Am J Nucl Med Mol Imaging 2012;2(3):344-352
www.ajnmmi.us /ISSN:2160-8407/ajnmmi1204002

Original Article

Interim analyses in diagnostic versus treatment studies: differences and similarities

Oke Gerke[1,2], Poul Flemming Høilund-Carlsen[1], Mads Hvid Poulsen[3], Werner Vach[4]

[1]Department of Nuclear Medicine, Odense University Hospital, Sdr. Boulevard 29, DK-5000 Odense C, Denmark; [2]Centre of Health Economics Research, Department of Business and Economics, University of Southern Denmark, Campusvej 55, DK-5000 Odense C, Denmark; [3]Department of Urology, Odense University Hospital, Sdr. Boulevard 29, DK-5000 Odense C, Denmark; [4]Clinical Epidemiology, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Str. 26, D-79104 Freiburg, Germany

Received April 29, 2012; accepted June 25, 2012; Epub July 10, 2012; Published July 30, 2012

Abstract: The purpose of this paper was to contrast interim analyses in (randomized controlled) treatment studies with interim analyses in paired diagnostic studies of accuracy with respect to planning and conduct. The term 'treatment study' refers to a (randomized) clinical trial that aims to demonstrate the superiority or noninferiority of one treatment compared with another, and the term 'diagnostic study' to a clinical study that compares two diagnostic procedures, using a third diagnostic procedure as the gold standard. Though interim analyses in treatment studies and paired diagnostic studies show similarities in *a priori* planning of timing, decision rules, and the consequences of the analyses, they differ with respect to (1) the need for sample size adjustments, (2) the possibility of early decisions without early stopping, and (3) the impact of keeping results secret. These differences are due, respectively, to certain characteristics of paired diagnostic studies: the dependence of the sample size on the agreement rate between the modalities, multiple aims of diagnostic accuracy studies, and the advantages of early unblinding of results at the individual level. We exemplified our points by using a recent investigation at our institution on the detection of bone metastases from prostate cancer in patients with histologically confirmed prostate cancer in which ^{99m}Tc-MDP whole body bone scintigraphy was compared to positron emission tomography/computed tomography with ^{18}F-fluorocholine as tracer, using magnetic resonance imaging as a reference.

Keywords: Study design, diagnostic imaging, PET/CT, efficacy studies, accuracy studies, sample size

Introduction

Nuclear Medicine has made substantial contributions to the progress in diagnostic imaging during several decades with modalities like single-photon emission computed tomography (SPECT), positron emission tomography (PET), and positron emission tomography/computed tomography (PET/CT) and recently positron emission tomography/magnetic resonance imaging (PET/MRI). However, regulatory authorities are increasingly reluctant to allowing introduction of new modalities in daily practice due to cost-effectiveness concerns. Hence, rigorous, comparative studies demonstrating the value of new imaging modalities are demanded [1, 2], which is costly and time-consuming. With respect to the latter, interim analyses may offer an advantage allowing early stopping of a study, and may play an important role in diagnostic research in the future. At first sight, the established framework for interim analyses developed for treatment studies can be applied in diagnostic studies, too. However, there are some subtle differences, which we would like to point out in this paper.

Nowadays interim analyses are an integral part of most late phase II and phase III trials comparing two treatments. They are conducted to ensure that patients are not exposed unnecessarily to clearly inferior or dangerous treatments and to accelerate scientific and clinical progress if clear evidence can be obtained much earlier than originally expected [3]. Similar considerations apply to studies comparing two diagnostic procedures if one procedure turns out to be ineffective [4]. Interim analyses are already well established in treatment studies, and the scientific, ethical, and organizational aspects of such interim analyses have been dealt with in the literature during the last two decades [5, 6]. There are guidelines on the requirements for the conduct of a clinical trial in general, i.e., in the spirit of the Declaration of Helsinki (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073122.pdf - last accessed in June 2012), and on statistical principles in particular (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073137.pdf - last accessed in June 2012). The latter comprises considerations on trial conduct (like trial monitoring and interim analysis, sample size adjustment, interim analysis and early stopping), data analysis (e.g. prespecification of the analysis, adjustment of significance and confidence intervals), and evaluation of safety and tolerability. Statistical procedures have been developed to solve the multiple inference issues, i.e., repeated testing of statistical hypotheses, resulting from the conduct of interim analyses. These comprise the so called 'α spending function approach' where the significance level is basically split into smaller pieces to be used in interim analyses and the final analysis at the end of the trial [7] and the boundaries approach in which test statistics are plotted against the sample size until certain stopping boundaries are crossed [8]. Both concepts keep the experiment-wise Type I Error at the nominal significance level. Further, the emergence of adaptive designs has widened the possible impact of interim analyses on the conduct of treatment studies, enabling, for example, treatment selection, treatment switch, or hypothesis selection at interim [9-13].

When the long-term impact of diagnostic modalities on patient management and time-to-event endpoints such as 'overall survival' or 'time to relapse' are investigated, randomized controlled trials (RCT) tend to be the design of choice [1, 2]. In the terminology of Fryback and Thornbury [14], these diagnostic studies are 'Level 5 efficacy studies', which measure the effect of the modalities on patient outcomes. Studies of this type, however, are rare [15]. Most of the studies in diagnostic research today are still accuracy studies, i.e., 'Level 2 efficacy studies', which address diagnostic accuracy, typically in terms of sensitivity and specificity. While phase II and III treatment studies are usually RCTs, comparative phase II diagnostic studies of accuracy are often designed as paired studies, in which all diagnostic imaging modalities are applied in each patient. At first glance interim analyses in paired diagnostic studies may be performed according to the established framework for treatment studies. However, due to the inherent differences between RCTs and paired diagnostic studies, the role of interim analyses in each may vary. It is the purpose of this article to point out three major differences, examine their consequences, and discuss how interim analyses can be used more efficiently and advantageously in paired diagnostic studies. We exemplify our points by using a recent investigation at our institution in which the detection of bone metastases from prostate cancer in patients with histologically confirmed prostate cancer by ^{99m}Tc-MDP whole body bone scintigraphy was compared to the detection by PET/CT with ^{18}F-fluorocholine, using MRI as a reference.

Materials and methods

Definitions

The term 'treatment study' refers to a (randomized) clinical trial that aims to demonstrate the superiority or noninferiority of one treatment compared with another treatment or a placebo. The classical statistical approach to show superiority (or noninferiority) is to apply a hypothesis test with the objective of rejecting the null hypothesis of no (or only a negligible) treatment difference. This kind of study is typically conducted in a parallel randomized design (e.g. [16]).

The term 'diagnostic study' refers to a clinical study that compares two diagnostic procedures, using a third procedure, i.e. the gold standard, as a reference. Here it is necessary to distinguish between two basically different designs: (a) the paired design, in which both procedures and the gold standard are applied in each patient, and (b) the unpaired design, in which only one randomly chosen procedure is applied in each patient in addition to the gold standard. The differences between the two study designs have already been presented in the literature [17, 18]. The paired design will be ruled out if it is either impossible or unethical to perform both procedures in each patient. However, as diagnostic procedures arising in the field of nuclear medicine are typically noninvasive and can be applied in addition to the standard practice, this restriction does usually not apply.

In a paired diagnostic study, blinding of the patient is often impossible, but the results can be temporarily blinded by requiring that each of the two diagnostic procedures be performed and assessed without knowledge of the results of the other one. This type of blinding is essential to ensure an unbiased comparison. However, once both procedures have been carried out and their results recorded (e.g. PET/CT images have been assessed by nuclear medicine physicians), it is no longer necessary for the results to remain undisclosed to physicians taking care of patient management, provided of course the study is focused solely on diagnostic accuracy without secondary objectives like risk evaluation or assessment of long-term prognosis. So, patients can benefit from both procedures which make the paired design very attractive from an ethical point of view. For example in comparing two imaging modalities to detect (local) metastases in cancer patients, lesions indicated by only one modality will, if accessible, typically be investigated further by some kind of bioptic procedure to ensure the true character of as many lesions as possible. If the results of the gold standard can also be obtained without delay, they too can be made available to the treating physicians for the same reasons.

It may sound inappropriate to use the results of a new, yet unapproved, diagnostic procedure in decisions on patient care. However, in nuclear medicine new imaging procedures have often a face advantage by providing images with better resolution and/or over a wider region (e.g. whole body scans) than established procedures. Indeed, diagnostic studies have the aim to demonstrate that this face advantage does also result in a higher accuracy, but it may be unethical to ignore this face advantage in the daily routine, although its existence should be dealt with some care.

Treatment studies vs. paired diagnostic studies

Table 1 contrasts the objectives and design issues of treatment studies and those of diagnostic studies, especially with respect to the use of interim analyses. Interim analyses in diagnostic studies focus often only on futility [19-23]. A very distinct and statistically highly significant improvement is often unlikely or impossible since an already substantially accurate standard procedure is already in use. Interim analyses in treatment trials are typically performed by external, independent Data Monitoring Committees. A similar tradition has not yet been established in diagnostic studies.

Results

Adjustment of sample size based on information about additional parameters

The power of a treatment study depends on the treatment effect and additional parameters. In the case of a continuous outcome (e.g. a decrease of blood pressure measured in mmHg), the power depends on the difference between the means and the population variance. In case of a binary outcome the power depends on the risk difference or risk ratio and the disease prevalence, and, finally, in case of a survival or another time-to-event outcome it depends on the hazard ratio and the event rate. An interim analysis allows us to detect that the original assumptions about the additional parameters used to complete the sample size calculation may be incorrect, and, hence, we can consider adjusting the planned sample size. However, this is usually not the main issue in an interim analysis for a treatment study as the influence of the additional parameters is often relatively limited compared with other factors, for example, the recruitment rate.

In paired diagnostic studies, on the other hand, the need for sample size adjustments is much greater. The simple reason for this is that the sample size required to reject certain hypotheses on the difference in sensitivity and specificity or to reach a certain precision in the corresponding estimates depends heavily on the degree of agreement between the two diagnostic procedures [24-31]. For example, if we assume a true difference of 10% in sensitivity between two modalities (and a significance level of 5%), a study including 100 patients can have a power of 56.7% if the agreement rate between the two modalities is 80%, and a power of 94.0% if the agreement rate is 90% [30]. The higher the agreement rate, i.e., the lower the number of patients showing different results, the higher the power (see Table 6 in [30]). Nonetheless, it is very difficult to achieve a reliable estimate of the degree of agreement be-

Table 1. Characteristics of phase II and phase III treatment and diagnostic studies.

	Treatment study		Diagnostic study	
	Early phase II	Late phase II / Phase III	Comparative accuracy studies	RCT
Objective	Dose-finding, administration form, preliminary efficacy, safety	Primary: efficacy; Secondary: safety	Accuracy: sensitivity, specificity, positive and negative predictive values	Patient benefit/ outcome such as survival
Design	Single-arm studies, food-drug interaction crossover trial	RCT	Paired	RCT
Blinding	Kept when study is randomized	To the fullest possible extent	Temporarily of results	Rarely possible
Input for sample size planning	Phase I studies of respective clinical project, literature	Phase II studies of respective clinical project, literature	Pilot study, possibly literature	Combining results from accuracy studies with expected benefit from improved diagnoses [2], literature
Use of interim analyses	No	Yes, with external, study-independent Data Monitoring Committees	Yes, internal assessment of interim results	Yes, internal assessment of interim results
Purpose of interim analyses	Not applicable	Primary: early stopping due to fertility or futility; Secondary: adjustment of sample size	Primary: adjustment of sample size due to lacking/limited *a priori* information for sample size planning; Secondary: early stopping due to futility	Primary: early stopping due to fertility or futility; Secondary: adjustment of sample size

tween two diagnostic procedures when planning a diagnostic study. Typically, we compare a new method with a standard procedure with which it has never been compared head-to-head. Therefore, the study is likely to be conducted initially with an incorrect assumption. Then, in an interim analysis, the true degree of agreement can be easily estimated even if the gold standard results are not disclosed. A discrepancy between the original assumptions and the interim results can be detected, and it may be deemed necessary to adjust the sample size. In an earlier publication, we made a concrete suggestion for how to perform such an adjustment [30], as will be exemplified later. A further source for a need of sample size adjustments may be updated information on the prevalence gleaned from the interim analysis.

Early decision without early stopping

A treatment study can be stopped early due to futility or fertility if the respective predefined stopping criteria have been met at interim, but the study will be continued if no conclusive results have been achieved. Investigating for example the ability of a new compound to decrease high blood pressure (measured in mmHg), the decrease of blood pressure is a mandatory condition for the success of this compound. If it fails to demonstrate this ability, the clinical development for this compound is over (at least for that particular indication). So, in treatment studies there is a clear relationship between early decision and early stopping: if we are sure about the inferiority or the superiority or the new treatment, the study must be stopped at once, as it would be unethical to continue to randomize and to offer half of the patients inferior treatment.

The situation can be very different in a paired diagnostic study. Firstly, accuracy is a two-dimensional concept described typically by sensitivity and specificity. In an interim analysis we may come to a firm conclusion with respect to one parameter, but not with respect to the other, in particular if the prevalence is not close to 0.5. Then, the power to determine sensitivity in diseased patients and specificity in diseasefree patients will differ. Secondly, whereas an improvement in diagnostic accuracy is regarded as sufficient for a change of the standard clinical routine, regulatory authorities require often also proof of clinical benefit, in particular when it comes to reimbursement decisions [2]. This may imply that we, for instance, may be able to demonstrate an increase in accuracy at the lesion level in an interim analysis, but have to continue the study in order to demonstrate also a benefit at the patient level (as will be exemplified later). Thirdly, even if we have demonstrated an improved diagnostic accuracy of the new modality, the question may remain whether we can become even better by applying both modalities in the future. This question requires typically a larger sample size than the comparison of the two modalities, as the gain from applying both procedures jointly is typically smaller than the difference between the two procedures. Hence, in diagnostic studies there can be many good reasons to perform an interim analysis reaching some firm conclusions of general interest to be published, but nevertheless to continue the study anyway. Consequently, interim analyses in diagnostic accuracy studies have to be considered differently from interim analyses in treatment studies. The central question is not whether to stop or to continue the study, but whether some results are already statistically significant and convincing enough to be made public whereas for others we have still to continue and wait and see. Therefore, we can regard interim analyses as a monitoring tool of important aspects of a study with the plan to continue until the last question of interest has been answered.

Of course, such a strategy has to be fixed a priori already in the study protocol. Ethical objections need also to be addressed, e.g. if we plan to continue to use an invasive procedure which has been proven to be inferior at the lesion level in order to confirm the result at the lesion level. Minor harm like exposure to a limited radiation dose may be justified by the continuing advantage for the patients have their care based on the results of both modalities.

Keeping results from interim analyses secret

In treatment studies, blinding at the individual levels ensures unbiased assessment of treatment outcome and adverse effects in each patient. Moreover, it protects the integrity of a trial by preventing premature conclusions: if blinding is successfully done and kept, there is little danger for rumors to arise about the results of the study. Consequently, results of interim analyses will typically be kept secret (except if the trial is stopped) to avoid any disturbance of the ongoing trial.

The situation is different in paired diagnostic studies. As pointed out above, it is one of the advantages of paired diagnostic studies that patients can benefit from the results of both modalities. Hence, it is common practice that the treating physician knows both results, and, typically that he or she will also know the result of the gold standard procedure or at least the follow-up data of the patient. In either case treating physicians may experience a success or failure of the modalities in some patients, and, hence, there is a risk for developing an opinion about the superiority of one modality over the other. If treatment decisions are made in interdisciplinary conferences, even more people than just the treating physician may experience successes or failures or at least discrepancies between the modalities. Therefore, there is some risk for rumors about the accuracy of each modality to arise.

This increased risk – compared to the classical treatment trail – may change our attitude about keeping the results from interim analyses secret. It may be an advantage to make the results of the interim analysis available for some or all people involved in the study just to avoid that rumors may reduce the willingness of clinicians or patients to participate in or to cooperate to the study. A common, but correct knowledge about the current state of the results with a clear indication of what has already been shown and what still needs to be shown may give a solid basis for a continued smooth conduct of the study and for a uniform interpretation of the results of the modalities in each patient.

Of course, such officially communicated and balanced information does not guarantee avoidance of any bias, as this information may still have an impact on those who evaluate the results of the diagnostic procedure. At the end of the day, we have to choose carefully between two options: a limited, but uncontrolled and heterogeneous spread of information (rumors, no interim analysis) or a controlled, but broader

Table 2. Sensitivity and specificity of ^{99m}Tc-MDP whole body bone scintigraphy and ^{18}F-fluorocholine PET/CT on a per-lesion basis, using MRI as a reference.

	^{99m}Tc-MDP whole body bone scintigraphy	^{18}F-fluorocholine PET/CT
Sensitivity (95% CI)*	36.4 % (22.4 %, 50.4 %)	87.5 % (76.8 %, 98.3 %)
Specificity (95% CI)*	81.8 % (72.5 %, 91.0 %)	89.9 % (81.2 %, 98.5 %)

*95% confidence intervals (95% CI) were adjusted for clustering of lesions in patients.

Table 3. Results of ^{99m}Tc-MDP whole body bone scintigraphy and ^{18}F-fluorocholine PET/CT in relation to the respective MRI results on a per-patient basis.

Results of test procedures		Results of MRI	
^{99m}Tc-MDP whole body bone scintigraphy	^{18}F-fluorocholine and PET/CT	Number of positive findings*	Number of negative findings**
Positive	Positive	20	0
Positive	Negative	1	0
Negative	Positive	2	0
Negative	Negative	0	19

*In these 23 patients at least one malignant lesion was found. **In these 19 patients only benign lesions were found.

and homogeneous spread of information by means of preplanned interim analyses.

Example

In a prospective diagnostic study at our institution we are currently investigating the detection of bone metastases from prostate cancer in patients with histologically confirmed prostate cancer and at least one bone metastasis at ^{99m}Tc-MDP whole body bone scintigraphy and no prior or active androgen deprivation [32]. ^{99m}Tc-MDP whole body bone scintigraphy is currently the method of choice for detecting bone metastases in these patients. However, the sensitivity and specificity of this image modality is suboptimal, and, therefore, we, as others, are looking for new diagnostic tools. We have examined the value of PET/CT with ^{18}F-fluorocholine, using MRI as a reference. For this interim analysis, data were collected from 42 consecutive patients from April 2009 to July 2011. The study was planned to be evaluated on a per-lesion basis, with positive and negative findings indicating malignant and benign lesions, respectively. However, due to recent developments about demands for demonstrating clinical benefits as well, we are today also interested in a patient-based analysis.

Both a lesion-based and a patient-based analysis were performed in the interim analysis. The per-lesion-based analysis using ^{18}F-fluorocholine showed a highly significant and clinically very relevant improvement in sensitivity, and even a moderate improvement in specificity (Table 2).

The results of the per-patient-based analysis with respect to the presence of at least one malignant lesion were less clear (Table 3).

The difference in sensitivity was estimated to be 4.4% with a 95% confidence interval of [-19%, 28%]. Its width of 47% indicates a rather imprecise estimate of the difference in sensitivity.

The results on the lesion level are very convincing, and there can be no doubt that the scientific community should be informed about this in a publication as soon as possible. However, assessment of the clinical benefit with respect to patient-related outcomes should be performed at patient level, and here the results are still inconclusive. So, a continuation of the study seemed justified.

The change from the originally lesion-based to a patient-based analysis has of course some impact on the sample size. If we stick to the original sample size assumptions with a power of 80% to detect a difference in sensitivity of 10% points, the formulas provided by Gerke et al. [30] would indicate that a total of 92 patients with at least one malignant lesion would have to be included. As we were able to observe that about half of the patients in the study population had at least one malignant lesion (which was also much higher than originally anticipated), this suggests that an overall sample size of 184 patients would be necessary. However, these numbers are based on the agreement rate of 3/23 observed in the patients with a malignant lesion, and this estimate is rather imprecise. Therefore, we would recommend performing a further interim analysis after the inclusion of 100 patients to obtain a more stable estimate of the agreement rate and to determine the final sample size.

Discussion

Interim analyses that are not carefully prepared in advance will always be a potential threat to the validity of a trial. For this reason they must be fully planned ahead of time and agreed upon by the members of the study group, and all details should be included in the study protocol. The general purpose of the interim analysis should be stated along with the procedures to be applied, the decision process, and the actions that should be taken based on the results of the analysis. Otherwise, interim analyses can be a source of inadequate decision-making and false-positive results, which hampers the scientific process and may even sometimes lead to fraud.

Interim analyses are commonly used in treatment studies for early stopping of trials if the treatment turns out to be beneficial or harmful. In diagnostic studies, interim analyses can serve different purposes, but their use is still uncommon. However, they can and should play a more prominent role as they can contribute (1) to a reduction of under- or overpowered diagnostic studies, (2) to a timely publication of the existing evidence, and (3) to minimize the risk inherent to rumors about emerging trends: 1) Paired diagnostic studies often suffer from the fact that *a priori* sample size calculations cannot be carried out properly because the agreement rate between the diagnostic modalities is normally not known in advance. This problem can be solved by recalculating the sample size after an interim analysis has been performed. 2) Paired diagnostic studies can serve multiple aims which can be addressed at different time points of a study's conduct, i.e., with different sample sizes. Therefore, there can be good reasons to continue the study in spite of convincing results concerning one sub-goal in an interim analysis, and sequential publication of partial results should be common practice. 3) The advantages of paired diagnostic studies are not compatible with long-term blinding. Therefore, there is a risk for rumors about emerging trends to arise, which can threaten the validity of a study in different ways. Informing all people involved in the study about the results of an interim analysis may contribute to a more uniform environment for the study and may antagonize unjustified rumors.

In general, we propose running a pilot study with 10 to 20 patients to check the feasibility of the diagnostic study and to test logistics (not least important with respect to interdepartmental collaborations). Such initial small pilot studies are a practical means to check logistics rather than a way of providing actual data that might influence the study design. Nevertheless, the results from such early pilot studies may at times be used as early tentative assumptions if a new procedure is being compared with a standard one for the first time. On the other hand, real interim analyses can and should be used for possible sample size adjustments since pilot studies generally do not produce sufficient data. Before starting the full study, the study protocol should be finished with a detailed plan for the interim analyses with respect to timing, analyses to be performed, and possible consequences with respect to sample size adjustment, internal and external dissemination of results, and continuation of the study.

Conclusion

Despite apparent similarities, interim analyses in treatment studies differ from interim analyses in paired diagnostic accuracy studies in at least three important ways. Namely the need for sample size adjustments, the possibility of early decision making without early stopping, and consequences of keeping interim results secret. If interim analyses are carefully planned, they can contribute to the reduction of under- and overpowered studies when initial reliable estimates of agreement rates cannot be obtained. In addition, they can ensure the early dissemination of relevant clinical results, and through controlled information sharing prevent inappropriate actions and promote a smooth study conduct.

Address correspondence to: Dr. Oke Gerke, Department of Nuclear Medicine, Odense University Hospital, Sdr. Boulevard 29, 5000 Odense C, Denmark. Phone: ++45 3017 1885, fax: ++45 6590 6192, E-mail: oke.gerke@ouh.regionsyddanmark.dk

References

[1] Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, Muti P, Jaeschke R, Guyatt GH. GRADE: assessing the quality of evidence for diagnostic recommendations. Evid Based Med 2008; 13: 162-163.

[2] Vach W, Høilund-Carlsen PF, Gerke O, Weber WA. Generating evidence for clinical benefit of PET/CT in diagnosing cancer patients. J Nucl Med 2011; 52 Suppl 2: 77S-85S.

[3] Golub HL. The need for more efficient trial designs. Stat Med 2006; 25: 3231-3235.

[4] Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. Stat Med 2003; 22: 727-739.

[5] DeMets DL. Clinical trials in the new millennium. Stat Med 2002; 21: 2779-2787.

[6] Ellenberg S, Fleming TR, DeMets DL. Data Monitoring Committees in Clinical Trials: A practical perspective. New York, Wiley Blackwell, 2002.

[7] DeMets DL, Lan KKG. Interim analysis: The alpha spending function approach. Stat Med 1994; 13: 1341-1352.

[8] Whitehead J. Design and Analysis of Sequential Clinical Trials. 2nd ed. New York, Wiley Blackwell, 1997.

[9] Bauer P, Koehne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994; 50: 1029-1041.

[10] Wassmer G, Eisebitt R, Coburger S. Flexible interim analyses in clinical trials using multistage adaptive test designs. Drug Inf J 2001; 35: 1131-1146.

[11] Vandemeulebroecke M. Group sequential and adaptive designs - a review of basic concepts and points of discussion. Biom J 2008; 50: 541 -557.

[12] Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biom J 2006; 48: 623-634.

[13] Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. Biom J 2006; 48: 635-643.

[14] Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11: 88-94.

[15] Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol 2012; 65: 282-287.

[16] Senn SS. Statistical issues in drug development. 2nd ed. Chichester, Wiley-Blackwell, 2008.

[17] Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York, Oxford University Press, 2003.

[18] Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. 2nd ed. New York, Blackwell-Wiley, 2011.

[19] Callstrom MR, Atwell TD, Charboneau JW, Farrell MA, Goetz MP, Rubin J, Sloan JA, Novotny PJ, Welch TJ, Maus TP, Wong GY, Brown KJ. Painful metastases involving bone: percutaneous image-guided cryoablation – prospective trial interim analysis. Radiology 2006; 241: 572-580.

[20] Jack CR Jr, Slomkowski M, Gracon S, Hoover TM, Felmlee JP, Stewart K, Xu Y, Shiung M, O'Brien PC, Cha R, Knopman D, Petersen RC. MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. Neurology 2003; 60: 253-260.

[21] Traugott AL, Dehdashti F, Trinkaus K, Cohen M, Fialkowski E, Quayle F, Hussain H, Davila R, Ylagan L, Moley JF. Exclusion of malignancy in thyroid nodules with indeterminate fine-needle aspiration cytology after negative ^{18}F-fluorodeoxyglucose positron emission tomography: interim analysis. World J Surg 2010; 34: 1247-1253.

[22] Edwards KR, Hershey L, Wray L, Bednarczyk EM, Lichter D, Farlow M, Johnson S. Efficacy and safety of galantamine in patients with dementia with Lewy bodies: a 12-week interim analysis. Dement Geriatr Cogn Disord 2004; 17 Suppl 1: 40-48.

[23] Rinnab L, Mottaghy FM, Simon J, Volkmer BG, de Petriconi R, Hautmann RE, Wittbrodt M, Egghart G, Moeller P, Blumstein N, Reske S, Kuefer R. ^{11}C-Choline PET/CT for Targeted Salvage Lymph Node Dissection in Patients with Biochemical Recurrence after Primary Curative Therapy for Prostate Cancer. Preliminary Results of a Prospective Study. Urol Int 2008; 81: 191-197.

[24] Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. Stat Med 1998; 17: 2635-2650.

[25] Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Stat Med 1998; 17: 891 -908.

[26] Alonzo TA, Pepe, MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. Stat Med 2002; 21: 835-852.

[27] Lu Y, Jin H, Genant HK. On the non-inferiority of a diagnostic test based on paired observations.

Stat Med 2003; 22: 3029-3044.

[28] Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clin Trials 2006; 3: 272-279.

[29] Bonett DG, Price RM. Confidence intervals for a ratio of binomial proportions based on paired data. Stat Med 2006; 25: 3039-3047.

[30] Gerke O, Vach W, Høilund-Carlsen PF. PET/CT in cancer – Methodological considerations for comparative diagnostic phase II studies with paired binary data. Methods Inf Med 2008; 47: 470-479.

[31] Newcombe RG, Altman DG. Proportions and their differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ: Statistics with confidence. 2nd ed. Bristol, BMJ Books, 2000, pp. 45-56.

[32] Poulsen MH, Petersen H, Høilund-Carlsen PF, Jakobsen JS, Gerke O, Karstoft J, Walter S. Detection of Bone Metastases from Prostate Cancer: A Prospective Study of ^{99m}Tc-MDP Bone Scintigraphy, ^{18}F-fluorocholine PET/CT, ^{18}F-fluoride PET/CT Compared with MRI. Moderated Poster at the 2012 Annual Meeting of the American Urological Association Education and Research Inc. Atlanta, Georgia, 19-23 May 2012.