**TWO META-ANALYSES OF NONCONTACT HEALING STUDIES**

Chris A. Roe, PhD* Charmaine Sonnex, BSc, MSc, and Elizabeth C. Roxburgh, BSc, PhD

**Objective:** Reviews of empirical work on the efficacy of noncontact healing have found that adopting various practices that incorporate an intention to heal can have some positive effect upon the recipient’s wellbeing. However, such reviews focus on ‘whole’ human participants who might be susceptible to expectancy effects or benefit from the healing intentions of friends, family or their own religious groups. We proposed to address this by reviewing healing studies that involved biological systems other than ‘whole’ humans (e.g., studies of plants or cell cultures) that were less susceptible to placebo-like effects. Secondly, doubts have been cast concerning the legitimacy of some of the work included in previous reviews so we planned to conduct an updated review that excluded that work.

**Data Sources:** The following databases were searched: Swetswise, ASSIA, Psych-NET, Web of Science, Cochrane Library, British Nursing Index, Cinahl Full Text, and Informaworld.

**Study Selection:** Only studies in English were eligible for inclusion. All studies must have examined the effects upon a biological system of the explicit intention to improve the wellbeing of that target; 49 non-whole human studies from 34 papers and 57 whole human studies across 56 papers were included.

**Data Synthesis:** The combined weighted effect size for nonwhole human studies yielded a highly significant $r$ of .258, but outcomes were heterogeneous and correlated with blind ratings of study quality; 22 studies that met minimum quality thresholds gave a reduced but still significant weighted $r$ of .115. Whole human studies yielded a small but significant effect size of $r = .203$. Outcomes were again heterogeneous, and correlated with methodological quality ratings; 27 studies that met threshold quality levels gave an increased $r = .224$.

**Conclusions:** Results suggest that subjects in the active condition exhibit a significant improvement in wellbeing relative to control subjects under circumstances that do not seem to be susceptible to placebo and expectancy effects. Findings with the whole human database suggests that the effect is not dependent upon the previous inclusion of suspect studies and is robust enough to accommodate some high profile failures to replicate. Both databases show problems with heterogeneity and with study quality and recommendations are made for necessary standards for future replication attempts.

**Key words:** Noncontact healing, Distance Healing, Meta-analysis, Reiki, Johrei, Therapeutic Touch, Intercessory Prayer and Wellbeing

(Explore 2015; 11:11-23 © 2015 Published by Elsevier Inc.)

---

**INTRODUCTION**

The supposed linkage between religious beliefs and practices and health has long been of interest to psychologists since it provides suggestive evidence for a connection between psycho-spiritual factors and physical wellbeing. This research is an extension of conventional accounts of the health benefits of religiosity and/or spirituality that supposes that they are mediated by cognitive and behavioral differences, with those expressing a religious faith tending to be more optimistic and resilient, to believe that the physical world is essentially orderly and meaningful, to engage in healthy behaviors such as regular exercise or meditation, and to avoid unhealthy behaviors such as drug and alcohol abuse and promiscuous or risky sex (for reviews see Fontana8 and Koenig et al.). More intriguingly, a number of reviews of the efficacy of healing have found that interceding on behalf of patients through prayer or by adopting various practices that incorporate an intention to heal can have some positive effect upon their well-being. However, these reviewers also raised concerns about study quality and the diversity of healing approaches adopted in the studies under review—ranging from techniques that usually involve close physical proximity between the practitioner and the patient, such as therapeutic touch and Reiki healing, through to techniques that work at a distance, such as psychic healing or intercessory prayer to a higher being—and this makes the findings difficult to interpret, since in some cases, the beneficial effects could be attributable to placebo effects or to the consequences of general lifestyle changes that are involved in holistic approaches to medicine. The diversity of approaches included under the rubric of healing also presents problems in explaining the observed effects, since there is so little common ground that it is difficult to conceive of a mechanism that they might all share.

Centre for the Study of Anomalous Psychological Processes, The University of Northampton, Park Campus, Boughton Green Road, Northampton, Northants NN2 7AL

* Corresponding author.

e-mail: chris.roe@northampton.ac.uk
Some of these concerns can be addressed by conducting double-blind randomized controlled clinical trials. These entail the random allocation of participants (or patients) to either a treatment or control condition so as to control for selection bias (or alternatively participants are matched on the basis of other variables that are thought to affect the prognosis of their health condition, such as age, gender, co-morbidity, and so on), with patients and attending physicians remaining blind to the allocation so as to control for placebo improvements. Such a design has been described by Astin et al. as meeting minimum standards for research quality.

Perhaps the first study (and certainly the most influential) that met these criteria is Byrd’s consideration of the effects of intercessory Judeo-Christian prayers with a population of 393 coronary care unit patients. Participants were randomly assigned on a double-blind basis to either a control or a prayer group on admission to the unit. Each participant in the prayer group was assigned to between three and seven intercessors, who were given the patient’s name, diagnosis, general condition, and updates on their condition throughout the trial (but not sufficient information to be able to trace the patient). The intercessors themselves were from a variety of Protestant and Roman Catholic churches, the only conditions to becoming an intercessor were that they had to be “born again” according to the Gospel of John 3:3 and that they should “lead an active Christian life as manifested by daily devotional prayer and active Christian fellowship with a local church” (p. 827). Intercessory prayer was conducted daily and involved asking for a “rapid recovery, and for prevention of complications and death, in addition to other areas of prayer they believed to be beneficial to the patient” (p. 827). Byrd found that the prayer group presented with significantly fewer cases of pneumonia, congestive heart failure, intubation/ventilation, cardio pulmonary arrest, and significantly less need for antibiotics and diuretics. Significantly more participants in the prayer group also showed a “good” hospital course, i.e., “no new diagnoses problems or therapies were recorded for the patient or if events occurred that only minimally increased the patient’s morbidity or risk of death” (p. 828).

Other well-controlled studies have also reported positive outcomes. For example, Sicher et al. conducted a study into distance healing using a population of people with advanced AIDS. In total, 40 participants were pair-matched by age, CD4 count, and number of AIDS-defining diseases (ADDs) before being randomly assigned to either the distance healing or control group. Four initial measurements were taken: CD4 count, psychological distress (measured using the Profile of Mood States), physical symptoms (measured using the Whaler Physical Symptoms Inventory), and quality life (measured using the Medical Outcomes Survey for HIV). These same measurements were also taken after the 10-week treatment period and 12–14 weeks later at the follow-up stage. During the study period, participants also reported doctor’s visits, hospitalization, illness recovery, and onset of new illnesses. Rather than working with traditional Christian groups, Sicher et al. recruited distance healing practitioners from different traditions or schools, but all with a minimum of five years regular ongoing healing practice, previous experience of distance healing with at least 10 patients, and previous experience of distance healing for patients with AIDS. Each practitioner treated five subjects for six hours in total (one hour daily for six days). Each participant received healing from 10 different practitioners. Sicher et al. found that during the six months of the study, patients in the treatment condition experienced significantly fewer doctor’s visits, hospitalizations, and new ADDs, as well as significantly shorter periods of hospitalization, significantly lower severity of illness, and significantly improved mood. However, no significant differences in physical symptoms or quality of life were found between the groups. Despite the marked differences in procedure (including the populations from which healers were drawn and the method by which healing was delivered), the positive findings have been regarded as a successful replication of Byrd (but see also Bronson for suggestions that the authors capitalized on data mining).

Some of this high-quality research has been summarized by Astin et al., who restricted their review to only those clinical studies that included random assignment of participants to conditions, a placebo-control condition, publication in full in a peer-reviewed journal, and use of participants who suffered from any medical condition [thus, excluding research involving direct mental interactions with living systems (DMILS) and staring detection studies such as those summarized by Schmidt et al., which reported significant effects of intention upon electrodermal activity in healthy participants]. Astin et al. identified 23 studies that met these criteria, collectively involving 2774 participants, which produced the predicted improvement in condition with a combined effect size of 0.40 (p < .001). Among these studies, 13 (57%) showed a positive treatment effect, nine showed no effect, and one showed a negative effect. Despite remaining concerns about the heterogeneity of the database and methodological limitations with some studies, the authors were able to conclude that the evidence was sufficiently strong to warrant further study.

A later review by Astin was restricted to prayer studies and consisted of 14 studies with a combined 2448 participants. These were mainly drawn from the earlier review (but with some additions, such as Abbott et al., 2001) and so does not provide much new information. Again, the outcome was positive, with six studies (43%) showing a positive treatment effect and the database generating an overall effect size of 0.30. This is somewhat lower than the effect size reported when studies of therapeutic touch are included, and other reviews have suggested that this approach may be of particular interest. It should be noted that Ernst also provided an update, consisting of 17 studies published after his 2000 review, and found that their outcomes “collectively…shift the weight of the evidence against the notion that distant healing is more than a placebo” (p. 241).

Rationale for the Current Study
Despite incorporating randomized control blinded studies, the studies included in the review by Astin et al. are still susceptible to counter explanations as a consequence of their inability to create an appropriate control condition (for
example, there can be no guarantee that control patients are not beneficiaries of healing intentions from friends, family, or their own religious groups). Additionally, putative relationships between healing intention and well-being might be obscured by reliance on relatively crude health outcomes (such as reduced depression scores) that themselves are open to influence from other mechanisms such as placebo and expectancy effects and are sensitive to other environmental and physical stressors that can vary over the course of a study. We would argue that related research that focuses on effects upon simpler biological systems than “whole humans” (such as growth of bacteria cultures, hemolysis of blood samples, or plant growth) would be less sensitive to the effects of such confounding variables and are likely to allow for “cleaner” control groups—plants seem unlikely to have expectancies concerning participation in a healing study, to have relatives sending them healing intentions, and to give rise to more straightforward and pre-specifiable well-being indicators. We therefore planned to conduct a quantitative review of healing studies that involve biological systems other than “whole” humans. Although some of this research has been reviewed previously (especially by Benor and Braud), these do not represent meta-analytic reviews and would benefit from the inclusion of more recent work. We proposed to blind-code such studies for methodological quality as well as other parameters so as to determine (i) whether there is an overall effect that cannot be explained in terms of Type I error, methodological flaws, or experimenter effects and (ii) whether effect sizes covary with other properties of the study design in a manner that might elucidate the mechanism of such effects.

Secondly, since Astin’s reviews have been published, serious doubts have been cast concerning the legitimacy of the work conducted by Daniel P. Wirth13,23 such that it would be unsafe to base conclusions on data that he has provided—Wirth contributed five studies to the review by Astin et al.,9 one study to Astin’s review,14 and five studies to Daley’s review.15 There is therefore a need to revisit these meta-analytic reviews of research with “whole” humans but with Wirth’s body of work removed. There has also been a “second wave” of replication attempts that have not been included in reviews to date; some of these have confirmed predictions24 but there are also some high-profile failures to replicate.27–29 To our knowledge, there has been no systematic meta-analytic review that has included these studies, and in our view, an updated and expanded review would be timely. Therefore, phase 2 of the current project consisted of a quantitative review of healing studies involving “whole” humans in a manner that addresses the shortcomings identified above.

METHOD

Identifying Qualifying Studies

A comprehensive literature search was conducted by one of the authors (C.S.) to identify studies of distant healing using the following databases: Swetswise, ASSIA, PsychNET, Web of Science, Cochrane Library, British Nursing Index, Cinahl Full Text, and Informaworld. Care was taken to ensure that nursing and medical journals were included in the search as well as those covering research in the social sciences. Search terms used were determined from a review of previous reviews and included the following: “Spiritual healing,” “Distance Healing,” “Noetic Healing,” “Intercessory Prayer,” “Laying on of hands,” “Therapeutic Touch,” “Johrei,” and “Reiki.” “Healing” was not used as a search term in order to avoid an excess of pharmaceutical research. For phase 1, this search was restricted by the inclusion of the qualifiers “Animals,” “Plants,” “Yeast,” “Bacteria,” and “Cells.” The articles resulting from these searches were then read by C.S., who decided which studies met the inclusion criteria outlined below. References lists of qualifying articles were then searched to identify further relevant studies, and this process was repeated until no new articles were identified. To minimize the file drawer effect, authors of included articles were also contacted to request details of any qualifying studies that were not listed in our database.

Inclusion/Exclusion Criteria

Only studies in English were eligible for inclusion in this review. All studies must have examined the effects upon a biological system of the explicit intention to improve the well-being of that target system. Thus, studies exploring the effects of intention upon physical systems or random number generators as their targets30,31 were excluded as were studies which looked at the effects of mental influence on movement of animals.32 Similarly, remote staring and DMILS studies were excluded on the grounds that they did not incorporate an intention to heal. The healing conducted must not involve direct touching, so as to be able to exclude the beneficial effects of contact/massage therapy.33–35 Articles that did not provide enough information concerning their methodology to allow for quality assessment were excluded as were studies that did not include a comparison condition (typically those involving only pre–post comparisons) and those that did not provide sufficient data to allow for an effect size calculation. To avoid systematic bias, where studies were reported as non-significant with no further statistical information, they were coded as having an effect size of zero.

Quality Assessment

In order to produce methodological quality assessments, C.S. produced versions of the method section for each qualifying study that excluded all information that might identify the researchers or give an indication of the study outcome. Each article was allocated a code number, and these numbers were randomized so that discrete studies in the same experimental series would not have consecutive code numbers and so would not be assessed one after another. Copies of these edited articles were provided in batches for judges in pdf format.

Three judges independently rated the studies for methodological quality using an adapted version of part 3 of the SIGN50 Methodology Checklist 2.36 The SIGN50

We would like to thank Sophie Drennan and Jacqueline Stone for their contribution as judges in phase 1 and Sophie Ridgway and David Saunders for their contribution as judges in phase 2. C.R. acted as a judge in both phases and was not involved in other aspects of the study until judging was completed.
scale was originally created in 2002 by the Scottish Intercollegiate Guidelines Network, which is responsible for producing evidence-based clinical practice guidelines for the Scottish National Health Service, as a tool to appraise the quality of clinical research. Judges are asked to rate the study along a number of dimensions using the following rating options: “well covered,” “adequately addressed,” “poorly addressed,” “not addressed,” “not reported,” and “not applicable.” The more rigorous the methods used, the higher the rating given for that item. For example, for item 1.2 “The assignment of subjects to treatment groups is randomised,” allocation by date of birth or by patient number were not considered to be true randomization processes, and so studies using such methods were given a rating of “poorly addressed”; randomization methods using hand-shuffled cards or hand-rolled dice, whilst somewhat more random are still subject to randomization methods using hand-shuffling; published tables of random numbers are rated as well covered.” In the original scale, the “not reported” option was defined as “mentioned, but insufficient detail to allow assessment to be made” and the “not addressed” option defined as “not mentioned, or indicates that this aspect of study design was ignored.” During the pilot phase, judges felt that it would make more sense if the definitions of these two items were swapped because ‘not reported’ suggested that that aspect of the methodology had been left out of the report altogether and that “not addressed” suggested that that aspect had been referred to but not effectively dealt with.

Following pilot work, some of the items were modified to tailor them to current needs. We removed item 1.3, “An adequate concealment method is used,” because concealment was covered under blinding (see below), and item 1.10, “Where the study is carried out at more than one site, results are comparable for all sites,” because it was not applicable. Item 1.4 “Subjects and investigators are kept ‘blind’ about treatment allocation” was expanded to give three separate items that better reflected levels of blinding: “subjects are kept blind about treatment allocation,” “investigators are kept blind about treatment allocation,” and “data analysts are kept blind about treatment allocation,” The following items were also added to the scale:

- “Controls in place for extraneous variables” (item 1.8).
- “Healers applied a consistent method of treatment” (item 1.9).
- “Rationale given for selection of healers” (item 1.10).
- “Controls in place to prevent Healers affecting participants/targets by conventional means” (item 1.12).

Judges were finally asked to give a rating out of 10 to represent the overall methodological quality of each study.

Calculating a Common Effect Size
The main outcome statistics were converted to the common effect size, $r^2$, by C.S. before judges’ quality ratings were used in identifying the amount of variance (e.g., in health outcomes) that can be explained by the intervention measure. It was collected so as to avoid any chance of bias influencing the conversions. Where no main outcome measure was identified, measures that were most similar to measures used in other studies were selected. If no such measure was utilized within a study, then the measure selected was the one which seemed most relevant to the condition being treated and which reported the most statistical information to allow for conversion, such as the number of participants in each group and the degrees of freedom. The statistics were converted by hand using formulae provided by Clark-Carter. Analyses were checked by C.R. once all judging had been completed.

**ANALYSIS (PHASE 1)**
Initially, 156 non-whole human sample studies were identified from 95 articles. Of these, 107 studies from 61 articles had to be eliminated from the meta-analysis as they were reviews of other studies, reported too little information, or did not fit with the above inclusion criteria. Thus, 49 studies from 34 articles were eligible for review. Effect sizes for these studies are illustrated in Figure 1. The combined effect size for the non-whole human studies weighted by sample size yielded an $r$ of 0.258 (CI$_{95}$ = 0.239–0.278), which is significant at the 5% level.

However, the effect sizes in the database are significantly heterogeneous ($\chi^2 (48) = 487.8$), and 10 "outliers" need to be cropped in order to reduce to non-significance, which reduces the weighted mean effect size for the cropped studies to $r = 0.204$, although this remains significant (CI$_{95}$ = 0.172–0.236).

**Effect Size and Quality Estimates**
In order to explore causes of variance in study outcomes, effect sizes were correlated against independent judges’ average ratings for various quality dimensions. This would evaluate the extent to which observed effects might be attributable to methodological flaws. Given the limited range for quality ratings, nonparametric correlations were calculated, and these are given in Table 1. A number of negative correlations can be observed that are consistent with an explanation in terms of methodological artifact; this association is significant for randomization method and suggestive for double blinding, control of extraneous variables, and clear specification of planned analyses. However, it should be noted that the average quality rating for these studies is low [mean = 4.3/10, standard deviation (SD) = 1.9] such that even relatively highly rated studies may still suffer from some methodological weaknesses.

To evaluate whether these weaknesses could account for the observed effects, we identified those studies that were rated as "well covered" or "adequately addressed" on all the following parameters: the assignment of subjects to treatment groups is randomized, investigators are kept "blind" about treatment allocation, the treatment and control groups are similar at the start of the trial, the only difference between the groups is the treatment under study, and values close to zero indicating no relationship. $r$ Values can be converted to $z$ scores using $r = z / \sqrt{N}$. 

---

*Footnote continued*
investigation, controls in place for extraneous variables, and controls in place to prevent healers affecting participants/targets by conventional means. The 22 studies that met these criteria gave a weighted effect size, \( r = 0.115 \), which remains significantly different from the null value of zero (CI95 = 0.090–0.141).

**Blocking Studies by Target System**

To further explore causes of variance in outcome, studies were blocked by target system type. Three categories had sufficient members for separate analysis: animal, plant, and in vitro studies. The largest category consisted of 22 in vitro studies (cell cultures and tissue samples). These gave a weighted mean effect size, \( r = 0.342 \) (CI95 = 0.319–0.363). The sample was significantly heterogeneous, \( \chi^2 = 271.19, p < .001 \), and 11 outliers had to be removed to reduce this to \( \chi^2 = 17.78, p > .05 \). The mean weighted effect size for the cropped sample reduces to \( r = 0.248 \) but remains significant (CI95 = 0.167–0.325).

Non-human animals (e.g., rats, mice, bush babies) were the subjects in 11 studies. These studies produced a significant weighted mean effect size of \( r = 0.277 \) (CI95 = 0.160–0.386). This sample was marginally heterogeneous, \( \chi^2 = 18.92, p < .05 \) (removing one outlier gives \( p > .05 \)). The mean weighted effect size for the cropped sample reduces slightly

---

**Table 1.** Spearman \( \rho \) Correlations Between Study Weighted Effect Sizes and Average Quality Ratings From Independent Judges for Non-Whole Human Studies

<table>
<thead>
<tr>
<th>Quality Criterion</th>
<th>( \rho )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>The assignment of subjects to treatment groups is randomized</td>
<td>-0.413</td>
<td>.004</td>
</tr>
<tr>
<td>Investigators are kept “blind” about treatment allocation</td>
<td>-0.281</td>
<td>.055</td>
</tr>
<tr>
<td>Data analysts are kept “blind” about treatment allocation</td>
<td>-0.239</td>
<td>.106</td>
</tr>
<tr>
<td>The treatment and control groups are similar at the start of the trial</td>
<td>-0.145</td>
<td>.332</td>
</tr>
<tr>
<td>The only difference between the groups is the treatment under investigation</td>
<td>0.083</td>
<td>.580</td>
</tr>
<tr>
<td>Controls in place for extraneous variables</td>
<td>-0.249</td>
<td>.092</td>
</tr>
<tr>
<td>Healers applied a consistent method of treatment</td>
<td>0.018</td>
<td>.904</td>
</tr>
<tr>
<td>Controls in place to prevent healers affecting participants/targets by conventional means</td>
<td>-0.161</td>
<td>.280</td>
</tr>
<tr>
<td>All relevant outcomes are measured in a standard, objective, valid, and reliable way</td>
<td>-0.129</td>
<td>.389</td>
</tr>
<tr>
<td>There is no scope within the design for optional stopping or otherwise capitalizing on chance variation in the outcome</td>
<td>-0.074</td>
<td>.620</td>
</tr>
<tr>
<td>Analyses are clearly pre-planned and corrected for multiple analyses where appropriate</td>
<td>-0.224</td>
<td>.099</td>
</tr>
<tr>
<td>Overall quality rating</td>
<td>-0.099</td>
<td>.507</td>
</tr>
</tbody>
</table>

---

Figure 1. Funnel plot of effect size by log \( N \) for non-whole human sample.
to \( r = 0.246 \) but again remains significant (CI\(_{95} \) = 0.123–0.361).

For 16 studies, the target systems were plants or seeds. These had a mean quality rating of just 3.22 and also gave a significant weighted mean effect size, \( r = 0.125 \) (CI\(_{95} \) = 0.098–0.153). This sample was also significantly heterogeneous, \( \chi^2 = 129.45, p < .001 \). Removing three outliers gives \( \chi^2 = 19.14, p > .05 \). The mean weighted effect size for the cropped sample increases to \( r = 0.197 \) and is significant (CI\(_{95} \) = 0.156–0.238). The weighted average effect sizes for the cropped in vitro and non-human animal studies falls outside this confidence interval, indicating that outcomes for the plant studies are significantly different.

**Publication Bias**

To evaluate whether the observed effect sizes might be affected by publication/availability bias, a funnel plot was constructed (Figure 1). Although the pattern is distorted somewhat by researchers’ greater tendency to give less statistical detail when outcomes were “non-significant” (in which case effect sizes were recorded here as zero to avoid loss of null data), it is clear that the plot is highly asymmetrical, with expected studies reporting null outcomes and (particularly) reversed effects being absent, which is suggestive of a publication bias. In order to estimate the number of unpublished non-significant studies that would be needed to render the database non-significant overall, Rosenthal’s failsafe \( N \) was calculated.\(^{23} \) This gives a value of 46,196\(^{\circ} \) where the critical number of studies is 240, suggesting that the file drawer effect alone cannot account for the observed results.

**ANALYSIS (PHASE 2)**

For the whole human meta-analysis, 182 studies were identified initially from 180 articles, of which 121 studies from 121 articles had to be eliminated from the meta-analysis as they were reviews of other studies, reported too little information, or did not fit with the above inclusion criteria, leaving 57 studies across 56 articles that were eligible for review. Weighted effect sizes were calculated and these are illustrated in Figure 2. When combined, these studies yielded a small effect size of \( r = 0.203 \) that was significant (CI\(_{95} \) = 0.180–0.232). As with the non-human meta-analysis, this database is significantly heterogeneous (\( \chi^2 = 754.7 \)); 11 outliers need to be removed to reduce this non-significance at \( p > .05 \). The mean weighted effect size for the cropped sample reduces slightly to \( r = 0.193 \) but remains significant (CI\(_{95} \) = 0.151–0.241).

**Effect Size and Quality Estimates**

Study outcomes for whole human studies were correlated against independent judges’ quality ratings, and these are given in Table 2. Of most concern is that judges’ overall ratings of study quality are negatively correlated with study outcome, suggesting that the observed effect might—at least in part—be attributable to methodological shortcomings (\( \rho = -0.253, p = .058 \)). Of the various quality dimensions, 11 of 13 also give negative correlations with study outcome, of which the strongest are suggestive associations with control of extraneous variables, rationale for healer selection, and explicit preplanning of primary analyses and a significant association with randomization.

In order to explore whether these factors could account for the observed effects, a subsample of methodologically superior studies was identified using the quality criteria described for phase 1. Of the original 57 studies, 27 met these threshold standards, giving a slightly larger weighted effect size, \( r = 0.224 \) (CI\(_{95} \) = 0.194–0.253).

**Blocking Studies by Healing Method**

Whole human studies could not be blocked by target system because the conditions being treated were too varied or were poorly specified. Instead, we categorized studies according to the reported healing method used, with four categories having sufficient members for separate analysis: intercessory prayer, therapeutic touch, Reiki or Johrei, and unspecified/other. The largest category consisted of 20 unspecified/other studies, which had a mean quality rating of 5.94 and gave a weighted mean effect size, \( r = 0.163 \) (CI\(_{95} \) = 0.105–0.219). The sample was significantly heterogeneous, \( \chi^2 = 57.34, p < .001 \); removal of three outliers reduces this to non-significance, with a mean effect size for the cropped sample that increases to \( r = 0.193 \) (CI\(_{95} \) = 0.115–0.267). Therapeutic touch was implemented in 19 studies (mean quality rating: 5.25), giving a weighted mean effect size, \( r = 0.371 \) (CI\(_{95} \) = 0.308–0.430). This sample was also significantly heterogeneous, \( \chi^2 = 217.58, p < .001 \); removal of three outliers reduces \( \chi^2 \) to 19.39, \( p > .05 \), giving a reduced effect size for the cropped sample of \( r = 0.203 \) (CI\(_{95} \) = 0.128–0.276). Overall, 11 studies incorporated intercessory prayer, giving the smallest weighted mean effect size, \( r = 0.173 \) (CI\(_{95} \) = 0.141–0.201). This sample was also significantly heterogeneous, \( \chi^2 = 446.47, p < .001 \); five outliers need to be removed to reduce this to non-significance, and the cropped sample gives a reduced effect size that remains significant, \( r = 0.138 \) (CI\(_{95} \) = 0.041–0.233). The smallest category consisted of seven Reiki or Johrei studies, which gave a weighted mean effect size, \( r = 0.320 \) (CI\(_{95} \) = 0.187–0.442). This sample was also significantly heterogeneous, \( \chi^2 = 33.36, p < .001 \), removal of one outlier reduces \( \chi^2 \) to 8.85, \( p > .05 \), and results in a reduced effect size, \( r = 0.224 \) (CI\(_{95} \) = 0.077–0.362).

**Publication Bias**

A more extreme pattern is evident in the funnel plot for whole human studies (Figure 2) than we saw for the meta-analysis of non-whole human studies (Figure 1), with the distribution affected by there being no null or reversed studies that reported effect size outcomes. Nevertheless, it is clear that the plot is suggestive of a publication/availability bias. Rosenthal’s failsafe \( N \) gave a value of 103,497 unpublished null studies needed to reduce the effect to non-significance where the critical number of studies is 255, again suggesting that the file drawer effect alone cannot account for the observed results.
DISCUSSION
We proposed to conduct a meta-analysis of distant healing studies that involved non-whole human target systems in order to ensure a clearer distinction between active and control conditions, given that studies involving patients as subjects are likely to involve “control” subjects who still benefit from the healing intentions from friends, family, or their own religious groups. We also had concerns about the inability to control for the effects of placebo and expectancy, since participants in control conditions tend to presume that they are in the active condition and so might experience placebo improvements in a manner that tends to reduce the difference between active and control conditions. These concerns can be addressed by the use of animal and tissue samples that presumably do not have expectancies about the effects of treatment or have communities of peers sending them positive intentions for their well-being. The combined weighted effect size for the non-whole human studies gave a weighted $r$ for the heterogeneous sample of 0.204, which indicates that those allocated to active healing conditions achieved better well-being outcomes than did those allocated to comparison conditions. Interpretation of this highly significant effect is not straightforward given that overall quality ratings were relatively low (mean = 4.3/10), and study quality criteria were not met.

Table 2. Spearman $\rho$ Correlations Between Study Weighted Effect Sizes and Average Quality Ratings From Independent Judges for Non-Whole Human Studies

<table>
<thead>
<tr>
<th>Quality Criterion</th>
<th>$\rho$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>The assignment of subjects to treatment groups is randomized</td>
<td>-0.330</td>
<td>.012</td>
</tr>
<tr>
<td>Participants kept “blind” about treatment allocation</td>
<td>-0.150</td>
<td>.264</td>
</tr>
<tr>
<td>Investigators are kept “blind” about treatment allocation</td>
<td>-0.078</td>
<td>.566</td>
</tr>
<tr>
<td>Data analysts are kept “blind” about treatment allocation</td>
<td>-0.165</td>
<td>.221</td>
</tr>
<tr>
<td>The treatment and control groups are similar at the start of the trial</td>
<td>-0.151</td>
<td>.262</td>
</tr>
<tr>
<td>The only difference between the groups is the treatment under investigation</td>
<td>-0.092</td>
<td>.498</td>
</tr>
<tr>
<td>Controls in place for extraneous variables</td>
<td>-0.233</td>
<td>.081</td>
</tr>
<tr>
<td>Healers applied a consistent method of treatment</td>
<td>0.017</td>
<td>.902</td>
</tr>
<tr>
<td>Rationale given for selection of healers</td>
<td>-0.238</td>
<td>.075</td>
</tr>
<tr>
<td>Controls in place to prevent healers affecting participants/targets by conventional means</td>
<td>-0.187</td>
<td>.164</td>
</tr>
<tr>
<td>All relevant outcomes are measured in a standard, objective, valid, and reliable way</td>
<td>0.089</td>
<td>.508</td>
</tr>
<tr>
<td>There is no scope within the design for optional stopping or otherwise capitalizing on chance variation in the outcome</td>
<td>-0.173</td>
<td>.198</td>
</tr>
<tr>
<td>Analyses are clearly pre-planned and correct for multiple analyses where appropriate</td>
<td>-0.233</td>
<td>.081</td>
</tr>
<tr>
<td>Overall quality rating</td>
<td>-0.253</td>
<td>.058</td>
</tr>
</tbody>
</table>
outcomes were significantly correlated with judges’ ratings of the quality dimension of randomization, and suggestively so with investigator blinding, control of extraneous variables, and preplanning of reported analyses. However, when analysis was restricted to those studies that were rated as “well covered” or “adequately addressed” for key quality dimensions, the subsequent database still gave a significant weighted effect, $r = 0.115$ ($CI_{95} = 0.090–0.141$). This suggests to us that further research is warranted but that research must meet methodological quality standards, particularly for aspects identified in Table 1.

We also had concerns that previous meta-analytic reviews had included work conducted by Daniel P. Wirth that has since been discredited and so we conducted an updated whole human analysis that omitted these studies and also included more recent publications. The resulting combined effect size for the homogeneous sample was small, with $r = 0.203$, but significant ($CI_{95} = 0.180–0.232$). As with the Phase 1 analysis, there are quality issues here with respect to investigator blinding, control of extraneous variables, and preplanning of reported analyses, but again, when these are addressed by selecting only those studies that are rated as “well covered” or “adequately addressed” with respect to key quality dimensions, the surviving studies still give rise to a significant weighted effect size, $r = 0.115$ ($CI_{95} = 0.090–0.141$).

Both databases included blocking of studies by type, and this suggested that some approaches had been more successful than others. For non-whole human research, similar effects were observed for non-human animals and in vitro samples, with the effect for plant studies being significantly lower. This may be a function of the complexity of the target system to be affected. For whole human studies, the largest effects were associated with Reiki and Johrei interventions followed by therapeutic touch, then unspecified healing, although effects were relatively similar. The outcome for prayer studies was somewhat (though not significantly) lower, giving the smallest effect size for any subsample. This is consistent with Astin’s earlier summary that found that prayer studies were less successful than therapeutic touch studies; it also reflects recent large-scale failures to capture effects of distant prayer.

It is possible that more proximal noncontact healing studies still afford some opportunities for blinds to be broken so that beneficial effects could be attributed to expectancy (although some studies are very impressive in the lengths to which they go in order to ensure that sham treatments are indistinguishable from active treatments). Alternatively, we might argue that prayer studies could take more care in ensuring that the healers they recruit constitute a homogeneous group that reliably applies a consistent method of healing; often the prayer groups are quite eclectic and little effort is made to ensure that standard practices are adhered to (Jonas has made similar observations).

It remains difficult to draw unequivocal conclusions based on this analysis because of the clear implication from the funnel plots that there exist missing studies. This combined with some associations between outcomes and quality parameters blunts our confidence that we are describing genuine noncontact healing effects. This will not be resolved by reanalysis and debate but rather by the execution and publication of further randomized controlled trials that explore this putative phenomenon. Findings are, in our view, sufficiently promising to justify that effort, and we would encourage colleagues to conduct such replications. With the design of those replications in mind, we make the following recommendations:

- Have a clearly circumscribed heelee population with explicit inclusion/exclusion criteria
- If heelees are randomly allocated to conditions rather than matched for potential confounds (such as co-morbid conditions), then researchers should pre-measure and report any significant differences in demographic data that could impact on the illness or its treatment
- All personnel who interact with heelees must be blind to condition allocation
- Researchers should state explicit criteria for the appointment of healers and intercessors that is related to the target population/illness (i.e., they should have experience of working successfully with that condition or should be able to show that previous success should generalize to the current situation)
- Homogeneity of approach across healers should be ensured through the production of explicit instruction and some attempt made to verify that this is adhered to
- Researchers should ensure that instruction given to healers regarding desired outcomes reflects the well-being factors that are actually measured in the course of the study
- Researchers should ensure that actors in the sham condition closely mimic behaviors used by healers in the active condition but precluding “inadvertent” healing effects by using actors who have no prior healing experience and who are prevented from developing positive thoughts toward the helee (for example, by having them complete mental arithmetic tasks). Interactions should be recorded to enable checks for perceptible differences between experimental and sham conditions
- Clear descriptions should be given of precautions to prevent normal communication with patients that could affect blindness, and interactions should be monitored to ensure no facility for normal communication
- Primary outcomes should be pre-specified; where multiple dependent measures are taken, these should be reported in the form of an appropriate omnibus test (e.g., MANOVA and multiple linear regression) before individual variable tests to avoid concerns over “cherry picking”
- Statistic effect sizes should be reported as well as $p$-values, and some indication given that study sizes have been designed to have sufficient power to detect the putative effect
- We had a poor response to requests for information about unpublished studies and so these are likely to be under-represented in this analysis. We would recommend that a repository is established and that researchers are encouraged to register studies with it at the design stage.
Acknowledgments

The authors would like to thank the Confederation of Healing Organisations for their kind support of this project.

APPENDIX 1. NON-WHOLE HUMAN DATABASE


**APPENDIX 2. WHOLE HUMAN DATABASE**


REFERENCES


