CONSORT statement, sample size calculations should be reported and justified in all published RCTs [10]. These calculations provide important information. Firstly, they specify the primary endpoint, which safeguards against changing the planned outcome and claiming a large effect on an outcome which was not the original primary outcome. Secondly, knowing the planned size alerts readers to potential problems, like problems with the recruitment or an early stop of the trial [8]. Readers of a published trial should be able to find all assumptions underlying the sample size calculation; the alpha, the power, the event rate in the control group and the treatment effect of interest (or the event rate in the treated group). Many studies only include statements like 'we calculated that the sample size in each treatment group should be 250 at an alpha of 0.05 and a power of 0.80'. However, such a statement is almost meaningless because it neglects the estimates for the effect of interest and the variability. Based on an example from the IDEAL Study, we can illustrate that a better way to report these calculations would

be: 'A clinically significant effect of 10% or more over the 3 years would be of interest. Assuming 3-year survival rates in the control group and the intervention group of 64% and 74% respectively, with a two-sided significance of 0.05 and a power of 0.8, a total of 800–1000 patients will be required' [11].

Although, ideally, all four components conventionally required for sample size calculation should be published, Charles *et al.* [12] recently showed that of 215 published reports of RCTs, 10 (5%) did not report any sample size calculation, and 92 (43%) did not report all the required parameters. Moreover, a Danish study by *Chan et al.* [13] demonstrated that, also in study protocols, sample size calculations are often poorly reported, and that explicit discrepancies between protocols and published papers are common. They found that only 11 of 62 (18%) identified studies described existing sample size calculations fully and consistently in both the protocol and the publication.

Further reading

Methods for sample size calculations are described in several general statistics textbooks, such as Altman (1991) [14] or Bland (2000) [15]. Specialized books which discuss sample size determination in many situations were published by Machin *et al.* (1997) [16] and Lemeshow *et al.* (1990) [17].

In addition, there are different software programs that can assist in sample size calculations. Examples of validated and user-friendly programmes that can be applied to calculate the sample size for several types of data and study designs are nQuery Advisor, PASS and 'Power and Precision'. For those programmes, a paid license is required. There are also a number of websites that allow free sample size calculations. However, those programmes are not always reliable. An example of a reliable, freely available website is: http://www.stat.uiowa.edu/~rlenth/Power/index.html [18].

Conclusions

In conclusion, the calculation of the sample size is one of the first and most important steps in designing a study. Although techniques for sample size calculations are described in many statistical textbooks, performing these calculations can be complicated. Because sample size calculations are sensitive to errors and because of the high costs related to performing an RCT, we recommend to perform the calculations with caution or to ask for statistical advice during the designing phase of the study.

Conflict of interest statement. The results presented in this paper have not been published previously in whole or part, except in abstract format.

References

1. Florey CD. Sample size for beginners. *BMJ* 1993; 306: 1181–1184
2. Altman DG. Statistics and ethics in medical research: III How large a sample? *Br Med J* 1980; 281: 1336–1338
3. Lui KJ. Sample size determination in case-control studies. *J Clin Epidemiol* 1991; 44: 609–612
4. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol* 1999; 28: 319–326
5. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. *Stat Med* 2008; 27: 5578–5585
6. Christensen E. Methodology of superiority *vs.* equivalence trials and non-inferiority trials. *J Hepatol* 2007; 46: 947–954
7. Wittes J. Sample size calculations for randomized controlled trials. *Epidemiol Rev* 2002; 24: 39–53
8. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005; 365: 1348–1353
9. Brasher PM, Brant RF. Sample size calculations in randomized trials: common pitfalls. *Can J Anaesth* 2007; 54: 103–106
10. Rothman KJ. *Epidemiology: An Introduction*. New York: Oxford University Press, 2002
11. Cooper BA, Branley P, Bulfone L *et al*. The Initiating Dialysis Early and Late (IDEAL) Study: study rationale and design. *Perit Dial Int* 2004; 24: 176–181
12. Charles P, Giraudeau B, Dechartres A *et al*. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* 2009; 338: b1732
13. Chan AW, Hrobjartsson A, Jorgensen KJ *et al*. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ* 2008; 337: a2299
14. Altman DG. *Practical Statistics for Medical Research*. London, UK: Chapman & Hall, 1991
15. Bland M. *An Introduction to Medical Statistics*. Oxford, UK: Oxford University Press, 2000, 3rd edn
16. Machin D, Campbell M, Fayers P *et al*. *Sample Size Tables for Clinical Studies*. London, UK: Blackwell Science, 1997, 2nd edn
17. Lemeshow S, Levy PS. *Sampling of Populations: Methods and Applications*. New York, US: John Wiley & Sons Inc, 1999, 3rd edn
18. Lenth R. V. *Java applets for power and sample size (computer software)* 2006–9. Retrieved October 12, 2009, from http://www.stat.uiowa.edu/~rlenth/Power

Received for publication: 15.10.09; Accepted in revised form: 3.12.09