Journal of Clinical Trials
ISSN: 2167-0870

Cohen KL, J Clin Trials 2015, 5:4
DOI: 10.4172/2167-0870.1000230

Research Article Open Access

Monitoring Reader Metrics in Blinded Independent Central Review of Oncology Studies

Kristin L Cohen[1], **Mithat Gönen**[2] **and Robert R Ford**[3*]

[1]*Janssen Pharmaceutical Research and Development, Titusville NJ, USA*

[2]*Memorial Sloan Kettering Cancer Center, New York, NY, USA*

[3]*Clinical Trials Imaging Consulting, LLC. Belle Mead, NJ, USA*

[*]**Corresponding author:** Robert R Ford, MD, Principal, Clinical Trials Imaging Consulting, LLC, 9 Raymond Lane, Belle Mead, NJ 08502, USA, Tel: 609I651I6887; Fax: 908I431I5940; E-mail: rfordmd@gmail.com

Rec date: March 20, 2015; **Acc date**: June 30, 2015; **Pub date**: July 02, 2015

Abstract

Purpose: Blinded independent central review (BICR) is advocated by regulatory authorities as a means of minimizing bias and independently verifying endpoints based on medical imaging when the data is intended to support pivotal trials. However, discordance between reviewers at the BICR raises concern with regulators. There are few published metrics related to discordance rates at the BICR and there is currently no standard metric which can be used to monitor reviewer performance in the BICR setting.

Methods: We analyzed BICR data from 29 oncology clinical trials including interpretations by 24 different radiologist reviewers of over 12,000 subject cases.

Results: The average reader acceptance rate was 48% and the rejection rate was 52%.

Conclusion: Based on our analysis, we propose the use of p-charts to monitor reviewer performance in oncology clinical trials employing BICR.

Introduction

Blinded independent central review (BICR) is the process by which radiographic exams and selected clinical data performed as part of a clinical trial protocol are submitted to a central location for blinded review by independent physicians not involved in the treatment of the patients. Regulatory authorities recommend BICR for oncology registration studies when the primary study endpoint is based on tumor measurements, such as progression-free survival (PFS), time to progression (TTP), or objective response rate (ORR) [1]. Clinical trial sponsors have also used BICR in Phase I and II studies to assist in critical pathway decisions including in licensing of compounds. There are different BICR review paradigms that are employed however current FDA guidance recommends multiple independent reviewers evaluating each subject [2]. BICR of industry-sponsored pivotal oncology studies generally includes the use of two independent radiologists evaluating each subject. This is commonly referred to as the "Two Readers and Adjudicator Paradigm" (Figure 1).

In the two readers and adjudicator paradigm, radiologist I (R1) reviews all images for a particular subject and determines an outcome based on the criteria that are established for the particular clinical trial. Commonly established validated response criteria include World.

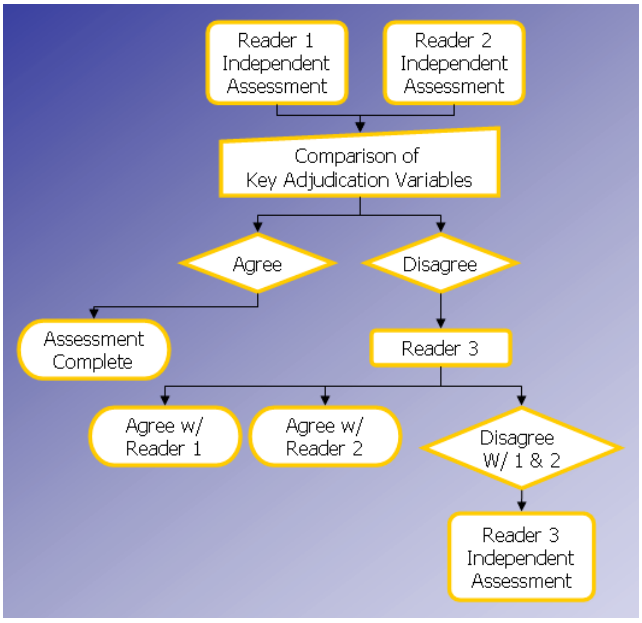

Figure 1: Shows one possible Imaging Core Lab approach to the two readers and adjudicator paradigm

Health Organization (WHO) [3], Response Evaluation Criteria in Solid Tumors (RECIST) [4,5], International Working Group (IWG) [6], International Harmonization Project (IHP) [7] and Macdonald Criteria8. Radiologist II (R2) performs the same function, reviews the same images independently, uses the same response criteria and determines an outcome. R1 and R2 are blinded to various components of the data. Reviewers are blinded to treatment arm (or any data that might un-blind the treatment arm), subject demographics, assessments made by the investigator, time point descriptions including whether scans are confirmatory or end of treatment, the total number of subject time points to be reviewed (preventing progression bias), the results or assessments of other reviewers participating in the BICR (except during adjudication), and any clinical data that may bias the independent reviewers.

At the completion of the review the outcomes are compared between the 2 reviewers. If there is agreement on the outcome variable, such as the date of progression in a study with a primary endpoint of progression-free survival (PFS) or time to progression (TTP), then two independent opinions have arrived at the same conclusion and no further review is needed. In the event there is discordance between the 2 reviewers on the defined outcome variable, a third radiologist (R3, the adjudicator) reviews the work completed by R1 and R2 and determines which of the 2 assessments is most accurate. In the event R3 does not agree with either R1 or R2, R3 re-reads the case from beginning to end and by default, this is the outcome.

Analysis of a single set of images by multiple reviewers will inevitably lead to some level of discordance if categorical variables are used to determine outcome. This has been reported in body imaging [8,9] and CNS imaging of high grade gliomas [10,11]. There are many factors influencing discordance rates among independent reviewers in the assessment of oncology clinical trial subjects. In a review of 31 oncology clinical trials across 10 different indications involving 8,752 subjects, we determined R1 and R2 agreed on the best radiographic response 77% of the time and on the date of radiographic progression 76% of the time (unpublished RadPharm data). Furthermore, statistical modeling studies we have reported (DIA Medical Imaging Continuum on October 2, 2008 and Food and Drug Administration on January 23, 2009) indicate the agreement rates for such endpoints have specific dependencies that include factors such as (but not limited to) therapeutic indication, average number of target lesions identified at baseline, average number of time points per subject, and the types of imaging exams required in the protocol. There are additional dependencies including lesion selection, inter-reader measurement variability, drug efficacy, duration of treatment, perception differences between reviewers, missing data, and image quality issues.

Discordance between reviewers raises concern among sponsors and regulators because the reasons for discordance are poorly understood and there are few published metrics related to BICR discordance rates. There is currently no standard metric that exists by which reviewer performance can be measured.

P Charts are engineering process quality control charts used to determine whether a business process is in statistical control. They are one of the seven basic tools used in quality control, the others being histograms, Pareto charts, check sheet cause and effect diagrams, flowcharts and scatted diagrams. The use of P Charts is optimal when counting outcomes of an event class (in this case, adjudication is the event class), an event has exactly 2 possible outcomes (in this case accept/reject), and each event is countable (in this case they are tracked). Additionally, P Charts take into account the size of the event class in setting limits (in this case the total number of subjects evaluated).

Materials and Methods

We selected 29 of the 350 oncology clinical trials on which we performed a BICR. Trial selection criteria included the use of a two readers and adjudicator reading paradigm and availability of data from a new company database in a format which could be queried. There were no other selection criteria considered and all trials satisfying these criteria were included. This review received an Institutional Review Board waiver as all data was blinded with respect to study sponsor, study protocol number, therapeutic agent under study, subject demographics and identifying information as required by the Health Insurance Portability and Accountability Act. The review was not blinded to indication, however all trials within a particular indication were blinded. The characteristics of each trial are summarized in Table 1.

Trail Id	Indications	Number of Subjects	Response Criteria
Clinical Trial 1	Bone	4458	OTHER
Clinical Trial 2	Brain	245	RECIST
Clinical Trial 3	Breast	2406	RECIST
Clinical Trial 4	Breast		RECIST
Clinical Trial 5	Breast		RECIST
Clinical Trial 6	Breast		RECIST
Clinical Trial 7	Breast		RECIST
Clinical Trial 8	Breast		RECIST
Clinical Trial 9	Breast		RECIST
Clinical Trial 10	Breast		RECIST
Clinical Trial 11	Colorectal	2575	RECIST
Clinical Trial 12	Colorectal		WHO
Clinical Trial 13	Colorectal		RECIST
Clinical Trial 14	Colorectal		RECIST
Clinical Trial 15	Colorectal		RECIST
Clinical Trial 16	Colorectal		RECIST
Clinical Trial 17	Colorectal		RECIST
Clinical Trial 18	GIST	117	RECIST
Clinical Trial 19	Head & Neck	22	RECIST
Clinical Trial 20	Kidney	820	RECIST
Clinical Trial 21	Kidney		RECIST
Clinical Trial 22	Lung	546	RECIST
Clinical Trial 23	Lung		RECIST
Clinical Trial 24	lymphoma	293	IWG
Clinical Trial 25	lymphoma		IWG

Clinical Trial 26	Lymphoma		IWG
Clinical Trial 27	Melanoma	365	RECIST
Clinical Trial 28	Melanoma		RECIST
Clinical Trial 29	Thyroid	167	RECIST

Table 1: Characteristics of clinical trials included in analysis.

A standard adjudication metrics report (Table 2) was generated for each trial. The report tabulated the following variables per independent reviewer: the number of cases read, the number of cases requiring adjudication, the adjudication rate for the particular reviewer, the number of adjudicated cases per reviewer which were accepted versus rejected by the adjudicating radiologist, as well as the associated reviewer acceptance and rejection rates.

Reader	#Cases Read	#Cases Adjudicated	Reader Adjudiottion Rate	I Cases Adjudication Accented	Cases Adjudication Rejected	Reader Acceptance Rate	Reader Rejection Rate
RI	309	94	30%	43	51	46%	54%
R2	284	81	29%	51	30	63%	37%
R3	188	83	44%	32	51	39%	61%
R6	31	6	19%	4	2	67%	33%
R7	253	68	27%	38	30	56%	44%
RIO	127	57	45%	22	35	39%	61%
R13	156	50%	32%	26	24	52%	48%
R21	245	68%	28%	33	35	49%	51%
R22	247	81%	33%	41	40	51%	49%
#A standard report was generated for each protocol which tabulated metrics for each radiologist assigned to read for the given study. This particular study had nine radiologists assigned as reviewersa each is identified with a unique reader ID.							

Table 2: Protocol specific adjudication metrics per reviewer.

The reviewer adjudication rates were calculated using the number of cases which required adjudication divided by the total number of cases read. The reviewer acceptance rates were calculated using the number of cases where the adjudicator accepted the given reviewer's assessment during adjudication divided by the number of cases which underwent adjudication. The rejection rates were calculated in the same manner, using the number of adjudicator rejected assessments as the numerator. In addition to determining protocols specific metrics, data from all individual trial reports were pooled and company-wide statistics were determined for each of the 24 reviewers. This is illustrated in Table 3.

Reader	*Cases Read	#Cases Adjudicated	Reader Adjudication Rate	*Cases Adjudication Accepted	#Cases Adjudication Rejected	Reader Acceptance Rote	Reader Rejection Rate
R1	4111	1232	30%	583	649	47%	53%
R2	2659	615	23%	328	287	53%	47%
R3	2252	616	27%	242	374	39%	61%
R4	103	31	30%	20	11	65%	35%
RS	94	29	31%	11	18	38%	62%
R6	582	169	29%	79	90	47%	53%
R7	2970	893	30%	499	394	56%	44%
R8	112	33	29%	14	19	42%	58%
R9	171	54	32%	26	28	48%	52%
R10	1113	427	38%	144	283	34%	66%

R11	153	51	33%	27	24	53%	47%
R12	82	26	32%	12	14	46%	54%
R13	2139	598	28%	298	300	50%	50%
R14	117	37	32%	15	22	41%	59%
R15	179	82	46%	36	46	44%	56%
R16	116	31	27%	16	15	52%	48%
R17	123	43	35%	19	24	44%	56%
R18	919	303	33%	145	158	48%	52%
R19	105	32	30%	12	20	38%	63%
R20	158	49	31%	34	15	69%	31%
R21	3158	840	27%	457	383	54%	46%
R22	2366	646	27%	369	277	57%	43%
R23	47	15	32%	5	10	33%	67%
R24	199	80	40%	40	40	50%	50%

Table 3 is an example of a company-wide adjudication metrics report. Data from the individual trial reports were pooled and overall metrics were tabulated for each of the 24 reviewers across the 29 studies. The pool of reviewers consisted of current BICR radiologists as well as past BICR radiologists. Each of the 24 radiologists was an assigned reviewer for a subset of the 29 studies there were no reviewers assigned to all 29 studies. This table summarizes data from 12,014 total subject cases each assessed by 2 readers, resulting in 3,466 adjudications, and 6,932 adjudication accept/reject outcomes.

Table 3: Company-wide adjudication metrics per reviewer based on pooled data from 29 clinical studies

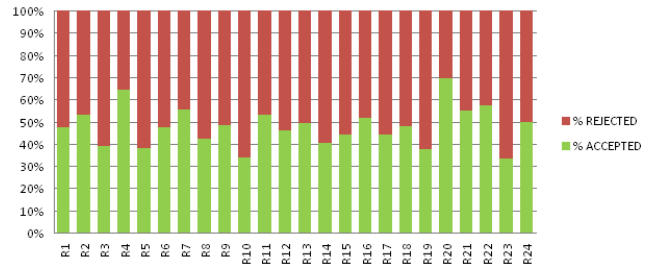

Figure 2: Summarizes the reviewer acceptance and rejection rates resulting from 3,466 adjudications.

Acceptance and rejection rates were plotted for each reviewer (Figure 2). P-charts [12,13] were used to analyze the percentage of adjudicated cases per reviewer which were accepted versus rejected by the adjudicating radiologist. P-chart analysis (PCA) was conducted to evaluate the distribution of acceptance rates across the 24 reviewers to identify if a given reviewer fell outside of acceptable boundaries. Control and warning limits were calculated as follows:

Control Limits = $p \pm 3\sqrt{p(1-p)/n}$and

Warning Limits = $p \pm 2\sqrt{p(1-p)/n}$,

where p is the proportion rejected and n is the number of samples for a given reviewer. Note: There are no widely-used thresholds for a minimum sample size so that the resulting limits are deemed to be reliable. Standard statistical recommendations for the binomial distribution include both np>5 and n(1,p)>5 for the approximately normality of the sampling distribution of p, so these can be used as a rough guide.

Results

Figure 3 demonstrates PCA can be used to analyze reviewer performance and identify outliers based on upper and lower limits generated during the analysis. Figure 3 represents a corporate-wide assessment of independent reviewers. The blue dots represent the adjudication acceptance rates for each individual reviewer. Based on the distribution of acceptance rates across reviewers and the number of cases completed per reviewer, PCA generates upper and lower limits. The upper and lower limits shown in green represent warning limits. The upper and lower limits shown in red represent action limits. Warning and action limits are two and three standard deviations on either side of the mean. A data point outside the warning limits has approximately 5.5% probability of coming from the same distribution as the other data point. This probability drops to 0.3% for action limits. These limits are different than confidence intervals. Control limits are boundaries for individual observations (calculated using the standard deviation) and confidence limits are boundaries for the mean (calculated using the standard error). The limits in Figure 3 are not the same across readers, because the sample size for each reader is different.

In this example, each of the 24 reviewers is identified on the horizontal axis. The vertical axis indicates the percentage of adjudicated cases per reviewer which resulted in an adjudicator accepted outcome. The blue dots represent the adjudication acceptance rates for each individual reviewer. The green lines

represent warning limits, whereas the red lines represent action limits. PCA takes into account the number of cases read, hence the saw tooth pattern of the warning and action limits. If all readers had read the same number of cases, the green and red limit lines would be horizontal. In this PCA, given the distribution of acceptance rates among all 24 reviewers and the number of cases completed by each reviewer, R3 and R10 are "outside the box" with an event rate crossing below the lower action limit. This means R3 and R10 have a lower than expected acceptance rate and further investigation is needed to determine the cause of the outliers. R7, R21 and R22 are "outside the box", as they have higher than expected acceptance rates. In this case, no action is required.

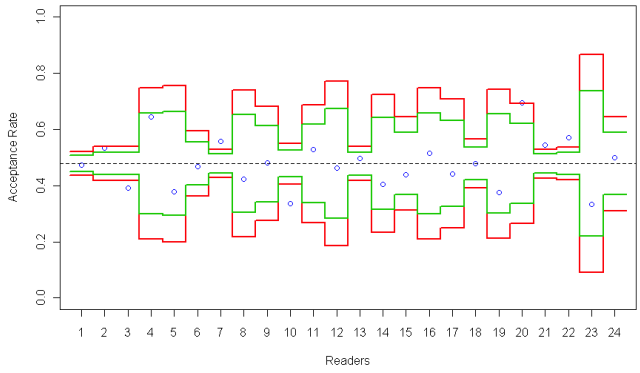

Figure 3: PCA consensus agreement rates of all readers across 29 clinical studies

Readers with acceptance rates which fall outside of the upper and lower boundaries may require investigation and/or action. The corporate-wide PCA shown in Figure 3 indicates 22 of the 24 reviewers are performing within acceptable limits given the distribution of acceptance rates among all 24 reviewers as well as the number of cases each reviewer has completed. The PCA indicates 2 of the 24 reviewers (R3 and R10) are "outside the box" with acceptance rates falling below the red action limit. Based on this finding, investigation would be needed and action may need to be taken depending on the outcome of the investigation. R7, R21 and R22 are "outside the box", however in a positive direction. These reviewers have higher than expected acceptance rates given the distribution of the data and the number of cases each has completed. Hence they are known as "overachievers" and no action is required.

A protocol-specific PCA representing a clinical trial with 920 subject reads completed is shown in Figure 4. The PCA indicates the performance of all reviewers is within acceptable limits, demonstrating the review process is within statistical control.

Using P-chart Analysis to Monitor Reviewer Performance during BICR. There are multiple factors that can result in outcome discordance between independent radiology reviewers. Some factors are due to justifiable interpretation differences among readers such as lesion selection, inter-reader measurement variability, perception of new lesions, and the qualitative assessment of non-measurable disease. Missing data and image quality issues also impact response assessment which can contribute to the adjudication rate. Separate from the above factors which are unavoidable in the independent review setting, are reader assessment errors that are occasionally identified.

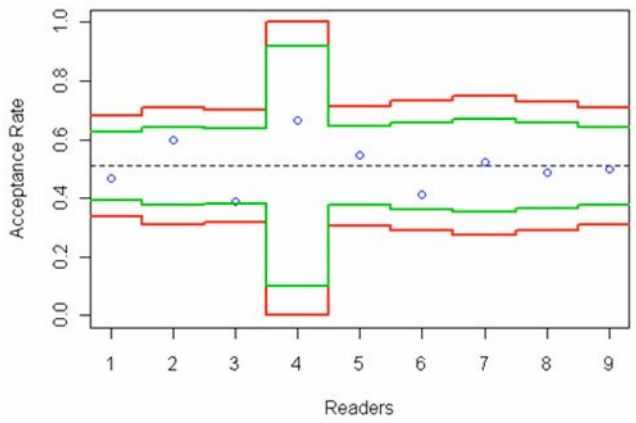

Figure 4: PCA consensus agreement rates for a specific protocol (In this protocol, there were nine radiologists assigned as reviewers on the study, with a total of 920 subject reads completed. All readers were "inside the box" indicating the review process is within statistical control).

The two reader and adjudicator paradigm allows surveillance and resolution of these assessment errors. This is supported by Figure 2 which summarizes the acceptance and rejection rates for the 24 independent reviewers. The fact that the acceptance and rejection rates tend to fall along the 50th percentile suggests many adjudications are the result of a justifiable difference between reviewers where neither reviewer is incorrect in their assessment. However, there is still discordance in outcome and a "coin flip" decision must be made by the adjudicator. Inter-reader reliability (based on different reviewers assessing the same set of images independently) and intra-reader reliability (based on the same reviewer assessing the same images at a later date once memory effect has resolved) is an area of concern among sponsors and regulators. We are proposing the use of p-chart analysis (PCA) in qualifying readers and monitoring reviewer performance in the BICR setting.

A p-chart is a control chart, commonly used in engineering process quality control to determine if a measurement process is within statistical control. P-charts summarize data collected in subgroups (subgroups can be of varying sizes). The "P" in p-chart stands for the p (proportion of successes) of a binomial distribution. PCA can be used when an event class has exactly two possible outcomes which can be counted or tracked. In this case, the event class was adjudication and the two possible outcomes included adjudicator acceptance and adjudicator rejection. In Figure 3, the subgroups along the horizontal axis represent radiologists with varying numbers of cases read. The process attribute (or characteristic) of a p-chart is described in yes/no, pass/fail form. In Figure 3, the attribute (along the vertical axis) represents the percentage of "accepted" assessments resulting from adjudication (relating to an adjudicating radiologist accepting or rejecting the assessment of the given reviewer).

P-charts analyze event rates by taking into account the distribution of the data (acceptance rates) among subgroups (individual reviewers) as well as the subgroup sample sizes (in this case, the number of cases completed by each reviewer). Figure 3 illustrates the percentage of adjudicated cases (per reviewer) where the adjudicator accepted the specified reader's assessment. Of the 12,014 cases read across the 29

studies, 3,466 cases required adjudication. Each adjudication resulted in two outcomes, an accepted outcome for one reader and a rejected outcome for the other reader. Therefore, the corporate-wide PCA of consensus agreement rates shown in Figure 3 consists of data from 6,932 adjudication outcomes.

PCA can be used to monitor the proportion of reader events in a sample and facilitate the decision as to whether the reviewer's performance requires attention. As illustrated in Figure 2, PCA can be conducted on a corporate-wide level across all protocols in order to qualify reviewers. PCA can also be used to monitor performance on a protocols specific basis as shown in Figure 4. In the event a reader was previously qualified as a reviewer but in a particular protocol was found to be "outside the box", investigation would be needed. Perhaps the reader requires additional training on a particular aspect of the criteria used to determine response. Alternatively, perhaps the radiologist was assigned as a reviewer on a trial that was outside his or her expertise, although one would have hoped to discern this as part of the reader vetting process.

In addition to monitoring reviewer acceptance rates, there are multiple other event rates which can be evaluated using PCA. For example, as shown in Figure 5, an analysis can be done regarding the number of times a reviewer's assessment required adjudication (the reviewer adjudication rate).

As PCA identifies outliers in a given process, it can be used to monitor performance across a group of reviewers who are following the same procedures. Therefore, PCA can be used to compare reviewer performance within an imaging core lab, but in some cases, it would be inappropriate to use PCA to compare reviewer performance across different imaging core labs. For example, there are different adjudication paradigms which are followed by different imaging core labs. Imaging Core Lab A may require the independent selection of lesions at baseline by R1 and R2, while Imaging Core Lab B may have a process in place which requires consensus of lesion selection by R1 and R2 at baseline. It would not be appropriate to plot the adjudication rates for each reviewer at Imaging Core Labs A and B on the same p-chart, as the reviewers are not following the same review process. Reviewer performance within Imaging Core Lab A could be monitored using PCA while the performance of reviewers within Imaging Core Lab B could be monitored using a separate PCA. There are, however, examples where reader performance could be compared across imaging core labs if the differences in process would not influence the event outcome. For example, the different processes followed by Imaging Core Labs A and B would not impact the PCA of acceptance rates across reviewers. Even though the different processes may influence the number of cases requiring adjudication, a given reviewer's acceptance versus rejection rate of cases which were adjudicated is not likely to be impacted by the different procedures. As a result, PCA may be used to evaluate reviewer performance within the same imaging core lab, but additional considerations may be needed if attempting to use the same PCA to evaluate performance across multiple imaging core labs.

Because PCA is used to identify outliers based on the distribution of the data across a number of subgroups (reviewers), a pitfall of PCA is that if a large number of reviewers are performing at an unacceptable level, PCA may incorrectly indicate their performance is within acceptable limits.

Conclusion

In summary, p-chart analysis (PCA) is a method of qualifying reviewers on a corporate-wide level and monitoring reviewer performance in BICR of oncology studies.

References

1. Ford R, Schwartz L, Dancey J, Dodd LE, Eisenhauer EA, et al. (2009) Lessons learned from independent central review. Eur J Cancer 45: 268-274.
2. United States Food and Drug Administration Guidance for Industry (2004) Developing medical imaging drug and biologic products part three: design, analysis and interpretation of clinical studies. Rockville, MD.
3. Miller AB, Hoogstraten B, Staquet M, Winkler A (1981) Reporting results of cancer treatment. Cancer 47: 207-214.
4. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, et al. (2000) New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 92: 205-216.
5. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, et al. (2009) New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 45: 228-247.
6. Cheson BD, Horning SJ, Coiffier B, Shipp MA, Fisher RI, et al. (1999) Report of an international workshop to standardize response criteria for non-Hodgkin's lymphomas. NCI Sponsored International Working Group. J Clin Oncol 17: 1244.
7. Cheson BD, Pfistner B, Juweid ME, Gascoyne RD, Specht L, et al. (2007) Revised response criteria for malignant lymphoma. J Clin Oncol 25: 579-586.
8. Macdonald DR, Cascino TL, Schold SC Jr, Cairncross JG (1990) Response criteria for phase II studies of supratentorial malignant glioma. J Clin Oncol 8: 1277-1280.
9. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, et al. (1996) Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol 167: 851-854.
10. Vos MJ, Uitdehaag BM, Barkhof F, Heimans JJ, Baayen HC, et al. (2003) Interobserver variability in the radiological assessment of response to chemotherapy in glioma. Neurology 60: 826-830.
11. Provenzale J, Ison C, DeLong D (2008a) Interobserver variability in bidimensional measurements of brain tumors. Am J Roentgenol 190: A43-A46.
12. Shewhard WA (1931). Economic Control of Quality of Manufactured Product. D. Van Nostrand Company, New York.
13. Grant EL, Leavenworth RS (1972) Statistical Quality Control. McGraw Hill, New York.