# Randomised Trials in Surgery: The Burden of Evidence

Kristoffer Lassen*, Anne Høye and Truls Myrmel

*Dept. of Gastrointestinal and HPB Surgery, University Hospital Northern Norway (KL), Centre of Clinical Documentation and Evaluation, Strategic Health Authority, Region North (AH), Dept. of Cardiothoracic and Vascular Surgery, University Hospital Northern Norway (TM) UNN HF, 9038 Tromsø Norway and Institute of Clinical Medicine, University of Tromsø, 9038 Tromsø, Norway (KL/TM)*

**Abstract:** A randomised controlled trial (RCT) is considered the hierarchical peak of evidence-based medicine and a general demand for any result to be evaluated by RCTs has evolved. Yet, many advances in operative surgery do not result from RCTs and many controversies remain without an RCT being conducted. A randomised comparison of laparoscopic versus open liver resection has recently been called for. Using such a trial and others as examples, we examine the limitations of randomised design in skill-dependant interventions. Surgical procedures are skill-dependant, constantly developing, irreversible and traumatising. Additionally, placebo control is usually unethical and adequate blinding difficult or impossible to accomplish. Under these circumstances, surgeon and patient participation will be problematic and the resulting data will tend to have low external validity. While some of these obstacles can be modified, others will remain. Non-randomised, prospective cohort comparison has other weaknesses, but may add complementary data with good external validity. An alternative hierarchy of evidence is warranted in this field.

**Keywords:** Randomised Controlled Trial/RCT/Surgery/Skill dependent interventions/Learning curve/Cohort study/Evidence Based Medicine.

## INTRODUCTION

A recently presented single-centre series of laparoscopically performed liver resections (LapLR) has shown reassuring results compared to a matched historical control group that underwent open liver resections (OpenLR) in the same centre [1]. There were fewer blood transfusions given and less need for opiates postoperatively in the laparoscopic group, as well as shorter time to oral diet and shorter length-of-stay [1]. The authors conclude that a randomised controlled trial (RCT) comparing LapLR to OpenLR is necessary to prove the potential benefits shown in their series.

Randomised controlled trials have clarified a number of important issues in surgery for primary or secondary malignancies. A large number of RCTs have compared laparoscopic to open colorectal surgery for cancer and several systematic reviews and meta-analyses have been conducted [2-6]. In many instances patient recruitment was slow and total costs to the communities are unknown. While level-1 evidence [7] is now available, most trials had shortcomings directly related to their design. Lack of blinding is the most obvious, but also insufficiently standardised interventions or lack of general applicability as a consequence of surgeons being either insufficiently trained; or too exclusive and few. These shortcomings were predictable. For comparison, an RCT recently established the superiority of an antecolic to a retrocolic gastrojejunostomy in pylorus preserving pan creatoduodenectomies [8].

*Address correspondence to this author at the Kristoffer Lassen, Department of GI and HPB Surgery University Hospital Northern Norway 9038 Tromsø, Norway; Tel: + 47 776 26601; Fax: +47 776 26605; E-mail: xtofero@gmail.com or lassen@unn.no

These situations differ in some important aspects that have a direct bearing on the appropriateness or quality of an RCT. In this paper, we discuss whether an effort to perform an RCT, and hence achieve traditional level-1 evidence [7], is advisable for all surgical interventions.

### The Randomised Controlled Trial

A well conducted, carefully controlled and adequately blinded RCT constitutes the highest level of evidence in the Evidence Based Medicine (EBM) hierarchy [7]. While non-randomised comparison of prospective cohort series may yield important data, one can only adjust for factors that are already known to influence the outcome. Adjusted for these factors, an association between intervention (exposure) and outcome may be shown - or not shown.

Randomisation is the one design that can adjust for all factors, known and unknown, and hence the best design to establish causal relationship between exposure and outcome.

Randomisation is one of the most important methodological principles in medicine. It was first used in its present form in the Medical Research Council's trial showing the effect of streptomycin in severe tuberculosis in the UK following the Second World War, and published in 1948 [9-11]. Streptomycin had just been made available, supply was limited and the unpredictable natural course of tuberculosis was a confounder well recognized at the time [9-11]. This seminal trial very clearly illustrates the optimal setting for a successful RCT: to test an intervention that is novel (no one is prejudiced for, or against it), standardised (requires no technical skill to deliver), stable (unchanged properties during trial period) and easily tolerated (non-traumatising and

1876-1038/12 $58.00+.00 

reversible), with an available placebo and where double (or actually triple) blinding is feasible. Read: a new drug!

The optimal target disease is, like tuberculosis, one where spontaneous improvements are probable and cannot confidently be attributed to intervention. A displaced fracture is the reversed situation: an anatomically healed limb after repositioning and plaster cast is not due to spontaneous improvement but a result of the intervention. We do not need an RCT to convince us and this is recognized in the EBM hierarchy as level 1c-evidence: "all-or-none" or "dramatic effect".

The blinded testing of a new drug against an existing one, or against placebo, is what earned the RCT its impressive track-record and hence its position at the top of the EBM hierarchy. Evaluating a novel or modified surgical intervention is a different situation. Most obviously, placebo control (sham operations in animal research) is usually unethical and successful blinding difficult to accomplish. Less obvious, but equally important: most surgical procedures are skill-dependant and hence have a learning-curve. This challenges standardisation and implies that the timing of a trial (relatively to the learning curve) will affect patient accrual and the validity of the results.

## Challenges to RCTs in Surgery

An RCT should not be attempted if it is unnecessary, inappropriate, inadequate or impossible [12], and this could be the case in 60 % of surgical treatment questions [13]. It may also apply to Ito and co-workers' call for an RCT to compare LapLR with OpenLR. One may argue that an explanatory trial is unnecessary and a pragmatic RCT is inadequate. Both designs might prove impossible as they will probably accrue patients very slowly.

A trial cannot answer all questions pertaining to a new intervention under scrutiny. It will be designed either to evaluate causality, i.e. whether the intervention may cause the desired outcome under optimal circumstances; or to evaluate feasibility, i.e. whether this is achievable in a general context and hence the effect of changing clinical practice "across the nation". In the first instance, one gets an explanatory-, expert-, or efficacy trial dependant on adequate standardisation and absence of bias to ensure optimal internal validity. In the latter case, one gets a pragmatic-, feasibility-, or effectiveness trial, dependant on patients and surgeons being representative of the general community to produce generalizability, also referred to as external validity.

One design is not more important than the other; they produce reciprocally complementing data, and most clinical challenges need to have both issues assessed. It is the first issue however, that of causality, which has generally received most focus. And it is the question of causal relationship that an RCT is the supreme tool to answer. The success of high-quality RCTs to ascertain causal relationships has to some extent overshadowed the question of whether it transfers into a large or small treatment effect when generally applied [14].

This applies to all trials, but RCTs are more vulnerable because problems with standardisation and timing negatively affect patient accrual. And slow accrual is an Achilles' heel of this design [15]. The reason for this is inherent in the method: experimentation, i.e. randomisation, implies that patients and surgeons cannot decide on treatment from preference, but must accept a random choice between options considered reasonably equal. If patients have a strong preference for one of the two, they are considered to lack equipoise, and will be difficult to recruit [15]. Surgeons may equally lack equipoise and be reluctant to participate [15]. If causality is the prime issue, an explanatory RCT should always be attempted. If feasibility of a nationwide change of practice is the aim, optimal generalizability must be sought and hence a design that ensures wide patient and surgeon recruitment. We have every reason to believe that only a very small minority of eligible patients participate in randomised trials [16].

Most trials are not absolutely explanatory or pragmatic; there is a balance between the two, but the nature of skill-dependant interventions with a learning curve implies that you cannot boost internal validity – which is generally the main priority – without loosing external validity, and vice versa. This represents a fundamental difference from a drug-trial where standardisation and blinding is effortless and complete.

Randomisation is inappropriate if the target outcome (event) rate is extremely low, e.g. for certain rare complications to surgery, as this will necessitate a prohibitive number of patients to achieve statistical significance [17]. The difference in 5 year-survival after LapLR or OpenLR for liver metastasis is believed to be none or very small. An RCT will therefore require a very large number of patients and probably need more than 10 years to acquire conclusive statistical power [18]. Showing reductions by 50 % for complications like bile leakage, liver failure or operative mortality is equally prohibitive, requiring 1300 to 6200 patients in each arm with a power of 0.90 and a two-tailed test [17]. If smaller, but clinically relevant effects are targeted, the increase in numbers needed will be exponential.

## Learning-curve and the Timing of Trials

The technical success of any RCT depends on participants not having too strong preference (i.e. the presence of equipoise) for one of the alternatives as this will bar recruitment [15]. Lack of surgeon- and/or patient equipoise has halted trials where academical equipoise was considered to be present [19,20].

When an intervention has a substantial learning curve, yet another factor is introduced: an intervention performed at the beginning of the curve is different from one performed at the plateau-phase at a later time-point [21,22]. This adversely affects standardisation, but also equipoise changes over time as a novel technique becomes gradually refined and generally accepted in both the medical and the general society. Hence, deciding on the correct time to evaluate a novel intervention is difficult and affects both standardisation and patient accrual. Sufficient expertise must be achieved to perform a standardised procedure, but surgeon and patient equipoise must still be present [23]. As Buxton has observed on the timing of evaluation: "It's always too early until, unfortunately, it's suddenly too late" [24].

The absence of an optimal time-window for a trial (or one already lost) and a predictable lack of patient and surgeon equipoise could detrimentally slow patient accrual in a proposed RCT on laparoscopic versus open left lateral liver resections. The world-wide introduction of laparoscopic cholecystectomies without any RCTs is the most conspicuous example and illustrates the loss of a time-window, patient- and surgeon preference and also market forces at work.

### Validity

The timing of a trial with a learning-curve also affects validity. Internal validity is reduced if the procedure under evaluation is not standardised within the trial. This will happen both as participating surgeons will position themselves differently on the learning-curve, but also because a complex procedure will be developed and refined as each surgeon moves onwards along the curve [25]. Standardisation can be increased through vigorous teaching and control [26], but it will always be a challenge in operative technique trials as improvisation is generally encouraged in surgery [27], and to some degree even a necessity for innovation [28]. External validity is reduced if the results have a low generalizability, as will be the case if only expert surgeons far out on the curve participate. Any time-point chosen for the trial will have its advantages and its problems, mostly in the balance between internal and external validity. In skill-dependant procedures with a learning curve, there will always be a trade-off between the two, and the choice will also affect patient accrual.

### Narrow Inclusion Criteria

The degree to which a trial is pragmatic versus explanatory is influenced by the timing of the trial and choice of participants and can hence be manipulated. A very homogenous sample of patients, e.g. without any co-morbidity, may be selected to boost standardisation in an explanatory trial, but this will automatically reduce external validity. Major, right-sided liver resections are considered challenging to perform laparoscopically and participating surgeons (and patients) are likely to be chosen with utmost care. Hence, neither patients nor surgeons will be representative. Data from patients that were for some reasons not invited, or who declined, are lost and these patients cannot be assumed to be similar to those participating [12]. It is important to acknowledge this as it is an inherent feature in experimentation that always results in a loss of external validity.

### Bias

Bias is the conscious or subconscious skewing of information. Selection-bias occurs when a patient sample, believed to be representative, proves not to be. It is prevented by randomisation. Outcome-assessor bias acknowledges that any outcome assessor in a trial is prone to be influenced by pre-trial perceptions and colleagues' preferences. This may affect interpretation of all data that are not completely objective, and few are. The only way to prevent this is blinding. While not impossible [29], blinding of a surgical access-techniques might be cumbersome and incomplete. Furthermore, blinding is not a magic bullet: it only prevents assessor bias. Two RCTs on laparoscopic versus open surgery (cholecystectomies and colonic resections) that are widely cited as examples of successful blinding, both concluded that laparoscopy did not result in shorter length-of-stay [29,30]. While previous trials with historical controls clearly had overestimated the benefit, few would subscribe to the conclusions of these RCTs today.

### Is There an Alternative to RCT?

The difficulties in performing robust RCTs evaluating craft-based interventions are many and not unique to surgery. What makes surgery an area of concern in this respect is that "many of these challenges coincide" [23]. In the latest two decades, the influence from EBM has been strong and important. It is reasonably agreed that the methodological quality of trials in operative surgery tends to be low in too many instances [28]. It is important, however, that the virtues of the RCT in ideal conditions do not overshadow the fact that there are situations where an RCT might not be the best alternative. In these cases we might instead plan for high-quality, non-randomised, protocol-driven prospective comparison [31-34]. For the sake of our patients, we should use all data from our "biased surgical laboratory". It should suffice to point to the overwhelming success of for instance congenital heart surgery. Without RCTs, the surgical mortality for these complex procedures have been reduced from prohibitive to almost nil [35,36], based almost exclusively on categorising and shearing observational data.

Non-randomised cohort comparison will generally ensure an almost complete participation as interventions are provided according to preference. While the important issue of an unknown confounder and hence a risk of selection bias cannot be overstated, these observations will nevertheless provide the real-world effects of surgeons putting their craft of best performance into each technique. A recent systematic comparison found meta-analysis of well-designed non-randomised trials to be as good as RCTs [37].

The quality of observational (non-randomised) trials will improve with the adoption of core elements of RCT design like a publicly registered protocol, pre-defined entry criteria, rigid prospective inclusion, intention-to-treat analysis, complete follow-up etc. [38]. Good observational data will not obviate the need for an RCT, but provide robust and complimentary data alongside those of a possible future RCT [38].

Core elements of trial design that must be considered are shown in Fig. (**1**).

## CONCLUSION

A well conducted RCT remains the most powerful tool to assess an intervention. It should always be considered, but there are situations where it is not particularly well suited. While surgical innovation cannot proceed without rigorous testing, evaluation and control, this must be performed with designs and against standards that acknowledge the challenges posed by key features of operative surgery [39,40]. These issues need not be irreconcilable. A mandatory participation in a national prospective database with patient characteristics and outcome for any major innovation in sur-

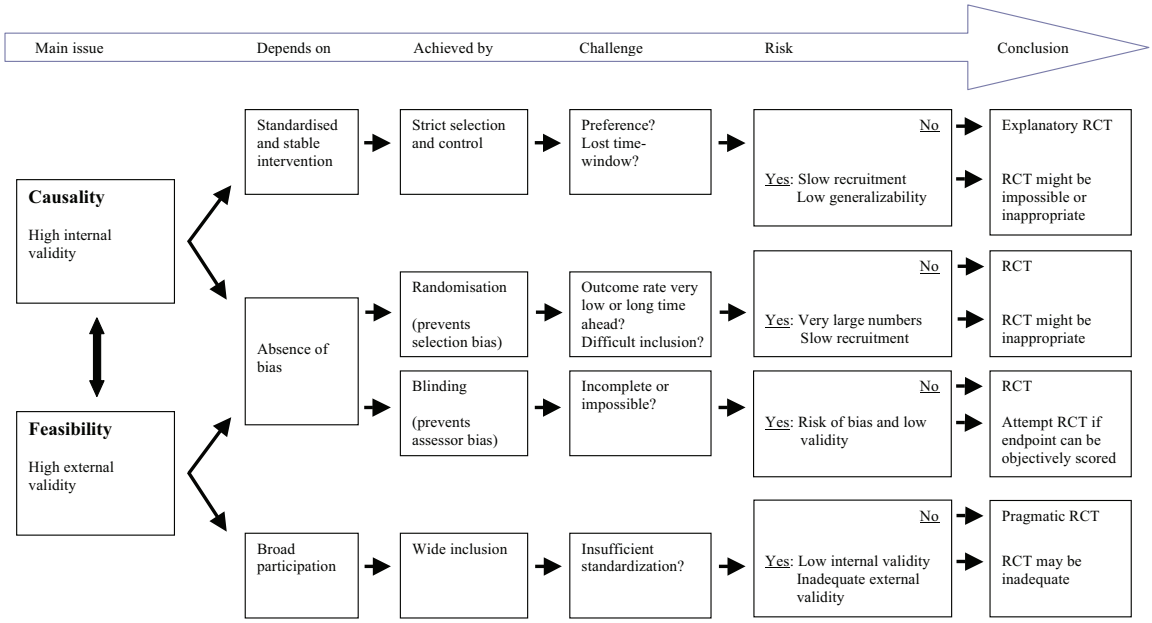


**Fig. (1).** Suggested flow-chart for planning process taking into consideration the balance between causality and feasibility.

gical technique may be proposed. Such a database should be financed and controlled by the health authorities for a limited number of years and may prove to new surgical procedures what the RCT has become in drug trials. For skill-dependant interventions in general, one is also tempted to advocate a change of paradigm, from today's hierarchical pyramid of evidence and to a circular model as suggested by Walach and co-workers [41]. A change from the narrow focus on internal validity and absence of bias, to an acknowledgment of the reciprocal value of different kinds of evidence for different situations and a wider view including external validity and general applicability.

## ACKNOWLEDGEMENTS

We would like to thank B. Vonen and A. Revhaug, University Hospital Northern Norway, and CHC Dejong, Maastricht University Medical Centre, the Netherlands, for their valuable comments on an earlier version of the manuscript.

## CONFLICTS OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

## AUTHORS' CONTRIBUTIONS

KL: Conceived the general idea, wrote the first draft and outline of the argument and participated in redrafting and rewriting.

AH and TM: Critically read and revised the manuscript and added scientific and argumentative content.

All authors read and approved the final manuscript.

## REFERENCES

[1] Ito K, Ito H, Are C, *et al.* Laparoscopic versus Open Liver Resection: A Matched-Pair Case Control Study. J Gastrointest Surg 2009; 13(12): 2276-83.

[2] Tjandra JJ, Chan MK. Systematic review on the short-term outcome of laparoscopic resection for colon and rectosigmoid cancer. Colorectal Dis 2006; 8(5): 375-88.

[3] Kuhry E, Schwenk W, Gaupset R, Romild U, Bonjer J. Long-term outcome of laparoscopic surgery for colorectal cancer: a cochrane systematic review of randomised controlled trials. Cancer Treat Rev 2008; 34(6): 498-504.

[4] Ma Y, Yang Z, Qin H, Wang Y. A meta-analysis of laparoscopy compared with open colorectal resection for colorectal cancer. Med Oncol 2011; 28(4): 925-33.

[5] Kahnamoui K, Cadeddu M, Farrokhyar F, Anvari M. Laparoscopic surgery for colon cancer: a systematic review. Can J Surg 2007; 50(1): 48-57.

[6] Bonjer HJ, Hop WC, Nelson H, *et al.* Laparoscopically assisted vs open colectomy for colon cancer: a meta-analysis. Arch Surg 2007; 142(3):298-303.

[7] Phillips B, Ball C, Sacket D, *et al.* Levels of evidence and grades of recommendations. 2007 ed. Oxford: Oxford Centre for Evidence-Based Medicine 2007.

[8] Tani M, Terasawa H, Kawai M, *et al.* Improvement of delayed gastric emptying in pylorus-preserving pancreaticoduodenectomy: results of a prospective, randomized, controlled trial. Ann Surg 2006; 243(3): 316-20.

[9] Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ 1948; 2: 769-82.

[10] [Editorial]. The controlled therapeutic trial. BMJ 1948; 2: 791-2.

[11] Yoshioka A. Use of randomisation in the Medical Research Council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. BMJ 1998; 317(7167): 1220-3.

[12] Black N. Evidence-based surgery: A passing fad? World J Surg 1999; 23(8): 789-93.

[13] Solomon MJ, McLeod RS. Should we be performing more randomized controlled trials evaluating surgical operations? Surgery 1995; 118(3): 459-67.

[14] Al Refaie WB, Pisters PW, Rothenberger DA. Surgical oncology trials and surgeons in the real world! Ann Surg Oncol 2010; 17(7): 1727-8.

[15] Abraham NS, Young JM, Solomon MJ. A systematic review of reasons for nonentry of eligible patients into surgical randomized controlled trials. Surgery 2006; 139(4): 469-83.

[16] Abraham NS, Byrne CJ, Young JM, Solomon MJ. Meta-analysis of well-designed nonrandomized comparative studies of surgical procedures is as good as randomized controlled trials. J Clin Epidemiol 2010; 63(3): 238-45.

[17] van den Broek MA, van Dam RM, Malago M, Dejong CH, van Breukelen GJ, Damink SW. Feasibility of randomized controlled

trials in liver surgery using surgery-related mortality or morbidity as endpoint. Br J Surg 2009; 96(9): 1005-14.

[18] Buell JF, Cherqui D, Geller DA, *et al*. The international position on laparoscopic liver surgery: The Louisville Statement, 2008. Ann Surg 2009; 250(5): 825-30.

[19] Barkun JS, Barkun AN, Sampalis JS, *et al*. Randomised controlled trial of laparoscopic versus mini cholecystectomy. The McGill Gallstone Treatment Group. Lancet 1992 Nov 7; 340(8828): 1116-9.

[20] Meakins JL. Innovation in surgery: the rules of evidence. Am J Surg 2002; 183(4): 399-405.

[21] Vigano L, Laurent A, Tayar C, Tomatis M, Ponti A, Cherqui D. The learning curve in laparoscopic liver resection: improved feasibility and reproducibility. Ann Surg 2009; 250(5): 772-82.

[22] Dagher I, O'rourke N, Geller DA, *et al*. Laparoscopic major hepatectomy: an evolution in standard of care. Ann Surg 2009; 250(5): 856-60.

[23] Ergina PL, Cook JA, Blazeby JM, *et al.* Challenges in evaluating surgical innovation. Lancet 2009; 374(9695): 1097-104.

[24] Buxton MJ. Managing new technology: economics research and practical decisions. Health Serv Manage Res 1988; 1(1): 43-50.

[25] Vigano L, Laurent A, Tayar C, Tomatis M, Ponti A, Cherqui D. The learning curve in laparoscopic liver resection: improved feasibility and reproducibility. Ann Surg 2009; 250(5): 772-82.

[26] Attwood SE, Lundell L, Ell C, *et al*. Standardization of surgical technique in antireflux surgery: the LOTUS Trial experience. World J Surg 2008; 32(6): 995-8.

[27] Biffl WL, Spain DA, Reitsma AM, *et al*. Responsible development and application of surgical innovations: a position statement of the Society of University Surgeons. J Am Coll Surg 2008; 206(3): 1204-9.

[28] Barkun JS, Aronson JK, Feldman LS, *et al*. Evaluation and stages of surgical innovations. Lancet 2009; 374(9695): 1089-96.

[29] Basse L, Jakobsen DH, Bardram L, *et al.* Functional recovery after open versus laparoscopic colonic resection: a randomized, blinded study. Ann Surg 2005; 241(3): 416-23.

[30] Majeed AW, Troy G, Nicholl JP, *et al.* Randomised, prospective, single-blind comparison of laparoscopic versus small-incision cholecystectomy. Lancet 1996; 347(9007): 989-94.

[31] Young JM, Solomon MJ. Improving the evidence-base in surgery: evaluating surgical effectiveness. ANZ J Surg 2003; 73(7): 507-10.

[32] Slim K. Limits of evidence-based surgery. World J Surg 2005; 29(5): 606-9.

[33] McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ 2002; 324(7351): 1448-51.

[34] Lilford R, Braunholtz D, Harris J, Gill T. Trials in surgery. Br J Surg 2004; 91(1): 6-16.

[35] Cardiac Surgery in Norway. Norwegian Association of Cardiothoracic Surgeons 2008; available at: www.legeforeningen.no/thorax

[36] Cardiac Surgery in Sweden. Swedeheart 2008; available at: www.ucr.uu.se/hjartkirurgi

[37] Abraham NS, Byrne CJ, Young JM, Solomon MJ. Meta-analysis of well-designed nonrandomized comparative studies of surgical procedures is as good as randomized controlled trials. J Clin Epidemiol 2009; 63(3): 238-45.

[38] Farrokhyar F, Karanicolas PJ, Thoma A, *et al.* Randomized controlled trials of surgical interventions. Ann Surg 2010; 251(3): 409-16.

[39] McCulloch P, Altman DG, Campbell WB, *et al.* No surgical innovation without evaluation: the IDEAL recommendations. Lancet 2009; 374(9695): 1105-12.

[40] Cooper JD. Randomized clinical trials for new surgical operations: square peg in a round hole? J Thorac Cardiovasc Surg 2010; 140(4): 743-6.

[41] Walach H, Falkenberg T, Fonnebo V, Lewith G, Jonas WB. Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. BMC Med Res Methodol 2006; 6: 29.

Received: February 08, 2012 Revised: May 07, 2012 Accepted: May 08, 2012