Journal of Clinical Epidemiology

Journal of Clinical Epidemiology 54 (2001) 541–549

COMMENTARY

The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf?

Ted J. Kaptchuk*

Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, KW-400, Boston, MA 02215, USA

Received 29 March 2000; received in revised form 1 September 2000; accepted 5 October 2000

Abstract

The double-blind randomized controlled trial (RCT) is accepted by medicine as objective scientific methodology that, when ideally performed, produces knowledge untainted by bias. The validity of the RCT rests not just on theoretical arguments, but also on the discrepancy between the RCT and less rigorous evidence (the difference is sometimes considered an objective measure of bias). A brief overview of historical and recent developments in “the discrepancy argument” is presented. The article then examines the possibility that some of this “deviation from truth” may be the result of artifacts introduced by the masked RCT itself. Can an “unbiased” method produce bias? Among the experiments examined are those that augment the methodological stringency of a normal RCT in order to render the experiment less susceptible to subversion by the mind. This methodology, a hypothetical “platinum” standard, can be used to judge the “gold” standard. The concealment in a placebo-controlled RCT seems capable of generating a “masking bias.” Other potential biases, such as “investigator self-selection,” “preference,” and “consent” are also briefly discussed. Such potential distortions indicate that the double-blind RCT may not be objective in the realist sense, but rather is objective in a “softer” disciplinary sense. Some “facts” may not exist independent of the apparatus of their production. © 2001 Elsevier Science Inc. All rights reserved.

Keywords: Double-blind randomized controlled trial; Gold standard; Bias; Artifacts; Placebo

1. Introduction

The double-blind randomized controlled trial (RCT) seeks to confer the ideal of scientific exactitude onto clinical experimentation in an effort to attain the objectivity of the laboratory model. A placebo-controlled RCT is considered medicine’s most reliable method for “representing things as they really are” [1]. While random error is mathematically estimated, systematic error is minimized by the rigorous application of methodological safeguards, especially randomization and blinding. Randomization aims to eliminate both unconscious and deliberate human influence on the assignment of subjects to different groups. Blind assessment ensures that treatment and analysis of outcomes are not colored by prejudice. Without these precautions, according to the standard epidemiological rationale, deliberate subversions (albeit well intentioned) or “subtle and intangible...subconscious” processes will affect the work of even the most conscientious researcher [2]. Assumed to be stripped clean of human bias, the masked (blind) RCT is accepted as the gold standard and thus above scrutiny as a potential source of systematic error.

However, it may be that experiments on humans by humans cannot circumvent the distortions and subversion of human consciousness and subjectivity. Do we need to consider “the effect of the experiment on the subjects themselves?” [3]. Is there a possibility that we need to “begin to speak of a ‘Heisenberg Principle of [Human Experimental] Sciences,’ where the very act of setting up controls can alter the phenomenon sufficiently to yield quite different results?” [4]. The general adoption of the double-blind RCT was based on theoretical reasons and intuitive attractiveness rather than a compelling body of data [5,6], and attempts to systematically investigate its assumed objectivity have been relatively scarce [7]. This article summarizes empirical evidence pertaining to the RCT’s capacity to produce undistorted and objective information. Shortcomings of imperfect RCTs are rarely examined in this essay; its focus is on possible systematic errors intrinsic in even an ideal RCT. Concealment in placebo-controlled trials is especially discussed in detail. It may be that every research methodology has inherent and random artifacts and that “truth” lies buried underneath multiple approaches.

* Corresponding author.

PII: S0895-4356(00)00347-4

2. The discrepancy argument

The primary empirical evidence for the objectivity of the double-blind RCT lies in the differential outcomes it detects compared with other research designs. Until very recently, there was a widespread perception that the absence of the usual components of the masked RCT will "exaggerate estimates of treatment effects" [8]. Often called a "measure of bias," this discrepancy among results achieved through different methodologies was accepted as evidence of the objectivity of a masked RCT [9]. It was generally believed that identical treatments "are much less likely to be judged efficacious in double-blind, randomized trials than in uncontrolled case series or unblinded, 'open' comparisons with contemporaneous or historical series of patients" [10]. The difference among treatment effects assessed by different methodologies can be of "the same magnitude as the observed treatment effect" compared to placebo in a double-blind, randomized trial [11].

The discrepancy argument was born, both theoretically and practically, at almost the very inception of concealed assessment (in fact in the same research report that coined the phrase "double blind") [5]. In 1950, Harry Gold and colleagues justified their innovative double-blinding technique by comparing the experiment's results with a preliminary, single-blind pilot study. The pilot study of 19 patients demonstrated that the cardiac drug khellin was dramatically better than placebo pill; a subsequent double-blind study of 39 patients demonstrated no difference between the two groups [12]. This comparison was the only concrete "empirical evidence" offered by Gold at an influential 1954 conference intended to introduce the emerging research model to the medical profession [13]. (The conference was also notable for introducing Freudian vocabulary into the language of epidemiology by describing "a psychic effect...[and] subtle mechanism which contrary to best intentions may give rise to misleading results...and unconscious bias") [13].

As the blind RCT became more common in the late 1950s and 1960s, researchers often adopted Gold's validation approach for demonstrating the new method's objectivity. A series of increasingly sophisticated papers (utilizing better statistical analyses and a larger number of patients) was published amidst much rhetoric about the need for scientific rigor. Invariably, it seemed, the more stringent the methodology (both in terms of randomization and blinding), the less efficacious the therapy (e.g., [14–19]). These early reports were used to advocate the adoption of the new research methods [20] and were the basis of the widespread belief that less rigorous evidence produces higher estimates of outcomes and favors new treatment.

However, three recent systematic reviews of the discrepancy evidence present a more complex picture. These new reviews make a compelling case that poor methodology could either overestimate or underestimate treatment effects. One analysis of these studies [21] included eight post-1977 studies that compared randomized and nonrandomized controlled trials of the same intervention; five showed that lack of randomization increased the estimate of treatment efficacy [22–26], two showed a decrease in efficacy [27,28], and one showed similar effects [29]. In studies that compared RCTs with non-RCTs across different interventions with diverse outcomes (converted to a standardized effect size) the results were inconsistent in two studies [30,31] and the third showed no difference between RCTs and non-RCTs [32]. Summarizing the evidence, this review found that "the deviation can go in either direction with the deviation of estimates of effect for non-randomized trials compared with randomized trials rang[ing] from an underestimation of effect of 76% to an overestimation of effect of 160% [21]. The two other systematic review, using slightly different entry criteria, one with 18 papers [33] and another with 14 papers [34], found roughly similar results [35]. It seemed that neither randomized nor nonrandomized methodologies consistently gave higher estimates of treatment effect and that variations between random and nonrandom evidence may not be greater than those between different RCTs [33].

3. A modified challenge to the discrepancy argument

The three recent comparisons of outcomes using different methodologies described above have slightly changed the discrepancy argument. Instead of consistently adjusting for bias in the direction of inflated estimates of effects, the methodological safeguards of randomization and blinding are now considered the "best protection against the unpredictability...of bias" [21]. All three studies comparing rigorous with less rigorous evidence agree that "a large, inclusive, fully blinded RCT...is likely to prove the best possible evidence of effectiveness" [33].

Two other recent reports challenge the discrepancy argument entirely (at least in terms of randomization). One based on 5 meta-analyses (99 reports) [36] and another on 19 treatment analyses (53 observational studies and 83 RCTs) [37], respectively, covering a wide range of therapeutic interventions, found that outcomes of observational studies and RCTs are "remarkably similar" [36]. Furthermore, one of these studies found that RCTs for the same condition, intervention and outcome produce more heterogeneous results then observational studies [36]. The editorial accompanying this pair of studies, however, has cast doubt on the validity of these conclusions [38]. The discrepancy debate has intensified.

In terms of explanations, the three studies that found discrepancies between randomized and nonrandomized evidence offered several potential reasons for any difference in outcomes with different methodologies. They included: "chance, bias, biased reporting and true heterogeneity (discrepancies due to differences in the participants or interventions in the randomized and nonrandomized studies)" [39] and preference effects. With the exception of patient preference effects (see discussion below), these distortions are primarily considered to be external to the apparatus of the masked RCT. The two reports that challenge the discrepancy phenomenon (at least in terms of randomization), while not denying the value of RCTs, argued that "research design should not be considered a rigid hierarchy" and that it is possible to perform accurate observational studies [36]. Also, it was thought that RCTs may be more valuable for "softer" outcomes where presumably bias operates more extensively [36].

In terms of blinding, recent comparative assessments remains consistent with the older evidence. Three studies showed that double-blind RCTs yielded significantly smaller treatment results than trials that were not double-blind [9,28,40]. Also, three studies showed that successful concealment of randomization (compared with inadequate concealment of randomization) produces smaller treatment outcomes [9,40,41]. Proper masking seems to create distinct outcomes; the discrepancy argument is intact in this domain.

Neither the modified discrepancy argument nor the absence of discrepancy in randomization argument raise the possibility that rigorous experimental conditions (especially blinding) can interact with humans and themselves cause departures from "truth." The rest of this article is an attempt to add an additional layer of complexity to the debate on what is "factual" evidence in medicine.

4. An overlooked aspect of the discrepancy debate: can an "unbiased" method produce its own distortions?

Overlooked in the discrepancy debate is that its logic is circular. It authenticates itself: "the truth is what we find out in such and such a way. We recognize it as truth because of how we find it out. And how do we know that the method is good? Because it gets at the truth" [42]. As one research team put it: "Unfortunately, there is no gold standard for judging the effectiveness of therapies apart from [double blind randomized] clinical trials" [43]. Is it possible that the exigencies of the double-blind RCT also contribute significant bias? While carefully designed experimental methods are critical to understanding reality, it may be that when it comes to conscious beings in health care situations, a residue of irreducible uncertainty can unpredictably cloud even an ideal scientific methodology.

There is no question that selection and measurement bias seriously distort research findings. However, the discrepancy argument traditionally goes one questionable step further: it assumes that any differences between masked RCTs and other research methods are due to deficiencies in the less stringent method. The ideal masked RCT is a priori considered a perfect tool and always innocent of any contribution to distortions from "reality." On the other hand, those who have recently challenged the discrepancy argument have not raised the possibility that ideal experimental conditions (especially the use of concealment) can influence clinical endpoints in unpredictable ways.

While most advocates of RCTs realize that outcomes in experiments in such a contrived set-up may not be the same as in ordinary clinical practice [44], the idea that the masked RCT "itself can give rise to biased results about outcomes [may] come as shock to many people" [45]. Does its demanding apparatus not eliminate the foibles of the mind? Yet human awareness and its potential distortion continues to operate even within the rarefied environment of a concealed RCT. One prominent researcher has called the conditions of an RCT "anathema to the human spirit" and has further said they "annoy human nature: [46]. Randomly subjecting a person to a milieu of hidden exposures and then spotlighting him or her with relentless observation does not nurture normalcy, nor does it isolate humans from their mental processes. Participants in RCTS are not immune to any "unconscious" processes that subvert science in less stringent experiments.

There are several ways in which the masking component in an ideal double-blind RCT could introduce bias into a trial. Knowledge that one has a chance of receiving placebo may introduce in a patient's perceptions uncertainty sufficient to decrease the magnitude of the response to either drug or placebo. Conversely, participation in an RCT may heighten sensitivity and vigilance on the part of either clinician or subject, thereby increasing the detection of beneficial (or adverse) responses. Participation in an RCT may create ambivalence, confusion, passivity, or absence of commitment among subjects (what researchers have called "resentful demoralization" [47] and "voluntary submission" [48]); any such factors could contribute to unpredictable reactions.

Further complicating matters, some of this reaction or reverberation may differentially affect the potency of drug and dummy controls. For example, vigilance may increase the drug effect and decrease the placebo response, or vice versa. Thus, there is potential for postrandomization forces to undermine the ability of the randomization process to create equivalent comparison groups [49,50]. This interference would also undermine the important assumption that in an RCT the placebo effect in the treatment arm equals the placebo effect of the placebo arm, an assumption that allows for commonly performed statistical tests [20].

5. Testing the temper of the gold standard: is there a "masking bias"?

Any external standard used in order to be "more objective" and verify the validity of the masked RCT would have to erect an even higher barrier to the subversive threat of human subjectivity. One theoretical example of this hypothetical "platinum" standard would be a trial in which both the patients and the dispensing physician were unaware that they were involved in a blind RCT. If patients were randomized to either "platinum" or routine RCT, one could compare results and thus test the temper of the gold standard. For ethical reasons, this option of concealment is generally not possible. Nonetheless, there have been several instances where versions, sometimes fragmentary, of this model have been tried. Although still rudimentary and seriously limited, collective results raise the possibility that the masked RCT, especially in relationship to concealment, does not ensure objective results: "reality" according to the gold standard may be different from the "reality" of the platinum standard.

Two experiments were performed in France immediately before informed consent became mandatory there in 1990. In the first trial, 30 carefully matched pairs of hospitalized patients with insomnia were randomly assigned to a double-blind RCT. Patients in one group and participating nurses were informed that the trial would compare a new hypnotic benzodiazepine drug to placebo (six patients in the informed group refused to participate). Patients in the matched control group were not informed that they were in a study.

Both groups of patients were given a single dose of placebo 1 or 2 h before sleep (by a nurse unaware of administering dummy pill to all) and were evaluated the next morning. The "control" patients (those unaware of being in an experiment) experienced hypnotic activity significantly higher than those in the informed group ($P < .05$) [51].

While interesting, this result is not necessarily surprising; it is sometimes thought that the magnitude of nonspecific effects may be different in RCTs [44]. Also, one could assume that any "distortions" of the placebo effect under informed consent would be similarly distributed to those on active treatment and those on placebo pill and therefore would be inconsequential in determining a drug–placebo difference. Still, the mitigation of therapeutic benefit by informed consent could well be a hidden source of "bias."

A subsequent French trial expanded on this earlier trial and concurrently examined both active drug and placebo under the two different standards. Secretly, 49 consecutive hospitalized patients with mild or moderate cancer pain (not requiring opiates) were randomly chosen to be informed—or not informed—of their participation in a randomized, double-blind, placebo–controlled, crossover experiment [52]. Completely unaware of the experiment, 25 "control" patients received naproxen and placebo pill in random order (they thought they were receiving routine care for pain). Eighteen of 24 in the informed group (seven refused to participate) agreed to enter an identically designed experiment under routine double-blind RCT conditions. The outcome measured was visual analog scales of pain 60, 120 and 180 min after taking the pill. In the informed group, *both* the naproxen and placebo were significantly more effective than in the unaware group (P=.012) (Fig. 1). In fact, the placebo intervention in the informed group was significantly more effective than the naproxen in the unaware group!

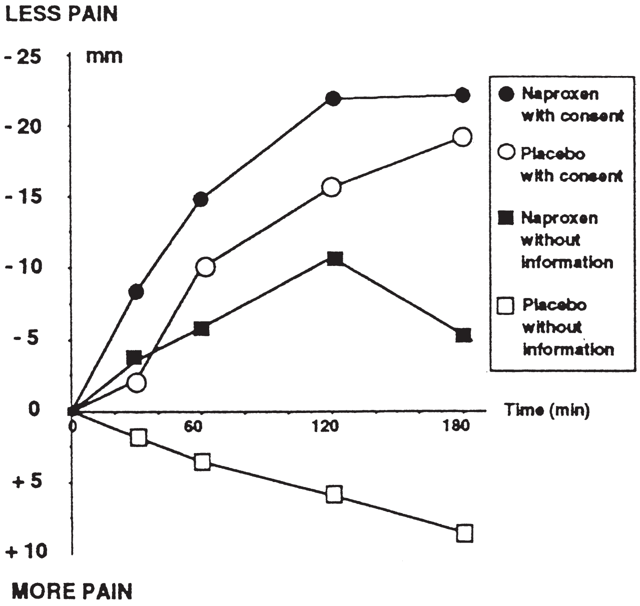

Fig. 1. Changes (mm) in 100 mm pain visual analogue scales after naproxen and after placebo in patients with (n = 18) and without (n = 25) information concerning the crossover placebo-controlled study (zero is given by the pain score at time 0). Reprinted from Bergmann JF et al. A randomised clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer pain. Clinical Trials and Meta-Analysis 1994;29:41–47, with permission from Elsevier Science.

In terms of the differential effect of informed consent on placebo, the outcome of this experiment is exactly the opposite of the previous sleep study. Here, concealment decreased placebo response. (In this case, informed consent seemed to initiate a swamping effect reminiscent of the "Hawthorne effect" first reported by industrial psychologists [53,54]). More importantly, however, this second trial provides evidence that knowledge of being in a masked RCT versus concealed randomization (i.e., absence of knowledge of informed consent) can affect the treatment results of the placebo arm and active drug disproportionately. If knowledge of concealment differentially affected the active drug and placebo, then it seems that masking can introduce a secondary, postrandomization variable into the experiment. It seems that the platinum standard can produce an outcome very different from the outcomes produced by the gold standard! Also the circumstantial sensitivity and variability of the placebo effect (in this trial, in the previous one and in the examples below) raises the possibility that a different amount of placebo response can be embedded in the drug arm versus the placebo arm of an RCT. (It should be noted that this trial could almost be considered platinum standard but for a lack of clarity in the report as to whether the researchers were unaware of the patient's group assignment at the time of evaluation).

The French researchers jettisoned informed consent to enhance the concealment of a normal blind RCT. A second line of research in the United States used another sort of deception to study whether dummy controls affect cognitive awareness sufficiently to cause placebo intervention and/or drugs to have actions different from those in nonexperimental settings.

In the first of these experiments, 100 volunteers were recruited and told they might receive a real drug or a placebo pill. Half were then told that they would be enrolled in a double-blind trial in which they would receive either placebo (decaffeinated coffee) or active drug (caffeinated coffee). The other half were deceptively informed that they would be receiving real coffee. All subjects were then administered placebo. Subjects were tested for alertness, tension, motor performance, pulse rate, systolic blood pressure, and certainty of having consumed caffeine.

Unexpectedly, the double-blind administration of the placebo produced effects that in most instances were different from—and usually opposite to—those produced by clinical administration of placebo, when people are not given reason to doubt that they are receiving a pharmacologically active agent [55]. For example, deceptive administration of the placebo resulted in a significant increase in pulse rate ($P<.05$) and increase in alertness ($P<.003$) compared with double-blind administration. In this experiment, which—like the first French trial—examined only placebo, it appears that the certainty of receiving a drug maximizes a placebo effect, and that the introduction of doubt as to whether or not one has received a real treatment diminishes the effect.

To examine both real drug and placebo in a masked RCT another team of researchers elaborated on this method of deception. Seventy-two smokers quit smoking and gave informed consent indicating that they would receive either nicotine gum or dummy treatment. They were then randomly assigned in a 3 × 2 design to six distinct groups. The three horizontal groups were distinct cognitive sets: a) told they were receiving nicotine gum (i.e., clinical practice conditions); b) told they were receiving placebo gum; and c) told they would receive either nicotine or placebo gum (i.e., placed under normal double-blind conditions). Vertically, the groups were divided by receiving medication or placebo. Outcome measures included smoking behavior, withdrawal effects, amount of gum used, and self-reported effects of the gum. Nicotine appeared to increase quit rates and perceived drug effects under "experimental" conditions, but not under "therapeutic" conditions ($P<.05$) [56]. [A trend for interaction between cognitive state and other smoking variables (e.g., withdrawal symptoms and self-administration of nicotine gum) was also detected but did not reach statistical significance.] One of the conclusions of the authors was that the trial apparatus was not neutral: drug effects under clinical conditions were different from those under double-blind conditions.

A third experiment essentially replicated the 3 × 2 design of the nicotine experiment with caffeine. After informed consent, 100 subjects were randomly assigned to one of the following groups: a) informed they would receive caffeine citrate; b) informed that they would receive a decaffeinated preparation; or c) informed that they would receive either a caffeinated or a decaffeinated beverage (double-blind condition) [57]. They were then randomly given either drug or no active drug. Measurements were taken over 15, 30, and 45 min. Outcome measures included blood pressure, pulse rate, alertness, tension, and certainty of receiving caffeine.

Cognitive states interacted with drug and placebo on at least several outcome measures at all measurement points. For example, subjects who had been told they were not receiving caffeine reported being significantly less alert than those who were told they were receiving caffeine ($P<.01$) or those who were given double-blind instructions ($P<.05$). Subjects who were told they would receive caffeine and in fact did receive caffeine reported significantly greater tension than subjects in any other groups ($P<.01$). Subjects given double-blind instruction had significantly lower diastolic blood pressure than subjects who had been either informed or misinformed about the content of their beverages ($P<.05$). The presence of caffeine was reliably discriminated only by subjects who had been given double-blind instructions. Because of the radically different outcomes that the cognitive context of drug consumption produced, the authors considered that their result, in conjunction with the nicotine experiment above, showed that double-blind studies may lead to erroneous conclusions about the clinical effects of particular drugs.

These three experiments were preceded and inspired by an earlier line of controlled experiments where it was shown that many active drugs could be distinguished by subjects only if they knew what to expect: complete concealment dramatically changed pharmacological effects [58]. Experiments with such drugs and phenmetrazine [59], epinephrine [60], amphetamine [61] and chloral hydrate [61] consistently demonstrated the importance of expectation as therapeutic adjunct. Also bronchoactive substances (isoproterenol and carbachol in aerosol) produced greater airway reactivity when the cognitive state was consonant with the pharmacologic action of the drugs [62,63]. In all these cases, those deprived of expectation experienced significantly diminished physiological effects from drugs. The alcohol literature also points in the same direction: when individuals have expectations that are contrary to the pharmacological effects of the drug, it is expectations rather than the drug's pharmacology that usually prevail [64,65]. Also supporting this evidence that cognitive states can interact with drug effects are even earlier experiments that showed that potent physiological effects can be reversed under deceptive conditions. For example, Wolf showed that subject misled about a drug's effect could experience cholinergic reactions from atropine or an antiemetic effect from ipecac [66]. In all these cases, concealment or changes in cognitive awareness had impacts on active medication. Masking was not a "neutral" device.

These experiments on concealment are intriguing—but even taken together—are not ultimately persuasive. They mainly concern short-term effects that may not be applicable to a genuine long-term RCT. Also, they mainly concern subjective end-points, which are especially susceptible to placebo effects {measurement error. It is unclear in what circumstances the results would apply. Nonetheless, they suggest that masking may not be neutral and that concealment in an RCT may produce a "masking bias."

6. "Investigator self-selection," "preference" and other sources of bias in RCTs

Besides a possible "masking" bias, the RCT apparatus can generate other sources of potential bias. Some of these potential problems are rarely discussed. Others are well known (especially those that deal with circumstances that affect the external validity of the trial and are perhaps less pertinent in a discussion of an "ideal" RCT) and have been extensively described elsewhere (e.g., [67]). A very brief review of some of these potential internal and external confounders may be helpful in summarizing the possibility that an RCT can introduce its own deviations from "truth."

In RCTs, patients are always randomized, while researchers rarely are. This could create and "investigator self-selection" bias. It has been noted that clinical researchers are not typical of regular practitioners [44]. Some evidence exists that different providers elicit different placebo responses. For example, one study has shown that two different researchers when they gave placebo pills either consistently increased or consistently decreased patient's gastric secretions [68]. Such differential effects have been confirmed in other studies that made the physician the interdependent variable [69,70]. Prospective RCTs have demonstrated that different styles of health care produce measurable differences in outcomes [71,72]. Two experiments even observed that practitioner behavior and attitude can interact unequally with active drug and placebo and reverse the ability of an RCT to register a drug–placebo difference [73,74]. Some researchers have even begun to speak of the necessity "to select random samples of both patients and therapists" [75] or perform a physician "run-in" phase in RCTs to ensure that any special characteristics of providers not threaten the integrity of a trial [76].

Randomization necessarily eliminates preference, intentionality and a sense of control. However, the essence of clinical decision-making is choice, and the process in which a treatment regimen is negotiated between physician and patient may itself confer a therapeutic benefit [45]. The psychosocial literature gives indirect support to this thesis [77] and the adherence literature points out that the act of compliance with any treatment, including placebo, affects not only symptom relief, but also hard outcomes, such as survival [78,79]. It may be that the elimination of options, tailoring to individual preferences, and emotional investment in the selected treatment choice systematically distort some RCT outcomes. This potential "preference bias" may be even more critical in unblind "participative" interventions, including self-monitoring, diet, exercise programs, or counseling, where a strong subject preference for—or aversion to—a particular treatment may change treatment effect [80].

Human behavioral reactions to the randomization process itself may contribute to systematic error. As Alvan Feinstein has pointed out, unlike agricultural plots, human subjects in RCT volunteer, an act that may render them unrepresentative of a random sample [81]. Numerous studies have shown that many patients eligible for RCTs do not consent to participate, and that these patients are not homogenous with those who enter trials [82–84]. Those who refuse to volunteer may differ in important prognostic features [85,86]. A recent study suggests that subjects participating in RCTs evaluating treatment of medical conditions tend to be not as affluent, not so well educated, and not as healthy as those who do not and also that the opposite is true for prevention trials [33]. Because subjects in an RCT are a subset of the patient population, "nonconsent bias" may create a discrepancy between the RCT outcome and the clinical outcome because the clinical population may contain a different spectrum of patients than those who participated in the trial [36,87–89]. Some have argued that such a bias could conceivably change a significant result into a nonsignificant one or even reverse the direction of outcome [90,91].

Another "consent bias" can also exist. Evidence suggests that the extensive disclosure requirements of informed consent in RCTs may significantly increase adverse effects and drop-out rates [92–94].

7. Conclusion: is the double-blind RCT objective?

The masked RCT attempts to provide a method that can free medical research from the fallibility of the human mind. Some experimental evidence shows that masking cannot completely neutralize the potential distortions of human consciousness and subjectivity. Such bias may threaten the internal validity of information produced. Preference effects also have this potential. Human behavior, such as a patient's refusal to enter trials or the researcher's reluctance to randomize practitioners, can also generate bias that threatens the external validity of RCTs. The simple fact may be that humans are effected by experimental conditions in unpredictable ways. Although far from conclusive and with many limitations, the available evidence presented in this article concerning the potential of placebo–controlled RCTs to produce bias prompts the question whether the masked RCT may produce its own deviations from truth. It seems that the most "rigorous" evidence may produce deviations from the "truth."

The claim that the RCT is objective may fall short of a "hard" correspondence with reality. Still, the blind RCT may be objective in a "softer" or disciplinary sense: it is a standardized, explicit, replicable, and impersonal procedure that defines unambiguous and formal norms for medical researchers. Its system of rules minimizes the need for personal trust and subjective judgment and "limit[s] the exercise of [personal] discretion" [95].

Undoubtedly, the double-blind RCT is the least subjective and most impersonal procedure ethically possible now. It may be the closest thing medicine has to a "technology of trust" [96]. But "fairness" rather than "truth" may be its central value [97]. Its objectivity may be closer to what one modern philosopher calls "inner realism," a practical, more modest claim that takes "experimental facts" and acts as if they are truth [98].

The intention of this article has not been to discredit the masked RCT, but to foster a balanced discussion of research methods. Any method of evidence production (from gold to the baser metals) has its own potential distortions; each needs to be weighed and studied. It may be that in some circumstances even the best instruments of detection effect the phenomena being measured and that medical "facts" may not exist independent of the apparatus of their production. While a gold standard is valuable, any worshipping at an altar of a golden calf, like the Biblical Exodus story, may obscure "reality." Giving any research method, including double-blind RCTs "a sanctified scientific status can be an obstacle "to improving the basic structure and evidence of trials" [99]. Unless one is aware of a research methodology's potential weaknesses, scientific activity can become a mechanical ritual.

Acknowledgments

The critical feedback of Fred Mosteller, Al Fishman and John Emerson and the editorial assistance of June Cobb are gratefully appreciated. Also I wish to thank the Seminar on Effective and Affordable Health Care at Harvard University for allowing a presentation of an earlier version of this paper.

References

[1] Rorty R. Philosophy and the mirror of nature. Princeton: Princeton University Press, 1977.

[2] Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. St. Louis: Mosby, 1985.

[3] Weinstein MC. Allocation of subjects in medical experiments. N Engl J Med 1974;291:1278–85.

[4] Fisher S, Cole JO, Rickels K, Uhlenhutt EH. Drug-set interaction: the effect of expectation on drug response in outpatients. In: Bradley PB, Flügel F, Hoch PH, editors. Neuropsychopharmacology vol. 3. New York: Elsevier, 1964.

[5] Kaptchuk TJ. Intentional ignorance: a history of blind assessment and placebo controls. Bull Hist Med 1998;72:389–433.

[6] Gehan EA, Freireich EJ. Non-randomized controls in cancer clinical trials. N Engl J Med 1974;290:198–203.

[7] Kleijnen J, de Craen AJM, van Everdingen J, Krol L. Placebo effect in double-blind clinical trials: a review of interactions with medications. Lancet 1994;344:1347–49.

[8] Sibbald B, Roland M. Why are randomized controlled trials important? Br Med J 1998;316:201–2.

[9] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. J Am Med Assoc 1995;273:408–12.

[10] Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1986;89(Suppl.):2S–3S.

[11] Pocock SJ. Allocation of patients to treatment in clinical trials. Biometrics 1979;35:183–97.

[12] Greiner T, Gold H, Cattell M, Travell J, Bakst H, Rinzler SH, Benjamin ZH, Warshaw LJ, Bobb AL, Kwit NT, Modell W, Rothendler HH, Messeloff CR, Kramer ML. A method for the evaluation of the effects of drugs on cardiac pain in patients with angina on effort. A study of Khellin (Visammin). Am J Med 1950;9:143–55.

[13] Conference on Therapy. How to Evaluate a New Drug. Am J Med 1954;17:722–7.

[14] Foulds GA. Clinical research in psychiatry. J Ment Sci 1958;104:259–65.

[15] Glick BS, Margolis R. A study of the influence of experimental design on clinical outcome in drug research. Am J Psychol 1962;118:1087–96.

[16] Astin A, Ross S. Glutamic acid and human intelligence. Psychol Bull 1960;57:429–34.

[17] Wechsler H, Grosser GH, Greenblatt M. Research evaluating antidepressant medications on hospitalized mental patients: a survey of published reports during a five-year period. J Nerve Ment Dis 1965;141:231–9.

[18] Grace ND, Muench H, Chalmer TC. The present status of shunts for portal hypertension in cirrhosis. Gastroenterology 1996;50:684–91.

[19] O'Brien WM. Indomethacin: a survey of clinical trials. Clin Pharm Ther 1967;9:94–107.

[20] Kaptchuk TJ. Powerful placebo: the dark side of the randomized controlled trial. Lancet 1998;351:1722–5.

[21] Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomized and non-randomized clinical trials. Br Med J 1998;317:1185–90.

[22] Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med 1977;297:1091–6.

[23] Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. Am J Med 1982;72:233–40.

[24] Diehl LF, Perry DJ. A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid? J Clin Oncol 1986;4:1114–20.

[25] Pyorala S, Huttunen NP, Uhari M. A review and meta analysis of hormonal treatment of cryptorchidism. J Clin Endocrinol Metab 1995;80:2795–9.

[26] Carroll D, Tramer M, McQuay H, Nye B, Moore A. Randomization is important in studies with pain outcomes: systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. Br J Med 1996;77:798–803.

[27] Reimold SC, Chalmers TC, Berlin JA, Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic atrial fibrillation: observations on the role of trial design and implications of drug related mortality. Am Heart J 1992;124:924–32.

[28] Recurrent Miscarriage Immunotherapy Trialists Group. Worldwide collaborative observational study and meta-analysis on allogenic leukocyte immunotherapy for recurrent spontaneous abortion. Am J Reprod Immunol 1994;32:55–72.

[29] Watson A, Vanderkerckhove P, Lilford R, Vail A, Brosens I, Hughes E. A meta-analysis of the therapeutic role of oil soluble contrast media at hysterosalpingography: a surprising result? Fertil Steril 1994;61:470–7.

[30] Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: medical. Stat Med 1989;8:441–54.

[31] Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: surgical. Stat Med 1989;8:455–66.

[32] Ottenbacher K. Impact of random assignment on study outcome: an empirical examination. Control Clin Trials 1992;13:50–61.

[33] McKee M, Gritton A, Black N, McPherson K, Sanderson C, Bain C. Interpreting the evidence: choosing between randomized and non-randomized studies. Br Med J 1999;319:312–15.

[34] Reeves BC, MacLehose RR, Harvey IM, Sheldon TA, Russell IT, Black AMA. Comparison of effect size estimates derived from randomised and non-randomised studies. In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. London: BMJ Publishing, 1998.

[35] Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Three systematic reviews—not so different answers? [Letter] eBMJ. http://www.bmj.org/cgi/eletters/319/7205/312.

[36] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, obser-

vational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–92.

[37] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–88.

[38] Pocock SJ, Elbourne DR. Randomized trials or observational tribulations. N Engl J Med 2000;342:1907–9.

[39] Kunz R, Oxman A. Two systematic reviews-two different answers? [Letter] eBMJ. http://www.bmj.org/cgi/eletters/319/7205/312.

[40] Moher D, Ba'Pham, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does the quality of randomized trials affect estimates of intervention efficacy reported in meta-analysis? Lancet 1998;352:609–13.

[41] Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983;309: 1358–61.

[42] Hacking I. Statistical language, statistical truth and statistical reason: the self-authentication of a style of scientific reasoning. In: McMullin E, editor. The social dimensions of science. Notre Dame: University of Notre Dame Press, 1992.

[43] Sacks HS, Chalmers TC, Smith H. Sensitivity and specificity of clinical trials. Randomized *v* historical controls. Arch Intern Med 1983; 143:753–5.

[44] Black N. Why we need observational studies to evaluate the effectiveness of health care. Br Med J 1996;312:1215–18.

[45] McPherson K. The best and the enemy of the good: randomised controlled trials, uncertainty, and assessing the role of patient choice in medical decision making. J Epidemiol Comm Health 1994;48:6–15.

[46] Schulz KF. Subverting randomization on controlled trials. J Am Med Assoc 1995;274:1456–8.

[47] Torgerson DJ, Sibbald B. What is a patient preference trial? Br Med J 1998;316:360.

[48] Silverman WA, Altman DG. Patients' preferences and randomized trials. Lancet 1996;347:171–4.

[49] Urbach P. Randomization and the design of experiments. Philos Science 1985;52:256–73.

[50] Kempthorne O. Why randomize? J Stat Plan Inf 1977;1:1–25.

[51] Dahan R, Caulin C, Figea L, Kanis JA, Cauline R, Segrestaa JM. Does informed consent influence therapeutic outcome? A clinical trial of the hypnotic activity of placebo in patients admitted to hospital. Br Med J 1986;293:363–4.

[52] Bergmann JF, Chassany O, Gandiol J, Deblois P, Kanis JA, Segrestaa JM, Caulin C, Dahan R. A randomized clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer pain. Clin Trials Meta-Anal 1994;29:41–7.

[53] Roethlisberger FJ, Dickson WJ, Wright HA. Management and the worker. Cambridge: Harvard University Press, 1946.

[54] Bouchet C, Guillemin F, Briançon S. Nonspecific effects in longitudinal studies: impact on quality of life measures. J Clin Epidemiol 1996;1996:49:15–20.

[55] Kirsch I, Weixel LJ. Double-blind versus deceptive administration of a placebo. Behav Neurosci 1988;2:319–23.

[56] Hughes JR, Gulliver SB, Amori G, Mireault GC, Fenwsick JF. Effect of instructions and nicotine on smoking cessation, withdrawal symptoms and self-administration of nicotine gum. Psychopharmacology 1989;99:486–91.

[57] Kirsch I, Rosadino MJ. Do double-blind studies with informed consent yield externally valid results? Psychopharmacology 1993;110: 437–42.

[58] Dinnerstein AJ, Lowenthal M, Blitz B. The interaction of drugs with placebos in the control of pain and anxiety. Perspect Biol Med 1966; 10:103–14.

[59] Penick SB, Hinkle LE. The effect of expectation on response to phenmetrazine. Psychosom Med 1964;26:369–73.

[60] Penick SG, Fisher S. Drug-set interaction: psychological and physiological effects of epinephrine under differential expectations. Psychosom Med 1965;27:177–82.

[61] Lyerly SB, Ross S, Krugman AD, Cylde DJ. Drugs and placebos: the effects of instruction upon performance and mood under amphetamine sulphate and chloral hydrate. J Abnorm Soc Psychol 1964;68: 321–7.

[62] Luparello TJ, Leist N, Lourie CH, Sweet P. The interaction of psychologic stimuli and pharmacologic agents on airway reactivity in asthmatic subjects. Psychosom Med 1970;32:509–13.

[63] Sodergren SC, Hyland ME. Expectancy and asthma. In: Kirsch I, editors. How expectations shape experience. Washington, DC: American Psychological Association, 1999.

[64] Kirsch I. Response expectancy as a determinant of experience and behavior. Am Psychol 1985;40:1189–202.

[65] Hull JC, Bond CF. Social and behavioral consequences of alcohol consumption and expectancy: a meta-analysis. Psychol Bull 1986;99: 347–60.

[66] Wolf S. Effects of suggestion and conditioning on the action of chemical agents in human subjects—the pharmacology of placebos. J Clin Invest 1950;29:100–9.

[67] Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. Health Technol Assess 1998;2:1–124.

[68] Wolf S. Part IV. Placebos: problems and pitfalls. Clin Pharmacol Ther 1962;3:254–7.

[69] Sarles H, Camatte R, Sahel J. A study of the variations in the response regarding duodenal ulcer when treated with placebo by different investigators. Digestion 1977;16:289–92.

[70] Joyce CRB. Differences between physicians as revealed by clinical trials. Proc Soc Med 1961;28:12–4.

[71] Thomas KB. General practice consultations: is there any point in being positive? Br Med J 1987;294:1200–2.

[72] LeBaron S, Reyher J, Stack JM. Paternalistic vs. egalitarian physician styles: the treatment of patients in crisis. J Fam Med 1985;21:56–62.

[73] Uhlenhuth EH, Canter A, Neustadt JO, Payson HE. The symptomatic relief of anxiety with meprobamate, phenobarbital and placebo. Am J Psychiatry 1959;115:905–10.

[74] Uhlenhuth EH, Rickels K, Fisher S, Park LC, Lipman RS, Mock J. Drug, doctor's verbal attitude and clinic setting in the symptomatic response to pharmacotherapy. Psychopharmacologia (Berl) 1966;9: 392–418.

[75] Martindale C. The therapist-as-fixed-effect fallacy in psychotherapy research. J Controlled Clin Psychol 1978;46:1526–30.

[76] Shapiro AK, Shapiro E. The powerful placebo: from ancient priest to modern physician. Baltimore: The Johns Hopkins University Press, 1997.

[77] Siegrist J, Peter R, Junge A, Cremer P, Seidel D. Low status control, high effect at work and ischaemic heart disease: prospective evidence from blue-collar men. Soc Sci Med 1990;31:1127–234.

[78] Horwitz RI, Viscolli CM, Berkman L, Donaldson RM, Horwitz SM, Murray CJ, Ransohoff DF, Sindelar J. Treatment adherence and risk of death after a myocardial infarction. Lancet 1991;336:543–5.

[79] Horwitz RI, Horwitz SM. Adherence to treatment and health outcomes. Arch Intern Med 1993;153:1863–8.

[80] Brewin CR, Bradley C. Patient preferences and randomised clinical trials. Br Med J 1989;299:313–5.

[81] Feinstein AR. Statistics versus science in the design of experiments. Clin Pharm Ther 1970;11:282–92.

[82] MIAMI Trial Research Group. Patient population. Am J Cardiol 1985;56:10G–14G.

[83] Smith P, Arnesen H. Mortality in non-consenters in a post-myocardial infarction trial. J Int Med 1990;228:253–6.

[84] Fairhurst K, Dowrick C. Problems with recruitment in a randomized controlled trial of counselling in general practice: causes and implications. J Health Serv Res Policy 1996;1:77–80.

[85] Llewellyn-Thomas HA, McGreal MJ, Thiel EC, Fine S, Erlichman C. Patients' willingness to enter clinical trials: measuring the association with perceived benefit and preference for decision participation. Soc Sci Med 1991;32:35–42.

[86] Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. Br Med J 1984;289:1281–4.
[87] Schooler NR. How generalizable are the results of clinical trials? Psychopharmacol Bull 1980;16:29–31.
[88] Marcus SM. Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. J Clin Epidemiol 1997;50:823–8.
[89] Pringle M, Churchill R. Randomised controlled trials in general practice. Gold standard or fool's gold? Br Med J 1995;311:1382–3.
[90] Schooler NR, Levine J, Severe JB, Brauzer B. Prevention of relapse in schizophrenia. Arch Gen Psychol 1980;37:16–24.
[91] Edlund JM, Craig TJ, Richardson MA. Informed consent as a form of volunteer bias. Am J Psychol 1985;142–624–7.
[92] Levine RJ. The apparent incompatibility between informed consent and placebo-controlled clinical trials. Clin Pharmacol Ther 1987;42: 247–9.
[93] Myers MG, Cairns JA, Singer J. The consent form as a possible cause of side effects. Clin Pharmacol Ther 1987;42:250–3.
[94] Simes RJ, Tattershall MHN, Coastes AS, Raghavan D, Solomon HJ. Randomised comparison of procedures for obtaining informed consent in clinical trials of treatment for cancer. Br Med J 1986;293:1065–8.
[95] Porter TM. Objectivity as standardization: the rhetoric of impersonality in measurement, statistics, and cost-benefit analysis. Ann Scholar 1992;9:19–60.
[96] Mathews JR. Quantification and the quest for medical certainty. Princeton: Princeton University Press, 1995.
[97] Megill A. Introduction: Four senses of objectivity. In: Megill A, editor. Rethinking objectivity. Durham: Duke University Pres, 1994.
[98] Putnam H. Reason, truth, and history. Cambridge: Cambridge University Press, 1981.
[99] Feinstein AR. Meta-analysis: statistical alchemy for the 21st century. J Clin Epidemiol 1995;48:71–9.