Clinical Research and Regulatory Affairs, 2010; 27(2): 42–51

informa
healthcare

REVIEW ARTICLE

Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol

Heejung Bang[1], Stephen P. Flaherty[2], Jafar Kolahi[3], and Jongbae Park[2]

[1]Division of Biostatics and Epidemiology, Department of Public Health, Weill Medical College of Cornell University, New York, NY, USA, [2]Asian Medicine and Acupuncture Research, Department of Physical Medicine and Rehabilitation, University of North Carolina at Chapel Hill. Chapel Hill, NC, USA, and [3]Torabinejad Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

Abstract
There is strong consensus in the clinical trial community that blinding is an important issue in randomized controlled trials. At present grossly incomplete reporting of procedures and the use of *any* assessment for blinding still prevails. The term 'double-blind' has almost become a convention without any checks or balances. Also there is a lack of consensus on quantitative procedures for evaluating the success of blinding in the literature. This article reviews statistical methods of blinding assessment along with software options, and discusses some of the most pressing issues surrounding the acquisition, interpretation, and reporting of blinding data. Finally, it proposes a sample blinding assessment protocol to address some of these issues.

Keywords: *Blinding assessment; blinding index; blinding protocol; blinding questionnaire; masking*

Introduction

Human behavior is influenced by what we know, what we believe, and our temptation to find out what is going on. Naturally, patients and physicians/investigators who participate in randomized controlled trials (RCT) are likely to want to know what treatment they received or assigned in experimental settings. Maintaining blindness in relevant parties (e.g. patients, physicians, and/or data analysts) throughout the trial, whenever feasible and justifiable, is an important undertaking to assure the internal validity of the study findings, yet it can be difficult to accomplish due to varied reasons. Compared to the highly rigorous standards of clinical trial planning and operation, RCT investigators often appear to be overly generous or cavalier on one element of the process, blinding.

Successful blinding could reduce ascertainment, performance, and information biases and improve compliance and retention. Particularly, information bias may occur at various points in a trial including data reporting, collection, and outcome assessment on the parts of physicians, study coordinators, and even statisticians. Also, if a patient was assigned to a placebo or sub-optimal treatment and she finds out this fact, she might be less enthusiastic about her participation in the trial (such as a higher probability of dropping out of the study or discontinuing to take the assigned medications.). Problems associated with unblinding can be more severe when the outcome measures are subjective (e.g. pain, cold, depression) or when a physician has a conflict of interest.

Some historical examples outlining why blinding is important provide a background for this review. In a single-blind, placebo-controlled trial investigating the effect of zinc on taste disorders, Henkin et al. (1) found evidence for a statistically significant benefit. However, an identical trial (with only one difference, a double-blind design) showed no such benefit. Among the problems identified were the responses were very

Address for Correspondence: Heejung Bang, PhD, 402 East 67th Street, New York, NY 10065, USA. Tel: (646) 962-8015. Fax: (646) 962-0281. E-mail: heb2013@med.cornell.edu

(Received 02 November 2009; revised 16 February 2010; accepted 14 March 2010)

ISSN 1060-1333 print/ISSN 1532-2521 online © 2010 Informa UK Ltd
DOI: 10.3109/10601331003777444

subjective and there was a suspicion of 'vested interest' (2). Another study reported that expectation can be more influential than treatment itself in some pain studies and reinforced the importance of the *placebo effect* in clinical research (3). Also, even the Women's Health Initiative (WHI), a landmark study as the largest RCT of its kind, which provided not only enormously useful data but also some tantalizing and conflicting findings, was found to have unsatisfactory blinding maintenance that justified further analyses (4, 5). Indeed, Garbe and Suissa (6) showed detection bias due to unblinding could lower the crude rate ratio of 1.28 to 1.02 in the WHI.

Many authoritative statements and/or recommendations on blinding have been put forth, including:

> [drug name]-related sided effects have the potential to unblind subjects and investigators. Unblinding may result in ascertainment bias of subjective study endpoints. We recommend that you administer a questionnaire at study completion to investigate the effectiveness of blinding the subjects and treating and evaluating physicians (Office of Therapeutics Research and Review, Center for Biologics Evaluation and Research, FDA 2003).
>
> DRUDP requests that subjects and investigators state at the end of the subject's participation as to what treatment assignment they think was made, in order to assess the adequacy of blinding (Office of Drug Evaluation III Center for Drug Evaluation and Research, FDA 2005).
>
> Investigators should report whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated (Item 11 among the 22 items in the Consolidated Standards of Reporting Trials (CONSORT) statement (7, 8) (http://www.consort-statement.org/).

The International Committee of Medical Journal Editors also provided a similar recommendation on blinding (9).

Although many seem to agree that it is important for RCTs to adopt some form of blinding (e.g. single, double, or triple), whenever relevant, blinding is usually simply reported as being accomplished (with a sentence or two routinely added in the protocol and publications, especially in the title and keywords). However, the success of blinding procedures and/or status is rarely measured or reported. Boutron et al. (10) reviewed blinding in a total of 90 RCT reports and concluded that methods and reporting were inconsistent and questionable. Fergusson et al. (11) showed that only seven of 97 (7%) general medicine RCTs and eight of 94 (8%) psychiatric RCTs reported the success of blinding. Also, Hróbjartsson et al. (12) showed that only 31 out of 1599 trials (2%) reported tests for the success of blinding. Based upon such results, the call for urgent improvement has been repeatedly made (13).

Responding to such a call for improvement in reporting of blinding assessment, we think it timely to identify potential causes of or difficulties in the lack of reporting of blinding-related data and analyses. We reckon informing clinical researchers of methods of analysis and reporting would improve this issue.

In this article, we focus on reviewing the existing statistical methods used to assess blinding and we hope to offer practical solutions to help clinical researchers understand how to use them, along with outlining the assumptions and properties (e.g. advantages and disadvantages) of each method. It is highly likely that some, if not many, clinical researchers may not be aware of the existence of blinding assessment methods and this may have hampered or discouraged researchers from collecting the data necessary for reporting blinding success. We will also discuss other issues pertaining to blinding that have not been fully resolved. Finally, we propose a sample blinding survey and assessment protocol for clinical trialists' consideration regarding the design of blinding questionnaires, the collection of the blinding data, and the assessment and reporting of blinding success (in Appendix; online version only).

Review of methods

Randomization and masking schemes

Various efforts have been directed toward the technical aspects of treatment masking and randomization. Investigators try to disguise the dissimilarities between treatments under comparison as much as possible (e.g. taste, smell, appearance, mode of delivery). Ideally, treatment identity should be tested prior to the trial in a (small) independent group of people who are not a part of the actual study (14). To help preserve the integrity of treatment assignment, blocks of random length (or large blocks) are used as a standard practice (http://www.fda.gov/cder/guidance/ICH_E9-fnl.pdf) with artificial treatment codes in documentation and statistical programming (e.g. A and B, possibly alternated with C and D).

Data required for blinding assessment

A blinding questionnaire or survey is the most prevalent method for obtaining the data needed for blinding assessment. Common formats of the blinding questionnaire ask both the control and experimental groups to guess which treatment they received. The two most common formats include: (1) three response categories for treatment guess: ‘New treatment’, ‘Placebo (or control)’, or ‘Don’t know (DK)’, and (2) five response categories: ‘Strongly believe the treatment is new treatment’, ‘Somewhat believe the treatment is new treatment’, ‘Somewhat believe the treatment is placebo’, ‘Strongly believe the treatment is placebo’, or ‘DK’. We may re-ask those who answered DK initially to make their best guess regarding their assignment (see Tables 1–3).

A unique methodological issue in blinding research is how to handle ‘DK’ data. Some investigators do not allow DK responses and force participants to guess from the beginning, although we generally believe this is not a good idea even though this strategy may simplify the ensuing statistical analyses. DK can be quite different from the missing or typical non-response data that are common in surveys and have been extensively studied. Indeed, DK is the preferred outcome of blinding as long as it is an honest answer, but it is prone to ‘social desirability bias’ or ‘lie’, especially when the data are collected during face-to-face interviews (15).

Statistical methods for blinding assessment

Existing statistical methods that can be utilized for blinding assessment include:

1. *Chi-Square and McNemar’s tests*: A traditional Chi-Square test for a contingency table was used by Hughes and Krahn (16) and Margraf et al. (17), among others. Kolahi et al. (18) used McNemar’s test. As test statistics, these methods basically test the independence of two variables (represented by row and column). Therefore, they provide *p*-values for statistical testing but not a numerical measure of blinding itself.
2. *Kappa statistic*: The standard Kappa statistic was also used (19). However, Kappa measures agreement rather than *disagreement*, which is a desirable outcome in blinding. Therefore, how to interpret the Kappa statistic proves tricky (e.g. how low is good enough?).

Beyond these standard statistical methods that may not handle DK responses and the emphasis of disagreement properly, two blinding indexes (BIs) have been developed.

Table 1. Data for blinding assessment (2 × 3 format).

	Guess			
Assignment	New treatment	Placebo	DK	Total
New treatment	n_{11} ($P_{1\|1}$)	n_{12} ($P_{2\|1}$)	n_{13} ($P_{3\|1}$)	$n_{1.}$
Placebo	n_{21} ($P_{1\|2}$)	n_{22} ($P_{2\|2}$)	n_{23} ($P_{3\|2}$)	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

$P_{j|i} = P$(guess j|assigned treatment i) for $i = 1$ (new treatment), 2 (placebo), and $j = 1$ (new treatment), 2 (placebo), 3 (DK), where DK denotes ‘Don’t know’. N is the total number of participants.

Table 2. Data for blinding assessment (2 × 5 format).

	Guess					
Assignment	1	2	3	4	5 (DK)	Total
New treatment	n_{11} ($P_{1\|1}$)	n_{12} ($P_{2\|1}$)	n_{13} ($P_{3\|1}$)	n_{14} ($P_{4\|1}$)	n_{15} ($\mathrm{P}_{5\|1}$)	$n_{1.}$
Placebo	n_{21} ($P_{1\|2}$)	n_{22} ($P_{2\|2}$)	n_{23} ($P_{3\|2}$)	n_{24} ($P_{4\|2}$)	n_{25} ($\mathrm{P}_{5\|2}$)	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	N

1: strongly believe new treatment; 2: somewhat believe new treatment; 3: somewhat believe placebo; 4: strongly believe placebo; and 5: DK.

Table 3. Data for blinding assessment. Data obtained by re-asking the subjects who answered DK from Tables 1 or 2.

	Guess		
Assignment	New treatment	Placebo	Total
New treatment	$\tilde{n}_{11}$ ($\tilde{P}_{1\|1}$)	$\tilde{n}_{12}$ ($\tilde{P}_{2\|1}$)	$\tilde{n}_{1.}$
Placebo	$\tilde{n}_{21}$ ($\tilde{P}_{1\|2}$)	$\tilde{n}_{22}$ ($\tilde{P}_{2\|2}$)	$\tilde{n}_{2.}$
Total	$\tilde{n}_{.1}$	$\tilde{n}_{.2}$	$\tilde{n}$

3. *James et al.'s BI*: James et al. (20) proposed a BI via modification of the Kappa statistic. The BI is defined as:

James' BI = $\{1 + P_{\text{DK}} + (1 - P_{\text{DK}})^{*}K_{\text{D}}\}/2$, where

$$P_{DK} = P(DK),$$

$$K_{D} = \left(P_{Do} - P_{De}\right)/P_{De},$$

$$P_{Do} = \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} P_{ij} / (1 - P_{DK}),$$

$$P_{De} = \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} P_{.j} (P_{i.} - P_{i3}) / (1 - P_{DK})^{2}, \text{when } P_{DK} \neq 1,$$

and $P_{ij} = n_{ij}/N$, and dot (.) in the subscript denotes summation.

The following weights were suggested: $w_{ij} = 0$ for correct guess; 0.5 for incorrect guess; and 1 for DK. Unknown parameters (P's) in the BI can be estimated from observed data.

We provide a brief summary of how to interpret the BI in the box below:

The asymptotic variance of the BI can be derived, or alternatively the Jackknife method can be employed.

$0 \leq$ James' BI ≤ 1.
If $P_{\text{DK}} = 1$, then BI = 1 (complete blinding);
If $P_{\text{DK}} = 0$ and $P_{\text{Do}} = 0$ (e.g. all responses are correct), then BI = 0 (complete unblinding);
If $P_{\text{DK}} = 0$ and $P_{\text{Do}} = P_{\text{De}}$ (e.g. 50% correct and 50% incorrect), then BI = 0.5 (random guessing).
Unblinding may be claimed: if two-sided confidence interval (CI) does not cover 0.5. (i.e. if the upper bound of the CI is below 0.5).

In this index, DK is very influential; it assumes that DK is truly DK, not just a socially desirable answer. (For example, it is possible that patient and/or physician may answer DK even though they found out somehow or are highly sure about the treatment identity.) The BI offers one summary measure that combines two arms, which may be useful as a simple index but can not detect different behaviors in the two arms that may represent qualitatively different scenarios due to the combining process.

4. *Bang et al.'s BI*: To address some of the limitations noted above, another index was developed by Bang et al. (21).

If we collect data in the 2 × 3 format (as in Table 1), let us define $r_{i|i} = P_{i|i}/(P_{1|i} + P_{2|i})$, ($i = 1$ for drug and $i = 2$ for placebo) which is the proportion of correct guesses among participants who provided certain guesses other than DK in the i^{th} arm. Bang's BI$_i$ is defined as $(2r_{i|i} - 1)^{*}(P_{1|i} + P_{2|i})$ and can be estimated by:

$$\text{Bang's}\,\hat{}\,\text{BI}_{\text{i}} = \left(2\frac{n_{ii}}{n_{i1} + n_{i2}} - 1\right)^{*}\left(\frac{n_{i1} + n_{i2}}{n_{i1} + n_{i2} + n_{i3}}\right)$$

and its variance formula is given by:

$$Var\left(Bang's\,\hat{}\,\text{BI}_{\text{i}}\right) = \left\{\text{P}_{1|i}\left(1 - P_{1|i}\right) + \text{P}_{2|i}\left(1 - P_{2|i}\right) + 2P_{1|i}P_{2|i}\right\}/n_{i.}$$

where variance can be estimated by substituting estimates for unknown parameters P's.
Note that:

1. Bang's BI$_i$ is identical to $P_{1|1} - P_{2|1}$ for the drug arm and $P_{2|2} - P_{1|2}$ for the placebo arm under trinomial distribution.
2. Without DK observations, Bang's BI$_i$ reduces to $2r_{i|i} - 1$.

For data in the 2 × 5 format (as in Table 2) with or without validation data for DK (as in Table 3),

Bang's BI$_i = P_{1|i} + w_{2|i}\ P_{2|i} + w3_{1|i}\ P3_{1|i} - P_{5|i} - w_{4|i}\ P_{4|i} - w3_{2|i}\ P3_{2|i}$
subject to $0 \leq w3_{1|i} = w3_{2|i} \leq w_{2|i} = w_{4|i} \leq 1$ and $P_{1|i} + P_{2|i} + P3_{1|i} + P3_{2|i} + P_{4|i} + P_{5|i} = 1$, where the authors suggested the symmetric weights of $w3_{1|i} = w3_{2|i} = 0.25$ and $w_{2|i} = w_{4|i} = 0.5$. Other weights may be used for sensitivity analyses.

A rule of thumb explanation on how to interpret Bang's BI is provided below:

$-1 \leq$ Bang's BI ≤ 1.
If $r_{i|i} = 1$ and $n_{i3} = 0$ (e.g. all responses are correct), BI$_i$ = 1 (complete unblinding).
If $r_{i|i} = 0$ and $n_{i3} = 0$ (e.g. all responses are incorrect), BI$_{\text{i}}$ = –1 (opposite guessing*).
If $r_{i|i} = 0.5$ (e.g. 50% correct and 50% incorrect among participants with certain identification),
BI$_i$ = 0 (random guessing).
BI$_i$ can be interpreted as 'the proportion of participants who answered correctly on the i^{th} arm beyond chance level.'
Unblinding may be claimed: if one-sided CI does not cover 0.**
* how to interpret can be subjective because this may represent complete blinding or complete unblinding in opposite direction.
** the authors attempted to de-emphasize its role in its use for statistical testing (i.e. rejecting vs not rejecting the null hypothesis of successful blinding). See more details in text.

Bang's BI is directly interpreted as the percentage of unblinding beyond chance and can capture different behaviors in different arms. Particularly, the 'wishful thinking' or 'lack of idea about control treatment' scenario that makes patients believe they are on active treatment may be revealed by this index. For example, in acupuncture studies, a majority of patients who received sham acupuncture tend to believe they received real acupuncture. This may be because some of the patients have no idea what sham acupuncture looks like or how it operates and/or tend to wish they were assigned to the real treatment despite their consent for randomization (3, 22, 23).

The Bang's BI tends to be more sensitive than the James' BI (i.e. more easily rejects the null hypothesis of successful blinding with a small difference as sample size increases), and this common statistical property may not be welcomed by investigators. However, the present authors believe that the most useful aspect of Bang's BI lies in estimation rather than significance testing (21, 24), since the ultimate determination of the success of blinding can be a rather complicated problem, where numerical estimation may serve only a part. Introduction of blinding scenarios to interpret the BIs in conjunction of the outcome finds a room in the rationale; see Other issues below. Later we highlight that not all unblinding scenarios are undesirable. [In this situation, the term 'unblinding' can be misleading. However, for consistent and simple presentation, we decided to use 'unblinding' to denote the situations where a high proportion of people guess their treatments correctly, where 'high' may be determined by a threshold set by investigators a priori or defined as 'beyond random chance level'.]

Examples

We introduce two clinical trials that assessed blinding in the literature.

Example 1: Acupuncture for sub-acute stroke rehabilitation

Acupuncture is used as an adjunctive therapy that may reduce persistent disability after stroke. Park et al. (22) conducted a sham-controlled, subject- and assessor-blinded randomized trial to evaluate the utility of acupuncture for recovery in activities of daily living (as the primary outcome, measured by the change in Barthel score) and health-related quality-of-life after stroke (as secondary outcomes). Blinding data from patients were collected at the end of 2 weeks of treatment and are summarized in the 2*3 format (see Table 4).

Note that James' BI indicates successful blinding, while Bang's BI shows that significantly higher proportions of patients in both arms (than would be expected by chance alone) reporting they received acupuncture. The observed blinding pattern may be understood as 'wishful thinking' and/or 'lack of idea about control treatment' scenarios, which are not infrequent in acupuncture studies.

Example 2: Warfarin-aspirin symptomatic intracranial disease (WASID) trial

Hertzberg et al. (25) investigated if the use of dose modification schedule is effective for blinding trials of warfarin in the WASID study (26). They also compared their analysis with the blinding data collected from the Stroke Prevention In Non-rheumatic Atrial Fibrillation (SPINAF) trial (27). The blinding data and evaluation from these two comparable trials are presented in Tables 5 and 6.

Based on James' BI, blinding was successful in all cases. Bang's BI, however, showed that 12–39% of people (i.e. physicians, coordinators, patients) could be unblinded in the warfarin arm, whereas ≤ 10% of people could be unblinded in the aspirin arm. In both the aspirin vs warfarin trials, Bang's BI uniformly showed increased unblinding for warfarin over aspirin. Summarizing a pattern can be important for different studies in consistency checking and meta-analytic perspectives. The authors explained that the observed trend may be explained in other occurrences associated with administration of warfarin (e.g. number of dose change, hemorrhage).

Other issues

Who should be blinded?

Some researchers argued that imperfect blinding or half blinding is preferable to an open design (2).

Table 4. Blinding data and assessment: acupuncture study.

	Guess, *n*			
Assignment	Acupuncture	Sham	DK	Total
Acupuncture	21 (47%)	0 (0%)	24 (53%)	45
Sham	19 (39%)	4 (8%)	26 (53%)	49
Total	40	4	50	94

James' BI = 0.73 [95% CI: 0.66–0.80]. Bang's BI = 0.47 [0.33–0.61] in acupuncture and −0.31 [−0.49 to −0.13] in sham.

Table 5. Blinding data and assessment: WASID study.*

	Guess, *n*			
Assignment	Aspirin	Warfarin	DK	Total
Patient/proxy[a]				
Aspirin	41 (18%)	55 (24%)	130 (58%)	226
Warfarin	14 (6%)	99 (45%)	106	219
Total	55	154	236	445
Study coordinator[b]				
Aspirin	42 (15%)	16 (6%)	213 (79%)	271
Warfarin	23 (8%)	97 (35%)	157 (57%)	277
Total	65	113	370	548
Principal neurologist[c]				
Aspirin	32 (12%)	10 (4%)	229 (85%)	271
Warfarin	18 (7%)	52 (19%)	205 (75%)	275
Total	50	62	434	546

[a] James' BI = 0.69 [95% CI: 0.65–0.73]. Bang's BI = –0.06 [–0.15 to 0.02] in aspirin and 0.39 [0.31–0.47] in warfarin.
[b] James' BI = 0.75 [95% CI: 0.72–0.79]. Bang's BI = 0.10 [0.04–0.15] in aspirin and 0.27 [0.20–0.34] in warfarin.
[c] James' BI = 0.85 [95% CI: 0.82–0.88]. Bang's BI = 0.08 [0.04–0.13] in aspirin and 0.12 [0.07–0.18] in warfarin.
* Missing data were excluded.

Table 6. Blinding data and assessment: SPINAF study.

	Guess			
Assignment	Aspirin	Warfarin	DK	Total
Patient[a]				
Aspirin	42 (16%)	16 (6%)	207 (78%)	265
Warfarin	42 (16%)	74 (28%)	144 (55%)	260
Total	84	90	351	525
Study coordinator[b]				
Aspirin	26 (10%)	10 (4%)	229 (86%)	265
Warfarin	10 (4%)	47 (18%)	203 (78%)	260
Total	36	57	432	525

[a] James' BI = 0.78 [95% CI: 0.76–0.80]. Bang's BI = 0.10 [0.04–0.15] in aspirin and 0.12 [0.04–0.20] in warfarin.
[b] James' BI = 0.86 [95% CI: 0.83–0.89]. Bang's BI = 0.06 [0.02–0.10] in aspirin and 0.14 [0.09–0.20] in warfarin.

Although double- or triple-blind trials may not be perfect, the general consensus is that they are the best methodology we currently have available. In contrast, 'prospective, randomized, open-label, the blinded-endpoint (PROBE) design' was advocated for objective endpoints (e.g. death) as a less precise yet cheaper and simpler alternative, especially for a complementary purpose (28–31).

How to decide between the two BIs?

It is clear that the two BIs may not be compared directly because they are based on different paradigms and assumptions. James' method (20) believed that the most important observations are 'DK', whereas the Bang's index does not rely as much on these observations, but places more focus on the balance in proportions of correct vs incorrect guesses. Recognizing this fact, our suggestion in practice is that the two BIs are *complementary* and that it may be reasonable to compute both BIs for each trial and present these results, especially when the two methods yield different conclusions. One rationale is that James' BI tends to be more conservative and may provide a summary measure of totality emphasizing DK and disagreement. Yet, even when DKs are truly DKs so that James' BI is well justified, we may still want to know the degree of unblinding in each arm and any imbalance between two arms that Bang's BI can offer. If one combines the BIs from different arms, as in James' method, some cancel-out effect can occur and that could lead us to an erroneous conclusion. Moreover, Bang's BI may enable us to classify a given scenario based on the 'classification rule of nine blinding scenarios' (see Table 7) (13, 21, 22). For example, the acupuncture study in the Example section may be classified as the (Experimental, Control) = (Unblinded, Opposite guess) scenario and it may mirror meaningful underlying scenario(s). Park et al. (32) computed frequencies of these scenarios for 63 trials that they found through systematic review and provided compelling evidence that unblinding is common.

When to perform the blinding survey?

Another important and controversial question upon which there is no real consensus is 'when to ask blinding questions?' Options can be 'before, shortly after randomization, during trial, and/or at study completion (i.e. after trial or after treatment administration, whichever more sensible)'.

Some researchers assert that assessment may be inappropriate after the trial due to confounding between efficacy/side-effects and correct guessing (33–35). Also, it is recognized that the timing of patient guess about allocation of treatment can be an important factor that must be explored. When the timing of patient guess is at a single time point (early, during, after) or at multiple time points, the results may differ. Longitudinal measurements would provide insight into the effect of time and study influences on the assessment outcomes, and indeed some studies have shown that the longer someone is on a treatment, the more frequently they are asked about their treatment belief, and the more likely they are to guess their treatment assignment (36, 37). We, however, generally advocate asking at the end of treatment or trial by arguing.

Table 7. Nine blinding scenarios.

Experimental arm	Control arm	Possible interpretations (on blinding and treatment effectiveness)
Random guess	Random guess	Ideal
Random guess	Opposite guess	(Psychologically/behaviorally) Rare
Random guess	Unblinded	Relatively rare—possibly, little treatment effect and completely no effect in control arm (e.g. no placebo effect)
Unblinded	Unblinded	Could be problematic. Possibly, clear treatment effect in experimental arm and no treatment effect in control arm (e.g. patients tend to know what to expect)
Unblinded	Opposite guess	Ideal (e.g. patients tend to have wishful thinking or patients don't know how control treatment looks)
Unblinded	Random guess	Could be problematic. Possibly, clear treatment effect in experimental arm and no treatment effect in control arm (e.g. patients do not know what to expect in the absence of treatment)
Opposite guess	Opposite guess	Rare or unlikely
Opposite guess	Random guess	Rare
Opposite guess	Unblinded	Possibly, no treatment effect at all or patients tend to be negative or unmotivated.

Random guess/unblinded/opposite guess may be classified based on statistical test or threshold set by investigators a priori (say, BI > 30%).

Statistically speaking, of course, the best approach is to ask twice or more because more data carry more information and longitudinal assessments can capture blinding patterns including belief change over time. However, we still prefer 'at the end' (in keeping with the FDA's recommendations). Although we may not be able to verify if unblinding is true unblinding or DKs are true DKs, we never want to make participants try to guess. As we ask more, they may become curious so that they may make efforts to break the blind, and this may cause behavioral change and some biases. Blinding may convey stories during the entire course of the trial including expectation, efficacy, side-effects, inadvertent disclosures or even lies, and the cause and effect relationship between blinding and treatment effect could be bi-directional. If you want to test the blinding at the beginning, it may be more relevant to allocation concealment (38) and more reasonable to do with the third party or in a pilot study. Moreover, when we say a person was blinded in a clinical trial, it generally means 'through the trial', not just at randomization.

An alternative strategy might be to ask each subject the blinding question *only once* but at different time points (say, randomly selected 25% before randomization, 25% shortly after randomization, 25% during, 25% at the end). Also, we may contact the participants who dropped out and assess their beliefs regarding blinding because some of them might have dropped out due to the fact that they were unblinded. When this is the case and blinding assessment occurs at the end of a trial, we may experience biased sample, yet we may also obtain an understanding of treatment effects through including these trial drop-outs.

Implications on trials of non-pharmacologic interventions

It is much more difficult to implement blinding techniques and to maintain blinding in trials assessing non-pharmacologic interventions such as device, surgery, psychosocial therapy, complementary and alternative medicine, among others. These studies could be fundamentally different from classical placebo-controlled trials and should frequently rely on non-standard, often creative, methods for blinding and allocation concealment. For instance, sham procedures, masking tools for patients (e.g. eye patches, curtain, specific positioning), blinding to the study hypothesis, blinded centralized assessment of primary outcome, and limiting communications among patients, physicians, and/or outcome assessors, have been employed; see Boutron et al. (39) for a detailed description of methods that could overcome some barriers of blinding in this field. The CONSORT statement has also been extended to randomized trials of non-pharmacologic interventions and discussed blinding-related issues and reporting standards in depth (40).

Concluding thoughts

Quantitative assessment of blinding can be quite straightforward statistically, but the final conclusion unavoidably relies on the subjectivity of the data, the interpretation, and the nature of the study. For example, subjectivity is exemplified in the question, *how much unblinding beyond chance is serious enough to be problematic 20%, 50% ...?* More importantly, the determination of the success of blinding could prove much more difficult than to be simplified into one test or measure, therefore a statistical method based on *observed data* may only be a part of the comprehensive evaluation that is truly needed but may not always be possible with data we normally collect.

We noted that not all correct guesses are problematic. Under correct or incorrect guesses, there can be different underlying stories or mechanisms. Moreover, the importance of blinding could depend on the medical condition and context. Clearly, blinding could be a more serious issue to studies with soft/subjective endpoints than those with hard/objective endpoints (e.g. mortality). Empirical evidence demonstrates that inadequate and unclear allocation concealment or blinding is associated with biased estimates of treatment effects (38, 40), whereas some meta-analyses reported that blinding was not associated with a treatment effect on hard endpoints (41, 42). Interestingly, Google Scholar (http://scholar.google.com/) identified more articles under 'single blind' than 'double blind' (1850 K vs 1500 K). In contrast, PubMed (www.ncbi.nlm.nih.gov/pubmed/) identified a significantly smaller number of articles under 'single blind' than 'double blind' (30 K vs 126 K). This may provide strong evidence that publication bias regarding the 'blinding' status is apparent—it is likely that some, if not many, editors/reviewers favor double-blind trials and authors add 'double-blind' in the title and/or keywords as a convention or a way to increase the success rate of their publications. Therefore, current meta-analyses based on this possibly inaccurate term may be limited.

For rigorous evaluations, we would need 'much more' additional data, possibly at multiple time points within a trial. However, as mentioned previously, this additional data may come at a greater cost than it is worth, because blinding is an important issue but, realistically speaking, may not be the most important issue in the conduct of a RCT and asking more can cause additional biases and behavioral changes. We may not know yet what we really need nor how to optimally collect data in the presence of confounding, psychological factors, and dishonesty. As such, blinding research may be destined to be *subjective*, *qualitative*, and *imperfect*. However, empirical (quantitative) evidence is almost always superior to ignorance in any decision-making process and understanding/characterizing scientific phenomena. Thus far, minimal attention has been paid to this important component in RCTs.

Perhaps the most essential question related to the maintenance of blinding could be: What is the extent of the impact that unblinding has on the primary analysis of treatment effects? Until much more research is done and much more data are collected, this question is extremely difficult to answer in a definitive manner. Perhaps the most realistic strategy may be to consider blinding assessments as a potentially useful 'exploratory' analysis that may play an important role in providing lessons to be learned for planning and conducting future trials. Also, if more researchers collect data on blinding and report this data in their publications, we would be able to synthesize the cumulative evidence and to know what kind of treatments are difficult to blind, what are common causes of unblinding, and to possibly understand its impacts on the primary analyses as a way to estimate the true effect of blinding. With currently available data, comprehensive, unbiased and robust literature synthesis and meta-analysis seem to be unachievable. Although the blinding protocol that we propose in this paper has not been validated yet, it may serve as a first step for promoting the collection of more blinding data in a standardized format and researchers and trialists may adapt or improve it further.

If unblinding occurs, it is crucial for investigators to find possible explanation(s) and identify the causes and fix any problems, whenever sensible and possible. Beyond investigators, new treatment/drug developers should take blinding issues more seriously. Our personal assessment regarding a good treatment/RCT is: 'Treatment effect' should be greater than 'Non-compliance effect', which should be greater than 'Unblinding effect'. Ideally, when one develops a new treatment, the research goal should be not only to trump the control treatment (e.g. placebo effect) but also to surpass non-compliance and unblinding effects.

In conclusion, we strongly recommend that more investigators collect blinding data if they wish to call their trial 'blinded', which otherwise can be a chronically abused and misused term, particularly 'double-blind'. Indeed, since the use of CONSORT has been shown to improve the quality of RCT reports (e.g. Moher et al. (8) found that the reporting of allocation concealment has been improved by 22% after CONSORT's publication), we would like to see the CONSORT group, funding agencies, FDA, journal editors, data monitoring committees, and protocol reviewers *require*, rather than 'softly recommend', some (preferably, numerical) evidence of blinding in the publication of all blinded RCTs. We are fully aware of the reality that blinding may not be always achievable throughout the trial, but investigators should do their best to protect their trials from bias and communicate their efforts to the reader. Ideally, readers should not have to assume or guess the methods used or if any was used (38).

Acknowledgements

We thank reviewers for constructive and thoughtful comments that improved this manuscript greatly.

Declaration of interest

Heejung Bang was partially supported by CTSC grant at Weill Cornell Medical College (UL1-RR024996).

References

1. Henkin RI, Schecter PJ, Friedewald WT, Demets DL, Raff M. A double blind study of the effects of zinc sulfate on taste and smell dysfunction. *American Journal of the Med Sciences* 1976;272:285–299.
2. Furberg CD, Soliman EZ. Double-blindness protects scientific validity. *J Thromb Haemostasis* 2008;6:230–231.
3. Bausell RB, Lao L, Bergman S, Lee WL, Berman BM. Is acupuncture analgesia an expectancy effect? Preliminary evidence based on participants' perceived assignments in two placebo-controlled trials. *Evaluation & the Health Professions* 2005;28:9–26.
4. Shapiro S. Risks of estrogen plus progestin therapy: a sensitivity analysis of findings in the Women's Health Initiative randomized controlled trial. *Climacteric* 2003;6: 302–310.
5. Shapiro S. Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? *Pharmacoepidemiol Drug Safety* 2004;13:257–265.
6. Garbe E, Suissa S.Issues to debate on the Women's Health Initiative (WHI) study. Hormone replacement therapy and acute coronary outcomes: methodological issues between randomized and observationals studies. *Human Reprod* 2004;19:8–13.
7. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz K, David S, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association* 1996;276:637–639.
8. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *JAMA* 2001;285:1987–1991.
9. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *Med Educ* 1999;33:66–78.
10. Boutron I, Estellat C, Ravaud P. A review of blinding in randomized controlled trials found results inconsistent and questionable. *J Clin Epidemiol* 2005;58:1220–1226.
11. Fergusson D, Glass KC, Waring D, Shapiro S. Turning a blind eye: The success of blinding reported in a random sample of randomised, placebo controlled trials. *British Medical Journal* 2004;328:432.
12. Hrobjartsson A, Forfang E, Haahr MT, Als-Nielsen B, Brorson S. Blinded trials taken to the test: An analysis of randomized clinical trials that report tests for the success of blinding. *International Journal of Epidemiology* 2007;36:654–63.
13. Kolohi J, Bang H, Park J. Towards a proposal for assessment of blinding success in clinical trials: up-to-date review. *Community Dental Oral Epidemiol* 2009;37:477–484.
14. Walter SD, Awasthi S, Jeyaseelan L. Pre-trial evaluation of the potential for unblinding in drug trials: a prototype example. *Contemp Clin Trials* 2005;26:459–468.
15. Bowling A. Mode of questionnaire administration can have serious effects on data quality. *J Public Health* 2007;27:281–291.
16. Hughes JR, Krahn D. Blindness and the validity of the double-blind procedure. *J Clin Psychopharmacol* 1985;5:138–142.
17. Margraf J, Ehlers A, Roth WT. How 'blind' are double-blind studies? *J Consult Clin Psychol* 1991;59:184–187.
18. Kolahi J, Soolari A, Ghalayani P, Varshosaz J, Fazilaty M. Newly formulated chlorhexidine gluconate chewing gum that gives both anti-plaque effectiveness and an acceptable taste: A double blind, randomized, placebo-controlled trial. *Int Acad Periodontol* 2008;10:38–44.
19. Wisner KL, Perel JM, Peindl KS, Hanusa BH, Findling RL, Rapport D. Prevention of recurrent postpartum depression: a randomised clinical trial. *Journal of Clinical Psychiatry* 2001;62:82–86.
20. James KE, Bloch DA, K.K. L, Kraemer HC, Fuller RK. An index for assessing blindness in a multi-centre clinical trial: Disulfiram for alcohol cessation - a VA cooperative study. *Statistics in Medicine* 1996;15:1421–1434.
21. Bang H, Ni L, Davis CE. Assessment of blinding in clinical trials. *Contr Clin Trials* 2004;25:143–156.
22. Park J, White AR, James MA, Hemsley AG, Johnson P, Chambers J. Acupuncture for subacute stroke rehabilitation: A sham-controlled, subject- and assessor-blind, randomized trial. *Archives of Internal Medicine* 2005;165:2026–2031.
23. Rimm AA, Bortin M. Clinical trials as a religion. *Biomedicine* 1978;28:60–63.
24. Bang H, Ni L, Davis CE. Response to Henneicke-von-Zepelin and Hemilä's comment. *Contr Clin Trials* 2005;26: 514–515.
25. Hertzberg V, Chimowitz M, Lynn M, Chester C, Asbury W, Cotsonis G. Use of dose modification schedules is effective for blinding trials of warfarin: Evidence from the WASID study. *Clinical Trials* 2008;5:23–30.
26. Chimowitz MI, Lynn MJ, Howlett-Smith H, Stern BJ, Hertzberg VS, Frankel MR, Levine SR, Chaturvedi S, Kasner SE, Benesch CG, Sila CA, Jovin TG, Romano JG. Comparison of warfarin and aspirin for symptomatic intracranial arterial stenosis. *New England Journal of Medicine* 2005;352:1305–1316.
27. Ezekowitz MD, Bridgers SL, James KE, Carliner CL, Gornick CC, Krause-Steinrauf H, Kurtzke JF, Nazarian SM, Radford MJ. Warfarin in the prevention of stroke associated with non-rheumatic atrial fibrillation. *New England Journal of Medicine* 1992;327:1406–1412.
28. Hansson L, Hedner T, Dahlöf B.Prospective randomized open blinded end-point (PROBE) study. A novel design for intervention trials. *Blood Pressure* 1992;1:113–119.
29. Smith DH, Neutel JM, Lacourcière Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint (PROBE) designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM measurements. *Journal of Hypertension* 2003;21:1291–1298.
30. Büller HR, Halperin JL, Bounameaux H, Prins M. Double-blind studies are not always optimum for evaluation of a novel therapy: the case of new anticoagulants. *Journal of Thrombosis and Haemostasis* 2008;6:227–229.
31. Casteels M, Flamion B. Open-label trials and drug registration: a European regulator's view. *J Thromb Haemostasis* 2008;6:232–234.
32. Park J, Bang H, Canette I. Blinding in controlled trials, time to do it better. *Complement Ther Med* 2008;16:121–123.

33. Sackett DL. Turning a blind eye: why we don't test for blindness at the end of our trials. *BMJ* 2004;328: 1136.
34. Henneicke-von Zepelin HH. Letter to the editor. *Contemp Clin Trials* 2005;26:512.
35. Hemilä H. Assessment of blinding may be inappropriate after the trial. *Contemp Clin Trials* 2005;26:512–514.
36. Lao L, Bergman S, Hamilton GR, Langenburg P, Berman B, Evaluation of acupuncture for pain control after oral surgery: A placebo-controlled trial. *Archives of Otolaryngology Head and Neck Surgery* 1999;125:567–572.
37. Smith MJ, Tong HC. Manual acupuncture for analgesia during electromyography: a pilot study. *Arch Phys Med Rehab* 2005;86:1741–1744.
38. Schulz K. Assessing allocation concealment and blinding in randomised controlled trials: why bother? *Evidence-Based Med* 2000;5:36–37.
39. Boutron I, Guittet L, Estellat C, Moher D, Hrobjartsson A, Ravaud P. Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. *PLoS Med* 2007;4:e61.
40. Boutron I, Moher D, Altman D, Schulz K, Ravaud P. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Annals of Internal Medicine* 2008;148:295–309.
41. Balk E, Bonis PAL, Moskowitz H, Schmid C, Ioannidis JPA, Wang C, Lau J. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973–2982.
42. Wood L, Egger M, Gluud L, Schulz K, Jni P, Altman D, Gluud C, Martin R, Wood AJG, Sterne JAC. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ. British medical journal (Clinical research ed.)* 2008;336:601–605.
43. Haynes RB. Incorporating allocation concealment and blinding in randomised controlled trials. *Evidence-Based Med* 2000;5:38.

informa
healthcare

2010 Informa UK Ltd.
doi:10.3109/10601331003777444
Online supplementary material published in conjunction with Heejung Bang et al., Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol. Clinical Research and Regulatory Affairs 2010; 27(2): 42–51

Supplementary material

Appendix: A sample blinding survey and assessment protocol

We recommend that the investigators mention the following points (all or relevant points) in the protocol as well as in the design and main papers. (Supplemental Documents may be used):

1. Who were blinded (the standard terms such as 'open or unblinded', 'single', 'double' may be used. We generally recommend that authors specifically state who were actually blinded such as patients, physicians, outcome assessors, and/or statisticians.)
2. How blinding was administered and what kind of efforts were made to maintain blinding. [Note that even in open trials, some parties (e.g. outcome assessors, investigators who do not treat patients, statisticians and/or data monitoring committee) could be blinded.]

At a minimum, all trials need to report #1 and 2.
If investigators decide to assess blinding,

3. When blinding data were collected (e.g. at the beginning, during, or at the end). If one time is to be chosen, we generally recommend the close-out visit (at the end of the trial or of treatment administration, whichever is more relevant; the latter is sensible when the endpoint is mortality or drop-out rate is high). If investigators want to collect the data more than once, it is certainly valid but some caution may be necessary (see the main text for explanations).
4. The following blinding questionnaire can be surveyed with proper adaptation as necessary:

   - Q1. *Which treatment do you believe you received?*

   a) New treatment
   b) Placebo
   c) Don't know

   Or

   - Q1. *Which treatment do you believe you received?*

   a) Strongly believe new treatment
   b) Somewhat believe new treatment
   c) Somewhat believe placebo
   d) Strongly believe placebo
   e) Don't know

   - Q2. *If you answer 'Don't know' above, are you willing to provide your best (or random) guess of a treatment you received anyway?*

   a) New treatment
   b) Placebo

In the survey questionnaire, we generally recommend that you ask some more qualitative questions together. For example,

- Q3. *Why do you believe you received this treatment?*
- Q4. *If you find out, are you willing to tell when and/or how?*

Other general questions (e.g. participants' satisfaction, problems, or other comments, or suggestions for the trial) can be queried together at the study close-out, preferably in the same questionnaire. [Rationale: we may not want to highlight blinding questions here. We do not want to make participants try to guess even in their future trials and give an impression that not knowing the treatment they received is wrong or bad.] We should encourage all the participants to provide their honest guesses and opinions.

5. Report the blinding data (see Tables 1–3 for sample tables in the main text).
6. Compute two BIs (i.e. James et al. and Bang et al.'s methods) and their 95% CIs and report them*. [Statistical testing may be conducted here. Yet, we emphasize estimation rather than testing in blinding assessment.]

You may classify your trial according to the nine blinding scenarios (as in Table 7).

7. Interpret the results from #6 and try to understand the underlying scenario. Also, report important problems identified, if any, and lessons learned

that may be useful for future trials or other investigators (notable findings may be reported in publications).

Other considerations:

1. If investigators are interested in the distinguishability of treatments, it may be tested prior to the trial in a (small) independent group of people who are not a part of the actual trial or in a pilot study.
2. Authors may describe allocation concealment and blinding separately—see the glossary in (43) for definitions.
3. If of drop-out/censoring or non-compliance are examined/recorded, unblinding should be considered as a possible cause.
4. Sub-group analyses based on the blinding status may be conducted.

* *Note*: A STATA module is available at http://ideas.repec.org/c/boc/bocode/s456898.html and a *R* function is available at http://biostat.mc.vanderbilt.edu/wiki/pub/Main/NM_R_FUNCTIONS/bi2.r.