2005

Changes in Beliefs Identify Unblinding in Randomized Controlled Trials: A Method to Meet CONSORT Guidelines

Judy R. Rees

Timothy J. Wade

Deborah A. Levy

John M. Colford

Joan F. Hilton

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Contemporary Clinical Trials 26 (2005) 25–37
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

Changes in beliefs identify unblinding in randomized controlled trials: a method to meet CONSORT guidelines

Judy R. Rees[a,*], Timothy J. Wade[b], Deborah A. Levy[c], John M. Colford Jr.[d], Joan F. Hilton[e]

[a]*Department of Community and Family Medicine, Center for Environmental Health Sciences, and Norris Cotton Cancer Center; Dartmouth Medical School, Lebanon, NH, United States*
[b]*United States Environmental Protection Agency, Epidemiology and Biomarkers Branch, Chapel Hill, NC, United States*
[c]*Centers for Disease Control and Prevention, Atlanta, GA, United States*
[d]*University of California Berkeley, School of Public Health, Berkeley, CA, United States*
[e]*University of California San Francisco, School of Medicine, San Francisco, CA, United States*

Received 7 April 2004; accepted 5 November 2004

Abstract

Double-blinded trials are often considered the gold standard for research, but significant bias may result from unblinding of participants and investigators. Although the CONSORT guidelines discuss the importance of reporting "evidence that blinding was successful", it is unclear what constitutes appropriate evidence. Among studies reporting methods to evaluate blinding effectiveness, many have compared groups with respect to the proportions correctly identifying their intervention at the end of the trial. Instead, we reasoned that participants' beliefs, and not their correctness, are more directly associated with potential bias, especially in relation to self-reported health outcomes.

During the Water Evaluation Trial performed in northern California in 1999, we investigated blinding effectiveness by sequential interrogation of participants about their "blinded" intervention assignment (active or placebo). Irrespective of group, participants showed a strong tendency to believe they had been assigned to the active intervention; this translated into a statistically significant intergroup difference in the correctness of participants' beliefs, even at the start of the trial before unblinding had a chance to occur. In addition, many participants (31%) changed their belief during the trial, suggesting that assessment of belief at a single time

* Corresponding author. Dartmouth-Hitchcock Medical Center, One Medical Center Drive, 7927 Rubin Building, Lebanon, NH 03756, United States.
E-mail address: judith.rees@dartmouth.edu (J.R. Rees).

1551-7144/$ - see front matter
doi:10.1016/j.cct.2004.11.020

does not capture unblinding. Sequential measures based on either two or all eight questionnaires identified significant group-related differences in belief patterns that were not identified by the single, cross-sectional measure.

In view of the relative insensitivity of cross-sectional measures, the minimal additional information in more than two assessments of beliefs and the risk of modifying participants' beliefs by repeated questioning, we conclude that the optimal means of assessing unblinding is an intergroup comparison of the change in beliefs (and not their correctness) between the start and end of a randomized controlled trial.

Keywords: Masking; Randomized controlled trial; Double-blind method; Research design; Placebo effect

1. Introduction and background

The randomized, placebo-controlled trial is often described as the gold standard for research involving human participants. An important aim of the study design is to allow comparisons between groups of participants whose characteristics are, on average, as similar as possible with the exception of the intervention being studied. This is achieved primarily through randomization, which aims to prevent selection bias and balance the distribution of measured and unmeasured confounding variables between the active and control groups. Even if randomization fails to distribute measured confounders evenly between groups, their effects can be adjusted for in the analysis. The benefits of randomization are further increased by blinding, a method to hide the true nature of the intervention assigned to each participant from participants and investigators and hence prevent the exposure–outcome relation under study from being influenced by knowledge or belief about the intervention.

It has been estimated that trials that do not attempt to use double-blinding exaggerate treatment effects by 14% compared with trials that do attempt to double-blind [16]. Because the latter group includes well-blinded and poorly blinded trials, it is likely that a comparison of treatment effects in successfully blinded versus unblinded trials would show an even larger bias than 14%. In recognition of the importance of effective blinding in such trials, the Consolidated Standards of Reporting Trials (CONSORT) statement describes many aspects of blinding methodology that may be included in published reports of randomized controlled trials; these include evidence for successful blinding among participants, those administering the intervention, outcome assessors and data analysts [1,2]. However, it is unclear what constitutes adequate evidence of blinding effectiveness. A recent study found that only 8% (15/191) of published randomized placebo-controlled trials reported any assessment of blinding [3].

Participants or researchers may become aware of the intervention assignment before allocation (failure of allocation concealment), or after allocation (unblinding), and the nature and magnitude of the resulting biases are potentially different [4]. Allocation concealment has been described as a means of preventing selection bias caused by differences in enrollment or early withdrawals of participants from the study, whereas blinding aims primarily to prevent ascertainment bias and attrition [4].

Participants' beliefs about the intervention to which they have been assigned may affect their experience or reporting of symptoms through a variety of mechanisms that probably differ from study to study and may be related to the placebo effect. This has been described as a mixture of factors including classical conditioning effects, spontaneous improvement in clinical course and the tendency for

participants to give polite or "expected" answers [5–7]. The magnitude of the placebo effect may be related to the participant's perceived likelihood of receiving an effective medication [8]. Beliefs about which intervention has been assigned can affect outcome reporting by participants and outcome assessment by researchers, and also lead to differences in adherence to the study protocol. Adherence can itself predict outcome within either the active or placebo arms [9]. Interactions between unblinding and patient preferences for a particular treatment may also influence outcome [10].

Blinding is not always possible to achieve and can be compromised in various ways. A key concern is the interplay between noticeable physical characteristics of the intervention (e.g., the smell, taste or texture of a pill) and participants' expectations, whether the latter are based on knowledge or assumption [11,12]. Side effects and/or the beneficial effects of a drug may lead to unblinding when they reach a discernible threshold [11]. It has been argued that "the more potent the therapeutic variable, the less likely its efficiency can be proven in a double blind study" [12]. This effect, termed Philip's paradox, describes the introduction of bias as a direct result of unblinding caused by the efficacy of the intervention; its implication is that unblinding is less likely if the study end-point is measurable but cannot be detected by the patient (e.g. a laboratory test) [11]. Unblinding may also result from participants' attempts to identify their intervention [13–15] or flawed protocol design or execution, which provide cues to participants [11].

Few studies have focused on strategies to evaluate blinding in detail. Most were based primarily on the proportions of participants guessing their group assignment correctly at the end of the study, analyzed either within treatment groups or overall [1,13,14,17–23]. Although blinded participants might be expected to have a 50% chance of guessing that they had received the active or placebo intervention, it has been argued that this can be expected only under exceptional circumstances [11]. Some investigators have supplemented primary guesses with "forced" guesses from participants who initially responded that they did not know (DK), or with measures of certainty [13,14,22,24]. Hughes and Krahn described a series of procedures to assess blinding, beginning with a χ^2 test comparing the proportions of correct and incorrect answers in each group, and including analyses stratified by the correctness of participants' guesses [16]. Howard et al. [14], James et al. [19] and Bang et al. [23] have described summary indices of blinding based on the proportions guessing correctly, incorrectly and DK (Appendix A). Many investigators have relied on a single evaluation of the correctness of beliefs at the end of the trial. Longitudinal approaches, based on sequential interrogation about beliefs during the trial, have been largely overlooked, although the importance of studying the temporal characteristics of unblinding has been raised previously [11].

In this study, we argue that cross-sectional analyses cannot capture unblinding and that, rather, one must account for participants' initial beliefs as well. We then test the hypothesis that unblinding can be captured adequately by a measure based on group-specific initial and final beliefs about group assignment (i.e., at weeks 2 and 16) against the null hypothesis that more interrogations are needed (i.e., at least six of eight possible biweekly responses).

2. Methods

We previously investigated participant blinding in a community-based pilot study in northern California to investigate the effect on gastrointestinal infectious illness of water treatment units installed at the kitchen sink [25,26]. To promote successful blinding, the active and placebo units were

designed to be externally identical and to produce water with similar characteristics such as temperature, taste and odor. The active unit consisted of a filter and ultraviolet light system in series; in the placebo unit, the filter casing was empty and a glass sleeve around the light prevented emission of ultraviolet wavelength [25]. The water treatment units, the primary evaluation of blinding, James' index [19], and the primary health outcome, "highly credible gastrointestinal illness" (HCGI) [27], were described in detail previously [25]. The sample size for the original study was based on our ability to identify a blinding index (BI) exceeding 0.5 with a type I error rate of 0.05 and type II error rate of 0.10 [25]. On eight occasions, we asked all participants aged ≥12 years to guess their group assignment in a self-administered questionnaire that they returned by mail. Based on Byington's approach [13], we asked them to choose one of five responses: "definitely the active water treatment device", "probably the active water treatment device", "probably not the active water treatment device", "definitely not the active water treatment device" or "I'm not sure". We asked participants who responded "I'm not sure" to make a guess, and this "forced guess" could be either "probably the active water treatment device" or "probably not the active water treatment device". For the analyses presented here, we combined "definitely" and "probably" responses to a single category. Participants were first questioned 2 weeks after the study began and every 2 weeks until 16 weeks were completed. They were also asked to write the reasons for their beliefs and answer questions about their health.

We examined several cross-sectional approaches to evaluate blinding effectiveness. Methods based on the *correctness* of answers included James' index [19], Howard's index [14] and a χ^2 test of the correctness of respondents' beliefs by intervention group. Methods based on *beliefs* about group assignment included a χ^2 test of respondents' beliefs (i.e., active, placebo, DK) by intervention group, and an analysis of the association between exposure and outcome stratified by belief, which was reported previously [25].

We devised two longitudinal measures to describe patterns of belief throughout the trial. For both measures, we replaced DK responses with the corresponding forced guesses. A response was defined as valid if the participant chose one of the initial five options described above, even if they refused to provide a forced guess. (1) Our two-point measure consisted of beliefs at weeks 2 and 16 (i.e., active–active, placebo–placebo, active–placebo, placebo–active); we compared these between groups using Fisher's exact test. (2) Our six-point measure used all valid responses given during the study for participants responding on at least six occasions, including weeks 2 and 16. We defined participants' beliefs as consistent (active or placebo) if all responses were identical, allowing up to one opposite belief during weeks 4 through 14 (e.g. if 0 is "believe active", 1 is "believe placebo" then 11111111 and 11011111 are consistent, but 10000000 is not). Participants were defined as switching belief (to placebo or to active) if they changed their belief exactly once during the study (e.g. 11000000, 00001111). This category aimed to identify participants with a clear time at which "unblinding" occurred. The remaining participants were classified as undecided (e.g. 01111110, 10111101). We compared these five categories between groups using Fisher's exact test. Note that the two-point and six-point measures are not independent: e.g. participants whose six-point measure is classified as consistently active must, by definition, have responded active–active at weeks 2 and 16, although some participants responding active–active at weeks 2 and 16 may have varied responses in the intervening period. The sensitivity of these results to the number of queries about beliefs was investigated by comparing the two sequential measures within the subset of participants who gave valid responses at weeks 2 and 16.

We compared measures of belief using likelihood ratio tests to identify differences between nested models. In addition, we investigated whether DK was synonymous with "blinding" and determined the reasons for participants' beliefs.

3. Results

Water treatment units were installed in 80 households, 3 of which were subsequently excluded from the study [25]; the remaining 77 households comprised 236 individuals. Participants aged ≥12 years (N=179) were asked to report their beliefs on 8 occasions and did so on average 6.6 times (median 8, interquartile range 6–8). Valid responses were provided at week 2 by 172 (96%) respondents, at week 16 by 145 (81%) and on all 8 occasions by 123 (69%); 132 (74%) reported their beliefs on at least 6 occasions including weeks 2 and 16.

3.1. Analyses of correctness of beliefs

At week 16, correctness of belief was strongly associated with intervention group; for example, after redistributing DK responses using forced guesses, 72% of active group participants and 40% of placebo participants guessed their assignments correctly (mean difference, 33%; 95% confidence interval [CI]=17–49%). Howard's "proportion who really knew" at week 16 was 0.09 (CI=0.04–0.18). James' BI was 0.64 (CI=0.57–0.72). Successful blinding is presumed for Howard's index if the confidence interval includes 0 and for James' index if it excludes 0.5 [14,19].

3.2. Analyses of beliefs

3.2.1. Beliefs at weeks 2 and 16 (cross-sectional analyses)

A majority of participants in both groups believed or guessed they were in the active group at the start of the trial (week 2, 76%; Fig. 1a) and also at the end of the trial (week 16, 67%; Fig. 1b). The distribution of beliefs did not differ significantly between groups at either time point (Fisher's exact tests, P=0.605 and P=0.139, respectively).

3.2.1.1. Certainty of beliefs and the "don't know" response. On the basis of beliefs ranked by categories of certainty, there was no statistically significant difference in certainty of beliefs between groups at week 16 (Wilcoxon rank-sum test, P=0.154). Participants who were unsure of their group assignment constituted 50% of respondents at week 2 and 33% at week 16. At week 2, we saw an increased tendency for forced guesses to be "active" among those who initially said they didn't know (72%, 57/79) as well as among those who initially expressed a belief (78%, 67/86) (P=0.393). In contrast, at week 16, the probability of forced guesses being "active" among those who initially said that they didn't know (49%, 21/43) was lower than it was among those who initially expressed a belief (75%, 73/97) (P=0.002).

3.2.1.2. Reasons for participants' beliefs. Of 179 participants, 171 (96%) stated ≥1 reason during the study for their beliefs about the intervention. Intervention-related factors (including taste, temperature, odor and appearance of the water, and physical characteristics of the filter) accounted

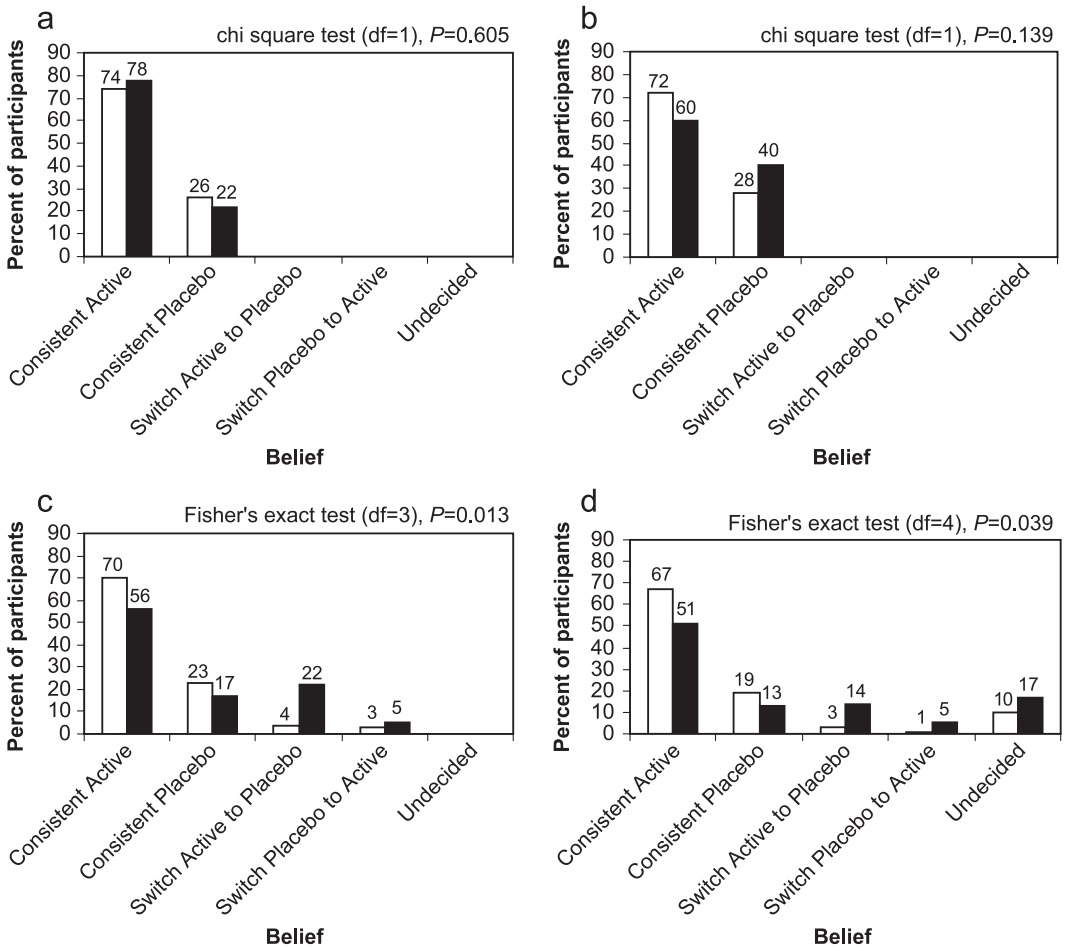

Fig. 1. The classification of belief patterns by cross-sectional and sequential measures of belief among 132 participants who reported beliefs at least 6 times including weeks 2 and 16 (N=132). (a) beliefs at week 2; (b) beliefs at week 16; (c) change in beliefs between two time points, weeks 2 and 16 ("2-point measure"); (d) change in beliefs over at least 6 of 8 time points, weeks 2 through 16 ("6-point measure"). (■) Placebo group; (□) active group; (a) consistent active=active at week 2; switch active to placebo, switch placebo to active, undecided=not applicable; (b) consistent active=active at week 16; switch active to placebo, switch placebo to active, undecided=not applicable; (c) consistent active=active at weeks 2 and 16; undecided=not applicable; (d) consistent active=active at weeks 2 and 16 and at all or all but one time points in between; switch active to placebo=active at week 2, placebo at week 16 and only one change of belief in between; undecided=none of the above.

for 77% (647/836) of reasons. Outcome-related factors (i.e., participants' health) accounted for only 6% (52/836) of reasons given for participants' beliefs about the intervention. After the study, participants in 6% (5/77) of families admitted trying to unblind themselves, either by trying to open the filter, testing the water or by other means.

3.2.2. Change in beliefs between weeks 2 and 16 (two-point analyses)

During the trial, 110/132 (83%) maintained their initial belief and 22/132 (17%) switched their belief. The trends in participants' beliefs during the study differed significantly by intervention group (Fisher's

exact test, P=0.013) (Fig. 1c). For example, a disproportionately high number of participants in the placebo group switched their belief from active to placebo (22%); 4% of participants in the active group switched their belief from active to placebo.

To examine further the relationship between beliefs and gastrointestinal illness, we stratified the intervention effect by change in beliefs (Table 2). Unexpectedly, the rate of illness was relatively low and the intervention effect was qualitatively different among those who consistently believed they were in the placebo group.

3.2.3. Sequential interrogation throughout the study (six-point analysis)

A large majority of participants (69%, 91/132) reported the same beliefs at all time points throughout the study. The proportion fulfilling our definition of consistent beliefs in the six-point analysis (75%, 99/132) was lower than the corresponding proportion (active–active or placebo–placebo) in the two-point analysis (83%, 110/132). Similarly the "switching" proportion according to the six-point analysis (11%, 15/132) is lower than that of the two-point analysis (17%, 22/132). The six-point measure, which was the only one capable of identifying participants with multiple changes of belief, classified 14% as "undecided". The more detailed categories defined by the six-point measure again provided evidence that belief patterns were distributed differentially between intervention groups (P=0.039, Fig. 1d). Repeating the analyses with six different definitions of consistency, we continued to find evidence of differences between intervention groups, although a seventh measure, in which "consistent" was defined as 100% identical beliefs, was less convincing (χ^2 (df=4), P=0.098) (data not shown).

Using likelihood ratio tests, we compared the group-related differences in beliefs using nested models. The difference between the two-point and six-point models was not significant (P=0.521), suggesting that the six-point model did not provide significant additional information over the two-point model. We estimate that, in our sample of 132, an intergroup difference in the proportions of participants who were undecided could have been identified in terms of a minimum risk ratio of 3.1 with a type I error rate of 0.05 and power of 0.8.

4. Discussion

In most evaluations of blinding, participants are only asked to try to identify their intervention group at the end of the trial. For a simple comparison of beliefs at the end of the trial to identify significant intergroup differences caused by unblinding depends on the assumption that participants are undecided at baseline with respect to beliefs. By actually measuring beliefs at baseline in the sequential analysis, we increased our sensitivity to identify changes in belief indicative of unblinding. Furthermore, by showing that use of the two-point measure gave results similar to the six-point measure, we provided evidence that repeated questioning of participants during this trial was unnecessary. This is important because of concerns that repeated questioning draws attention to the issue and may cause additional unblinding and bias.

It has been argued that "a double blind design can work only if the subject is clearly free from the influence of suggestion resulting from accurate information about his medication" [20]. This statement illustrates a common misconception in the evaluation of blinding; bias does not result simply because some participants can identify their intervention group *correctly*. Bias occurs when intergroup differences in *belief* about group assignment differentially affect the outcome and bias the relative risk

(RR) estimate (Appendix B). The relationship between beliefs and their correctness is such that when the two groups are similar with respect to the correctness of beliefs, the beliefs themselves must differ (Table 1). Our results illustrate that, if correctness of beliefs examined at baseline showed a strong association with group, this could not be attributed to unblinding since participants have no experience of the clinical trial at baseline. Furthermore, since many participants maintain their initial belief, we argue that beliefs, and not their correctness, should be used to evaluate blinding effectiveness.

Is a comparison of belief patterns by group a reliable method to identify bias caused by unblinding? If the beliefs themselves are similar in both groups and these beliefs influence health outcomes to the same extent in each group (i.e. there is no effect modification), one might expect the RR estimate to be unbiased; however, we note that bias can result even under these circumstances. It can be shown that, whereas (i) beliefs about the intervention that lead to non-differential *under*-reporting of the outcome in both intervention groups will not bias RR, (ii) beliefs that lead to non-differential *over*-reporting of the outcome will bias RR towards the null (Appendix B). These situations are equivalent to non-differential misclassification of the outcome with (i) perfect specificity and imperfect sensitivity and (ii) perfect sensitivity and imperfect specificity, respectively; these effects have been studied in detail [28,29]. Thus, bias in the RR may occur even if beliefs are comparable in the intervention groups; in the absence of effect modification, RR will be unbiased or biased towards the null. In the presence of effect modification, bias in either direction may result.

To investigate further the effects of beliefs in our study, we stratified the intervention effect by the beliefs observed at weeks 2 and 16 (active–active, placebo–placebo, active–placebo and placebo–

Table 1
Cross-sectional analysis of beliefs and correctness of beliefs about group assignment at week 16

	Intervention group		Mean difference	Total
	Active (*N*)	Placebo (*N*)	(95% confidence interval)	% (*N*)
Belief				
Active	57% (43)	43% (30)	13% (−3% to 30%)	50% (73)
Placebo	16% (12)	17% (12)	−2% (−14% to 11%)	17% (24)
Don't know	28% (21)	39% (27)	−12% (−27% to 4%)	33% (48)
Forced guess active	14% (11)	14% (10)		14% (21)
Forced guess placebo	11% (8)	20% (14)		15% (22)
Refused	3% (2)	4% (3)		3% (5)
Total	100% (76)	100% (69)		100% (145)
Don't combine forced with initial guesses (active, placebo, DK): χ^2 (*df*=2)=2.86, *P*=0.239				
Do combine forced with initial guesses (active, placebo, refused): Fisher's exact test (*df*=2), *P*=0.238				
Correctness of belief				
Correct guess	57% (43)	17% (12)	39% (24% to 54%)	38% (55)
Incorrect guess	16% (12)	43% (30)	−28% (−42% to −13%)	29% (42)
Don't know	28% (21)	39% (27)	−12% (−27% to 4%)	33% (48)
Forced guess correct	14% (11)	20% (14)		17% (25)
Forced guess incorrect	11% (8)	14% (10)		12% (18)
Refused	3% (2)	4% (3)		3% (5)
Total	100% (76)	100% (69)		100% (145)
Don't combine forced with initial guesses:χ^2 (*df*=2)=25.7, *P*<0.001				
Do combine forced with initial guesses: χ^2 (*df*=2)=16.4, *P*<0.001				

active; Table 2). Unexpectedly, the rate of illness was substantially lower among those who consistently believed they were in the placebo group, suggesting that beliefs may have affected the outcome or vice versa. Stratification also provided some qualitative evidence that beliefs systematically modified the effect of the intervention. This could be formally tested via the three-way interaction between group, outcome and belief strata, given a larger sample size. Nonetheless, in addition to reporting the intervention effect overall, it can be reported separately for those who consistently believed they were in the placebo group and those not holding this belief. In a stratified analysis, the stratum-specific RR should equal the overall RR if there is no confounding; otherwise it may be necessary to report belief-specific treatment effects [30]. Two other important situations can cause relative risks to differ by stratum; (i) effect modification (i.e. the effect of the intervention on outcome is modified according to a participant's belief) or (ii) if belief about group assignment is involved in a causal pathway including both exposure and outcome. For example, the physical characteristics of the intervention led participants to believe they are in the active or placebo group, and the belief decreases or increases the likelihood or severity of illness [31]. A primary motivation for blinding is to prevent this causal pathway. If, due to blinding failure, belief behaves as a time-dependent intermediate variable, then stratification and simple multivariate models should not be used to adjust statistically for the effects of belief, although stratified analyses may be helpful qualitatively in describing the relation between belief and outcome. In the future, it may be possible to adapt other statistical methods [32,33] to adjust for the effects of beliefs, by comparing beliefs and outcomes sequentially during the study.

It has been reported previously that "don't know" may not always represent successful blinding [13,14,21,24] although it is desirable because it suggests a weakly held view. Using "forced guesses", we showed that participants who responded DK at the start of the study held beliefs in similar proportions to those who provided responses without prompting (i.e. the majority believed active). In contrast, "forced guesses" made at the end of the study were more consistent with uninformed random guessing. This indicates that study-related factors may affect participants' willingness to express their opinion, and DK does not necessarily represent successful blinding. These observations justify the use of forced guesses rather than DK responses and suggest that the weighting system underlying James' blinding index is not ideal.

In view of the preceding discussion, we define a successfully blinded trial as one in which participants' or researchers' beliefs about the nature of the intervention assigned to each participant do

Table 2
Stratified analyses: incidence rates (episodes of highly credible gastrointestinal illness (HCGI) per person year at risk) by intervention group stratified by participants' beliefs at weeks 2 and 16 (adapted from Colford et al. [25] and Rees [26])

	Intervention group		Rate (95% CI)	Incidence rate ratio (95% CI)
	Active (95% CI)	Placebo (95% CI)		
Belief sequence				
Active–active	2.8 (1.7–4.5) (n=48)	3.3 (1.9–5.7) (n=35)	3.0 (1.8–5.0) (n=83)	1.2 (0.6–2.4)
Placebo–placebo	0.6 (0.2–1.9) (n=16)	0.3 (0.04–2.2) (n=11)	0.5 (0.1–2.0) (n=27)	0.5 (0.1–4.7)
Active–placebo	2.3 (1.0–5.4) (n=3)	3.6 (1.8–7.0) (n=14)	3.3 (1.7–6.7) (n=17)	1.6 (0.5–4.7)
Placebo–active	0 (n=2)	3.7 (0.5–25.2) (n=3)	2.2 (0.3–15.1) (n=5)	–
All	2.2 (1.4–3.4) (n=69)	2.8 (1.8–4.5) (n=63)	2.5 (1.6–3.9) (n=132)	1.3 (0.7–2.5)

not significantly bias the study's findings. This definition shifts focus away from the correctness of beliefs; instead, it emphasizes the beliefs themselves and their potential association with health outcomes. A variety of studies could fulfill our definition of a successfully blinded trial, including (i) studies in which groups are comparable with respect to beliefs and beliefs do not modify the effect of the intervention on outcomes; (ii) studies in which beliefs differ by group but are not associated with the outcome; and (iii) studies in which participants' beliefs are similar until a specific outcome assessment takes place that affects these beliefs. For the first two scenarios, successful blinding can be confirmed by demonstrating that belief patterns do not differ significantly by intervention group, and do not modify the exposure–outcome relationship. For the third scenario, beliefs must be ascertained before the outcome assessment to establish whether unblinding occurred at a time when it could have biased the outcome.

The main limitation of our study was that we did not ascertain baseline beliefs until two weeks' participation were completed; however, in our study beliefs at week 2 were remarkably similar by group. For future studies, we recommend ascertainment of beliefs after informed consent but before randomization to assess the effects of the enrollment and consent processes but not the intervention. In addition, we focused on participant blinding and did not measure beliefs among the research team. We acknowledge the importance of measuring blinding effectiveness among investigators, outcome assessors and analysts and propose that our methods be modified to suit these other important groups.

5. Conclusion

In summary, assessment of blinding effectiveness via changes in beliefs identified group-related differences in participants' beliefs that were not identified by cross-sectional analyses of beliefs at the end of the trial. The longitudinal evidence for unblinding was statistically significant and was supported by evidence that beliefs were associated with—and hence might bias—the primary study outcome, suggesting that the intervention effect should be reported within belief strata.

Our investigation raises concern about the sensitivity of the more commonly used cross-sectional analyses and the relevance of those based on the correctness of responses. The approaches we recommend to evaluate blinding effectiveness are (1) measurement of beliefs rather than correctness of beliefs, (2) a two-point sequential measure of belief patterns, (3) forced guesses for DK responders, (4) analysis of the main measure of effect and/or other reasons for participants' beliefs, stratified by belief patterns, and (5) a description of the reasons for participants' beliefs, to elucidate the possible role of beliefs in the causal pathway. Future studies should compare cross-sectional and longitudinal approaches in alternate settings, and attempt to measure the likely impact on trial results of the bias caused by beliefs about the intervention.

Acknowledgement

Funding for this work was provided through Cooperative Agreement U50/CCU915546-02-1 from the Centers for Disease Control and Prevention. For their contributions to this work, we thank Anne Benker, Sue Binder, Cliff Bowen, Susan Burns, Rebecca Calderon, Joshua Ergas, Kim Fox, Allen Hightower, Ron Hoffer, Alan Hubbard, Dennis Juranek, Asheena Khalakdina, Catherine Ma, Daniel

Mills, Gretchen Rothrock, Art Reingold, Sona Saha, Rick Sakaji, Suhminder Sandhu, Susan Shaw, Kate Steiner, Duc Vugia.

Appendix A

A.1. Howard's index [14]

The proportion of all N participants who "really knew" their group assignment is the difference between the proportions guessing their treatment correctly (n_1/N) and incorrectly (n_2/N) after "don't know" responses have been redistributed to the forced responses, "guess correctly" and "guess incorrectly":

$$H = (n_1 - n_2)/N.$$

A.2. James' blinding index [19]

$$BI = 1/2[(1 + n_3/N) + (1 - n_3/N)*K],$$

where n_3/N is the proportion of all N participants who don't guess (i.e., "don't know") and K is a measure of agreement between beliefs about assignments and actual assignments. One can see that participants who do guess are weighted by K, whereas participants who don't know are weighted by 1. It is not obvious that, within K, participants who guess correctly are weighted by 0 and those who guess incorrectly are weighted by 0.75.

Blinding is said to be adequate if BI and its confidence limits exceed 0.5.

Appendix B. Hypothetical data to show potential bias in relative risk caused by participants' beliefs

Exposure	Outcome	
	Disease	No disease
Truth		
Placebo group	40	20
Active group	20	40
		RR=0.50 (95% CI 0.34–0.75)
a. All participants believe "active"; under-reporting of the outcome		
Placebo group	32	28
Active group	16	44
		RR=0.50 (95% CI 0.31–0.81)
b. All participants believe "placebo"; over-reporting of the outcome		
Placebo group	44	16
Active group	28	32
		RR=0.64 (95% CI 0.47–0.87)

(*continued on next page*)

Appendix B (*continued*)

Exposure	Outcome	
	Disease	No disease
c. All participants respond incorrectly; differential reporting of the outcome		
Placebo group	32	28
Active group	28	32
		RR=0.88 (95% CI 0.61–1.25)
d. All participants respond correctly; differential reporting of the outcome		
Placebo group	44	16
Active group	16	44
		RR=0.36 (95% CI 0.23–0.57)

In a hypothetical study, 120 subjects were randomized to two groups of 60. We assume that, among participants who believe that they are receiving the placebo device, 20% of those without disease are misclassified as having disease. Among participants who believe they are receiving the active device, 20% of those with disease are misclassified as not having disease.

References

[1] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637–9.
[2] Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–94.
[3] 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. BMJ 2004;328:432.
[4] Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. Ann Intern Med 2002;136:254–9.
[5] Kiene H. A critique of the double-blind clinical trial. Part 1. Altern Ther Health Med 1996;2:74–80.
[6] Kiene H. A critique of the double-blind clinical trial. Part 2. Altern Ther Health Med 1996;2:59–64.
[7] Mattocks KM, Horwitz RI. Placebos, active control groups, and the unpredictability paradox. Biol Psychiatry 2000; 47:693–8.
[8] Diener H-C, Dowson AJ, Ferrari M, Nappi G, Tfelt-Hansen P. Unbalanced randomization influences placebo response: scientific versus ethical issues around the use of placebo in migraine trials. Cephalalgia 1999;19:699–700.
[9] Horwitz RI, Horwitz SM. Adherence to treatment and health outcomes. Arch Int Med 1993;153:1863–8.
[10] McPherson K, Britton AR, Wennberg JE. Are randomized controlled trials controlled? Patient preferences and unblind trials. J R Soc Med 1997;90:652–6.
[11] Desbiens NA. In randomized controlled trials, should subjects in both placebo and drug groups be expected to guess that they are taking drug 50% of the time? Med Hypotheses 2002;59(3):227–32.
[12] Ney PG, Collins C, Spensor C. Double blind: double talk or are there ways to do better research. Med Hypotheses 1986;21(2):119–26.
[13] Byington RP, Curb JD, Mattson ME. Assessment of double-blindness at the conclusion of the beta-blocker heart attack trial. JAMA 1985;253:1733–6.
[14] Howard J, Whittemore AS, Hoover JJ, Panos M. How blind was the patient blind in AMIS? Clin Pharmacol Ther 1982;32:543–53.
[15] Karlowski TR, Chalmers TC, Frenkel LD, Kapikian AZ, Lewis TL, Lynch JM. Ascorbic acid for the common cold. A prophylactic and therapeutic trial. JAMA 1975;231:1038–42.
[16] Juni P, Altman GA, Egger M. Assessing the quality of controlled clinical trials. BMJ 2001;323:42–6.
[17] Moher D, Schulz KF, Altman DG. The revised CONSORT statement for reporting randomized clinical trials. Ann Intern Med 2001;134:657–62.
[18] Hughes JR, Krahn D. Blindness and the validity of the double-blind procedure. J Clin Psychopharmacol 1985;5:138–42.

[19] James KE, Bloch DA, Lee KK, Kraemer HC, Fuller RK. An index for assessing blindness in a multi-centre clinical trial: disulfiram for alcohol cessation—a VA cooperative study. Stat Med 1996;15:1421–34.
[20] Marini JL, Sheard MH, Bridges CI, Wagner E. An evaluation of the double-blind design in a study comparing lithium carbonate with placebo. Acta Psychiatr Scand 1976;53:343–54.
[21] Moscucci M, Byrne L, Weintraub M, Cox C. Blinding, unblinding, and the placebo effect: an analysis of patients' guesses of treatment assignment in a double-blind clinical trial. Clin Pharmacol Ther 1987;41:259–65.
[22] Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology 1994;44:16–20.
[23] Bang H, Ni L, Davis CE. Assessment of blinding in clinical trials. Control Clin Trials 2004;25:143–56.
[24] Deyo RA, Walsh NE, Schoenfeld LS, Ramamurthy S. Can trials of physical treatments be blinded? The example of transcutaneous electrical nerve stimulation for chronic pain. Am J Phys Med Rehab 1990;69:6–10 [see comments].
[25] Colford JM, Rees JR, Wade TJ, Khalakdina A, Hilton J, Ergas I, et al. Participant blinding and gastrointestinal illness in a randomized, controlled trial of an in-home drinking water intervention. Emerg Infect Dis 2002;8:29–36.
[26] Rees JR. A drinking water evaluation trial. School of public health. Berkeley: University of California; 2001. p. 322.
[27] Payment P, Richardson L, Siemiatycki J, Dewar R, Edwardes M, Franco E. A randomized trial to evaluate the risk of gastrointestinal disease due to consumption of drinking water meeting current microbiological standards. Am J Public Health 1991;81:703–8.
[28] Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol 1996;25:1107–16.
[29] Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 1977;105:405–88.
[30] Miettinen OS, Cook EF. Confounding: essence and detection. Am J Epidemiol 1981;114(4):593–603.
[31] Weinberg C. Towards a clearer definition of confounding. Am J Epidemiol 1993;137:1–8.
[32] Robins J. The control of confounding by intermediate variables. Stat Med 1989;8:679–701.
[33] Robins JM. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60.