Available online at www.sciencedirect.com
ScienceDirect
The Veterinary Journal 180 (2009) 141–148
The Veterinary Journal
www.elsevier.com/locate/tvjl

Personal View

Homeopathy and the curse of the scientific method

Karen L. Overall [a,*], Arthur E. Dunham [b]

[a] *Center for Neurobiology and Behavior, Psychiatry Department, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[b] *Biology Department, University of Pennsylvania, Philadelphia, PA, USA*

Accepted 12 October 2008

In a recent article, published in *The Veterinary Journal*, Nina Cracknell and Daniel Mills of the Department of Agricultural Sciences at the University of Lincoln tested the effect of a homeopathic remedy for the treatment of fear associated with the noise of fireworks (Cracknell and Mills, 2008). They found that the placebo group responded as well as the homeopathic/treatment group when changes in the number of behavioral signs remaining were compared statistically.

Homeopathy is experiencing a resurgence of interest and is receiving attention in the field of veterinary behavioral medicine largely because of the perception that homeopathic treatments have no side effects whereas traditional treatments are often perceived to have unacceptable side effects, a justification noted by Cracknell and Mills (2008). Such perceptions appear to be seldom questioned, yet if we are to move veterinary behavioral medicine past the realms of observation and evidence-based medicine and into the realm of tests of mechanisms and hypotheses we must question our approach to such issues. This article provides a commentary on what a review of homeopathy can teach us about the thought process involved in assessment and treatment in veterinary behavioral medicine.

Issues affecting experimental design

There are two issues pertaining to the design of any treatment or group comparison, especially in behavior, namely conceptual issues and methodological issues. In veterinary behavior and behavioral medicine the latter often involves measuring things you cannot see or count, especially at the mechanistic level.

Conceptual issues

Conceptual issues address concerns like the effect of merely doing (or participating in) a study. Here, we must consider a version of the Heisenberg effect (where one changes something by measuring it) and population issues, including bias. The way patients are selected is critical, especially in studies where responses to putative behavioral treatments are the focus (Brown, 2006).

We seldom evaluate patients' views (in the case of human treatment studies) or caretakers' views (in the case of non-human treatment studies) about certain treatment classes, but it is likely that no one volunteers for something for which they do not 'believe' there is a high-probability of a positive effective. While treatment groups may be 'randomized' with respect to the pool of those solicited for or attracted to the study, such randomization is unlikely to include those who do not wish to use medication in pharmacological studies, or those who view homeopathy as 'rubbish', in the case of a study on a homeopathic treatment. The same logic pertains to the perception of side effects: if one 'believes' that a particular treatment has more side effects than another, one's probability of participating in a study would be affected by the perception of the risk thought to be posed by the side effects. In this case it is the perception of risk – not the measurement of it – that biases the participating population.

Because veterinary researchers cannot directly assess patients' views about their conditions, the randomization issue becomes a critical factor when we evaluate questions pertaining to the assessment of patients. This is especially true if the clients (the same individuals we solicited for the study) are involved in any way and at any level in the evaluation of the response, as is almost always the case with veterinary behavioral studies (King et al., 2000, 2004; Crowell-Davis et al., 2003; Cracknell and Mills, 2008) where the signs are non-specific and may not be obvious to clinicians examining the patients. In other words, in some populations of our clients and their animals, interpretation of the problem and the response to the putative treatment may depend on the client's belief rather than on an objective evaluation of the patient's

[*] Corresponding author.
E-mail address: overallk@mail.med.upenn.edu (K.L. Overall).

1090-0233/$ - see front matter
doi:10.1016/j.tvjl.2008.10.006

response to a particular treatment (Bausell, 2007). This is why – from the conceptual viewpoint – randomized, placebo-controlled trials are essential and may be the most likely explanation why the study reported by Cracknell and Mills (2008) generated completely different results from an earlier, open label, non-randomized, non-placebo-controlled trial conducted by the same group (E.D. Levine and D.S. Mills, unpublished data, cited in Cracknell and Mills, 2008).

Methodological issues

Methodological issues include whether studies use placebos and, if so, how. Placebos can be defined (OED online, 2008) '*as a drug, medicine, therapy, etc., prescribed more for the psychological benefit to the patient of being given treatment than for any direct physiological effect, especially one with no specific therapeutic effect on a patient's condition, but believed by the patient to be therapeutic (and sometimes therefore effective).*' We who engage in such treatments and experiments should restrict the placebo definition to '*a substance with no (known) therapeutic effect used as a control in testing new drugs, etc.*' (OED online, 2008), which then functions as a blank sample in a test. The placebo effect is '*the beneficial (or occasionally adverse) effect on health produced by a placebo that cannot be attributed to the properties of the placebo*' (OED online, 2008). It is often defined as a measurable, observable, or felt improvement in health or behavior that is not attributable to the true treatment administered.

Placebo controls are essential because the patient (in human medicine), and the pet owner (in veterinary medicine) know that they are participating in an experiment or a treatment study. Because of regulations governing informed consent in experimental medical science, there is no way to test the efficacy of a putative treatment compared to that of a placebo without either the patient or the client, respectively, being informed as to the nature of the experiment and providing a written consent to participation in the test.

Accordingly, it should not surprise us that placebos are very effective in treating pain, and that there is a postulated mechanism for this finding. Because the perception of pain is dependent on a series of complex neurochemical interactions, some of which can be modulated by thought processes, pain can be modulated by the belief that the treatment will help. However, the mechanisms by which this occurs differ. For example, medications may block the production of prostaglandins involved in the pain cascade, but belief in success may release endorphins which block the effect of the prostaglandins (Park, 2000; Hribjartsson and Gotzche, 2001; Bausell, 2007). From the viewpoint of the assessment of the clinical 'phenotype' – i.e. how the patient is doing – the outcome may be the same (Park, 2000).

It is important to remember that a study designed to evaluate phenotypic variation, for example, is not the same as an interventional study and may provide excellent information which can be completely unaffected by the 'beliefs' of the clients. In an interventional experimental design, a placebo treatment is included to control for the feeling of achieving or receiving a beneficial result that is due to the patient's (or, for animals, the owner's or caregiver's) belief that the study in which they are participating will have a positive effect. The placebo is given to control for any effect on the outcome of a study or trial of either participation in or anticipation of the outcome of participation.

However, two things are critical. Firstly, the placebo used should have no known effect on the pathology and there should be no known mechanism whereby the placebo could have such an effect. Secondly, it should be recognized that, because both treatment groups (i.e. the group receiving the placebo and the group receiving the putative treatment) are subject to the placebo effect, the essential statistical test is one that has the power to detect a difference between the effect measured on or reported by the group receiving the placebo and that receiving the putative treatment.

What is seldom appreciated or articulated is that, when accurately measured, the placebo effect is a valid measure of the variance of the effect of the trial or the study itself, regardless of the placebo's focus or hypothesized mechanism. The importance of including a placebo treatment in an experiment is that the existence and magnitude of a statistically significant difference between the response of the treatment group and that of the placebo (control) group in an experiment provides evidence for the efficacy of the putative treatment. In short, we have a way to assess our intervention, but seldom interpret our results, in this light. We need to begin to do so, because when signs are non-specific, not equally noted, and may be affected by clients' interpretations, we need to evaluate the effect that these factors have on our findings.

The application of the rules governing experimental design in medical science in general and to homeopathy in particular

In any study we should want to know firstly, is there a 'statistically significant' difference between the effects of the treatment and the placebo? Drug trials differ from other types of case-control experiments in this regard because they are interventional studies where the mechanism of change is linked to the mechanism by which the intervention is thought to work. In other words, you may be assessing effects on symptoms or signs, and not the underlying pathology. We also need to know if the difference in the direction is implied by a putative mechanism.

In determining whether the difference is large enough to matter, we need to recognize that statistical significance itself is not enough for us to be comfortable with the detection of a phenomenon. The magnitude of the statistical effect tells you about your design, not about the importance of a biological phenomenon; the design must be evaluated in terms of the probability of detecting a treatment effect of a given magnitude (statistical power) and the likelihood that the finding is dependent on sample size, for example. We seldom see reports of the latter, but we often see some attribution of certainty assigned to a certain given probability level without any understanding that a probability of 0.0002 is not equivalent to a probability of 0.05, which is not equivalent to a probability of 0.1. All of these could be considered 'significant', but an evaluation of this statistic must take into account the relationship between the design of the experiment and the underlying distribution of the data – a feature that is seldom discussed in any study reporting 'significance'.

Often, however, the significance is sufficient and believable, but we then fail to ask whether the magnitude of the experimental treatment effect is sufficient to be biologically meaningful. Merely having a statistically significant effect does not guarantee that the significant finding has any bearing on the actual focus of the study. A rare statistical association may have no effect on the biology of the system, but could misdirect incautious readers to believe that this was not the case. If the magnitude of the experimental effect is sufficiently large that it is likely to be biologically meaningful, then a hypothesis of the mechanism whereby the treatment effect is produced is required for completeness. Testable hypotheses of mechanisms of action are almost uniformly missing in any studies of complementary and alternative medicine (CAM), and indeed in many veterinary behavioral studies.

Such discussion about what we actually know is limited by what you can actually measure when seeking to evaluate an intervention. This is not as simple as asking whether something is the 'same or different' because how it is the same or different may matter in interventional studies (see above). For example, in evaluating the differential responses to a provocative noise you could do a case-control study and (1) ask about overall differences (presence/absence, frequencies), (2) look at different sequences of behaviors (one might affect the other), and (3) use a randomization model to evaluate whether the patterns in each of these steps and overall effect differs from that which would be expected if the intervention has no effect. With the latter you can look at where the actual test statistic values fall compared to those generated from repeated iteration of the randomization model in terms of magnitude of effect. This is a very useful, but underutilized technique for evaluating behavioral changes in populations of relatively small sample sizes (Overall et al., 2001).

The two major issues about which we should have statistical concern are (1) the distribution of effect sizes or the magnitude of the effect, and (2) the distribution of the test statistic values. The power of a statistical test $(1 - \beta)$ is not insignificant. When you evaluate power (the probability of making a type II error or accepting the hypothesis when it is false), you are asking what sample size is needed to actually detect an effect of a given magnitude; power should be calculated during the design of the experiment. If your experiment is not capable of detecting an effect of a particular magnitude with the data you have, any real effects smaller than that magnitude will likely be missed. In Cracknell and Mills (2008), an a posteriori calculation was made for power of the test, and it was used to confirm that there was no effect of the homeopathic treatment. Had an effect been found, however, the likelihood that someone would have checked for the effect of sample size would have been diminishingly small, again emphasizing the importance of good experimental design at the outset of the clinical trial.

Establishment of the validity of a homeopathic method requires (1) that it be held to the same standards for validity of non-homeopathic methods, (2) that anyone with a vested interest in the outcome must not be involved in conducting the trials (in part because of the problems associated with evaluation of effects), and (3) that the methodological issues discussed above should obtain in every trial: the trial must be blinded, must evaluate the placebo effect, and must evaluate the relative magnitude of any effect (Bausell, 2007).

Unless you can write down a putative hypothesis with a mechanism that can be tested scientifically, there is no point in going further. This point may be the one that distinguishes high-dilution homeopathic preparations from many other 'herbal' or 'natural' preparations that are often wrongly grouped under homeopathy. The 'herbal' and 'natural' types of interventions are more appropriately part of CAM. Virtually all commercially produced pharmaceuticals have their source and/or intellectual roots in active compounds isolated from plants. It is ironic that 'herbal' sources themselves are often considered 'safe', but that commercially produced derivatives – the compounds that are forced to undergo toxicity testing – are not. The finding that 1/5 Ayurvedic medicines purchased on the Internet have detectable and often toxic levels of lead, mercury and arsenic (Saper et al., 2008) should give those who think CAM interventions can do no harm serious pause.

High-dilution homeopathic preparations, like the one tested in Cracknell and Mills (2008), involve dilution to the point where it is very unlikely that even a single molecule of the base compound is present. This means that there is no conceivable mechanism of action by any currently accepted scientific standards, including those of chemistry and physics which govern serial dilutions.

Specific issues about related studies raised by Cracknell and Mills (2008)

Behavioral signs can be non-specific and are not equally appreciated by all who evaluate them; yet clients are often asked to evaluate changes in non-specific signs. Accordingly, not all behavioral measures are accurately assayed. For example, the non-specific signs of destruction, escape, and elimination are easier for clients to evaluate (and more of a problem for the client) than are the non-specific signs of freezing, salivating, and panting. Behavioral studies that seek to use client evaluations need an assessment of how readily people can assay the behaviors of import. In other words, we must ask not only how much can we trust clients' evaluations but also how much variance is there in how clients evaluate their pets? The latter issue can be dealt with by using repeated assessments and a repeated measures design, which allow pets to be evaluated as individuals that change across time, but not in a comparative study on variation of signs among the patients.

Of the 15 signs of interest in the firework study by Cracknell and Mills (2008), only two or three ('destructiveness', 'elimination', and possibly 'self-harm', depending on the social environment) could be recognized by clients who did not witness the events. If the dog was able to leave the home, 'bolting' could also be assessed. This means that for the vast majority of signs used to evaluate the intensity of the condition and its putative response to any treatment, no accurate independent assessment of the validity of the clients' reports is possible.

Such effects on assessment of signs can be controlled for by thinking about the conceptual issues involved in the study design. In a study on the effects of clomipramine on behaviors shown by animals affected with separation anxiety, the signs assessed were restricted to those everyone could see and evaluate (King et al., 2000, 2004), although these were not a complete compendium of all of the signs evinced by dogs with separation anxiety. In King et al. (2000), the effect of clients knowing that they participated in a study involving a medication and a placebo was modulated by a set of instructions affecting their response to their pets that were very similar to those used by Cracknell and Mills (2008). Simply, the clients were educated about what it means to accidentally or deliberately reward/reinforce anxious behaviors and were asked not to do so. To help clients comply with this request, they were asked to make the dog wait until it was sitting and calm before giving it any attention, and were tutored in how to do this. Such instructions have the advantage of allowing and encouraging – but not ensuring – cooperation from all clients. As was the case for Cracknell and Mills (2008), in the King et al. study (2000) the 'placebo' was actually a form of passive behavior modification designed to allow people who wished to (and were going to) do something that would still allow evaluation of the effect of their intervention.

There are still two conceptual flaws with such plans and these affect all placebo studies. Firstly, people lie. Secondly, even if we assume that our clients are all virtuous, there is an additional, more insidious problem that serves to emphasize how important understanding mechanism is for our patients. We cannot separate the effect of doing something passive – which might change the dog's behavior – from becoming more watchful, attentive, informed etc. – which might affect how you interpret the dog's behavior. A change in interpretation may or may not lead to a change in how the client interacts with the dog, but we almost never evaluate such effects. In a placebo group in veterinary behavioral medicine, we actually seek to control human behavior towards the animal. This does not address the issue of whether we have differential population biases within our treatment groups, but instead is an attempt to level the effects of some of the potential biases.

How and where we assess behaviors are also issues. Cracknell and Mills (2008) noted that one study on the use of behavior modification combined with two medications, clomipramine and alprazolam, for the treatment of storm phobia had reported improved outcomes in client evaluations of the dogs' behaviors, but did not show an effect in behavior ratings of signs from videotaped recordings (Crowell-Davis et al., 2003).

One of the problems with many behavioral studies involves methodology that is not comparable between studies or between populations of those studied. In the Crowell-Davis et al. (2003) study, some of the assessment documentation is available only from the author and these documents expand on the idea of scaled effects, as used by Cracknell and Mills (2008), to incorporate behavioral logs and questionnaires assessing specific aspects of frequency and intensity as part of the caregivers' global assessment. In such cases we actually obtain better information on actual behaviors by removing focus from clients' interpretations of such behaviors. The situation under which behaviors are evaluated is not unimportant.

In the Crowell-Davis et al. (2003) study, the responses of storm phobic dogs to a recording of three successively more intensive storms were videotaped. All videotaping occurred at the veterinary hospital and, despite clinician and client assessments of improvement in treatment, no differences in anxiety-associated event variables (lip-licking, yawning) or state variables (whining, hiding, panting and trembling) were noted. Unfortunately, no in-clinic videotaping of dogs that were not storm phobic was done, nor were the phobic dogs videotaped at home before and after treatment. This is only an issue because the effect of being in a university veterinary teaching hospital could not be evaluated in this study.

Cracknell and Mills (2008) commented that there was '*no change in behaviour ratings from videotape*' (p. 86). This conclusion must be evaluated in the context of the above constraint of the potential effect of the location in which the assessment occurs. In one study on vocalizations of dogs which were non-remarkable and those affected with separation anxiety, videotapes of both groups in both the home and hospital setting indicated that there was a large effect of treatment on vocalization at home, but that the same magnitude of effect was not found in the veterinary hospital, where, after treatment, affected dogs vocalized to the same degree as did non-remarkable dogs (Overall et al., 1999a, 1999b). All dogs barked and, or vocalized while in the stressful environment of a veterinary hospital, making interpretation complex.

These examples show the importance of transparency in design, and emphasize that the more objective the evaluation of the behavior is the less one has to rely on client impressions. In such cases, not only are the data more reliable, but the design can then be used and/or replicated by other researchers. If we are interested in beginning to elucidate and understand underlying mechanisms of behavioral problems this is an essential step.

Review of the findings of Cracknell and Mills (2008) for this study and other types of non-traditional treatments

As noted by Cracknell and Mills (2008), there have been few peer-reviewed studies on homeopathy and even fewer that utilize a placebo control. A few words on peer-review and on a technique commonly used to compare many of the homeopathic studies in human medicine, meta-analyses, are warranted here.

Peer-review

Peer-review guarantees that prior to (and as a pre-requisite to) publication, a paper will be read and evaluated by someone who is supposedly knowledgeable about the field, but successful publication does not mean that the paper is 'good' or 'right'. All journals publish papers with a range of quality. One would hope never to have to claim that any journal has published other papers as bad as one's own.

Glaring errors and falsehoods are usually – or hopefully – caught in the peer-review process, but not always (DeAngelis and Fontanarosa, 2008; Ross et al., 2008). But reviewers do not actually repeat the experiments to validate the outcomes and so we cannot expect publication to guarantee 'truth'. There is one notable exception to this comment on replication, and it involves a homeopathic study.

In 1988, *Nature* published a paper involving a study of a homeopathic approach using a high-dilution anti-IgE compound that supported the effect of this compound as measured by basophil degranulation (Davenas et al., 1988). Because of the uproar caused by this paper – which was hailed as one of the first and best proof of the principles of homeopathy – three investigators sought, with the original researchers' help, to replicate the study. The results of their replication were also published in *Nature* (Maddox et al., 1988) and could serve as a tutorial in what has been lacking even in the 'better' studies on homeopathic interventions. The authors found no support for any effect of the high-dilution anti-IgE and summarized their conclusions in five main points: (1) the care with which the experiments were carried out did not match the extraordinary character of the authors' claims; (2) the described phenomena were not reproducible and the originating laboratory had undertaken no investigation of the reasons; (3) the data actually lacked errors that would be expected in any set of measurement (e.g. random errors) – in essence, the data were too clean and too 'good'; (4) no real attempt had been made to eliminate systematic errors (as opposed to the random errors mentioned in point 3); and (5) the climate of the laboratory did not support objective evaluation of the exceptional data (Maddox et al, 1988).

At issue here is what these authors have chosen to call '*exceptional data*'. In high-dilution homeopathic solutions, the originally present compound is diluted to the point that it is highly likely that not a single molecule still exists. If such a solution can be shown to have any effect, the effect must be free of known and routinely used and referenced standards and mechanisms. In other words, in the absence of a putative mechanism, which can be explored and tested, any data would have to be exceptionally carefully collected and assessed to be believable.

We mention this because homeopathic interventions are not the only types of putative treatments that do not meet these standards for 'exceptional data'. Bausell (2007) stated that all CAM interventions fail to meet such standards. In veterinary medicine we have to look no further than many commonly used behavioral 'treatments', including some herbal supplements and pheromone products. In fact, Cracknell and Mills (2008) reference a non-placebo-controlled study from their own laboratory on treatment of fear of firework noises using the pheromonal product DAP, or dog appeasing pheromone (Sheppard and Mills, 2003), which found an improvement in signs similar to those they report for the non-placebo-controlled version of this study (E.D. Levine and D.S. Mills, unpublished data; cited in Cracknell and Mills, 2008). One has to wonder whether this 'effect', too, would vanish, were the intervention subjected to the rigors of a replicated, placebo-controlled, double-blinded study.

Were one to Google pheromonal interventions for animals one would find tens of thousands of 'hits', yet only a handful of carefully conducted studies have been published, and most of those are not placebo-controlled. In the face of the extraordinary label claims of many products, including the pheromonal ones, we have to understand the insidious and penetrating effect the 'folklore' (Maddox et al., 1988) has on any true assessment or evaluation of any effects, or the ability to obtain such assessments and evaluations. This problem is a property of unclear and undefined mechanisms of action which allow proponents to justify any findings as a set of special cases. In case anyone is still wondering, that is not science.

Meta-analyses

A number of reviews have appeared which tabulate the results of homeopathic studies (Linde et al., 1999; Linde and Melchart, 1999; Ernst, 2002; see Bausell, 2007 for other studies). Such tabulations are informative, but a simple show-of-hands approach to evaluate the effectiveness of an approach is inadequate. It matters little how many poorly designed (and thus unreliable) studies show positive effects of a homeopathic treatment. However, if a number of well designed, placebo-controlled, double-blinded experiments have shown positive results, then one gains confidence that the effect is reliable. Even in these cases, a simple show-of-hands approach lacks statistical power and rigor.

There is an established statistical methodology (meta-analysis) for combining the results of several studies into a single analysis. Meta-analysis is used extensively in epidemiology and evidence-based medicine and in many other non-medical disciplines, and there is an extensive literature on its use (Whitehead, 2002). This can be a powerful tool, but, as noted by Bausell (2007), what has been done with CAM studies, including those on homeopathy, are neither true meta-analyses nor they 'syntheses'-they are summaries, more commonly known as a show-of-hands. Science, done well and carefully, removes the need to vote on whether you think a method or treatment works.

So, what do we really know about homeopathy and other forms of CAM? In contrast to the sympathetic wording found in many papers evaluating homeopathy, there is no cause for sympathy. The clearest statement yet of the analysis of available data and the re-analysis and review of studies purported to examine other studies is found in Bausell (2007): no CAM therapy – including homeopathy – has a scientifically plausible, biochemical mechanism of action beyond that of the placebo effect.

Recommendations how behavioral data might best be measured and why choosing the appropriate measurement is important

So where do we go from here? Collection of data for behavioral and behavioral medicine studies is not trivial. The following methods for the assessment of behavioral data are designed to show what is needed to move towards sufficiently rigorous data collection that will permit mechanistic hypothesis formulation and testing, and allow us to believe the outcome of that process.

Method 1. Ask about people's impressions of behavior

This is not an optimal technique, but the quality of the data obtained very much depends on the quality of questions asked. Questionnaires absolutely must be rid of the need for those completing the questionnaire to make judgments and draw conclusions (Serpell and Hsu, 2001), including those involved in diagnosis, and people must be given the option of stating when they do not know something (see Overall et al., 2006 for one example). In the Cracknell and Mills (2008) study, the phrasing violates this tenet: '*owners stated that they felt their dog's fear...*'. It would be preferable to ask about actual events (e.g. '*an unknown, leashed dog approaches your leashed dog*') and some objective measure of them (e.g. '*does your dog back away or withdraw, does he growl, does he tuck his tail?*') (Overall et al., 2006). The same standard must be applied to evaluate outcomes. In the Cracknell and Mills (2008) study, clients were asked about their level of 'satisfaction' but not about their viewpoint on, for example, homeopathy, which could have provided some much needed information about biases.

Method 2. Ask questions about specific behaviors and their intensities/frequencies

This must be done in a way that assures that every observer means the same thing (e.g. '*does the dog growl in circumstance x, how often does he encounter circumstance x, and what proportion of the time does he react by growling? By 'growl' we mean the following....*') (Overall et al., 2006). This is an improvement on the first approach, but it works best only if the findings can be corroborated (e.g., video, a log etc.) and when the definitions of the behaviors are clear and can be queried when uncertain. The latter can be accompanied by reviewing the questionnaire with the clients using a standard set of definitions.

Method 3. Have people keep a log of Method 2?

Logs or diaries provide more objective data, if people are schooled in the type of data that they are to collect (e.g. time at beginning and end of behavior, behaviors exhibited, weather conditions etc.).

Method 4. Observe behaviors, record and do ethological an analysis

This method is better and can be improved upon by using tick sheets with objective categories and definitions of those categories (Martin and Bateson, 1986), but its efficacy is limited by human recording abilities.

Method 5. Video record behaviors and do a complete ethological analysis

This method is best if it is important to know which behaviors are used, in what combinations, for how long, and with what specific manifestation, and for those who wish to have the subsequent ability to formulate hypotheses about the behaviors.

Method 6. Look at co-varying patterns of behaviors and physiology

Behaviors do not occur in vacuums and co-occurrence of physiological changes may allow a more complete understanding of patterns of behaviors and putative underlying mechanisms (Overall, 2005). This expanded ability improves on Methods 2–5.

Method 7. Subject each dog to the same provocative test

A comparison of responses or outcomes can be done by subjecting each dog to the same provocative event (e.g. a handedness assessment in the sense of Branson and Rogers, 2006; a physiological test in the sense of Overall et al., 1999a, 1999b), whether it involves a disruption in their sensory environment (Crowell-Davis et al., 2003) or a putative test that asks about usage or problem solving (Branson and

Rogers, 2006; Batt et al., 2007). This method provides an unbiased assessment of each animal because they are all treated in the same way, and improves on Methods 2–6. The provocative test chosen needs to pertain to the questions the authors are asking, and without other information, inferences about mechanism must be made cautiously.

Method 8. A combination of the above

This is the superior approach to ensuring that behaviors are evaluated in as unbiased a way as possible, and in a manner that ensures collection of data that can help define a 'response surface' (Overall, 2005). The combination of Methods 4–7 are likely to produce the most informative data and allows you to ask about consistency of data obtained using various methods.

Conclusions

When an approach declares itself outside the accepted methodologies of science, it should not and cannot be taken seriously by scientists. Hypothesis testing and falsification are at the very core of the scientific approach. If homeopathy and other CAMs wish to be considered by scientists, they must be shown to be valid using methods that science uses to evaluate all treatment modalities. If these fields are not willing to comply with these rules they cannot be considered scientific and cannot be used in any set of scientific and medical best practices.

Quite simply, the onus is not on the rest of the scientific community to provide and prove others' experiments – it is on the proponents of CAM. If, for validation, one has to invoke another, non-specific way of thinking or authority separate from that used in science, one should not be surprised when the outcome of such invocations is that results are considered non-scientific and the proponents' assertions are considered invalid using the methods by which all of science is tested.

The ultimate findings of Cracknell and Mills (2008) could, we suggest, be restated as follows:

1. There is no evidence of any effect of the homeopathic 'treatment';
2. There was no effect of treatment using the homeopathic 'treatment';
3. Dogs suffering from fear associated with the noise of fireworks will not benefit from 'treatment' with the homeopathic preparation;
4. The homeopathic preparation will not help fearful dogs who worry about the noise of fireworks.

During the course of writing this editorial, one of us (KLO) received two e-mail exhortations (13 August 2008 and 12 September 2008) from HomeoPet[1] that their product known as Storm Stress had been shown to be 98% effective, with a 94% client satisfaction rating in placebo-controlled, double-blind studies. Their unpublished and apparently unreviewed placebo-controlled, double-blind study is available online, and it is not what it purports to be. Given that it cites Overall et al. (2001), implying support of some of the claims, this is not good. In fact, the online reference could be the exemplar by which every point made in this article could be taught.

By co-opting the status implied by the phrase 'placebo-controlled, double-blind studies', Homeopet asserts efforts not made and effects not found to gain them credentials they do not have in the eyes of a public that cannot know this. The veterinary community simply must not allow their intellectual and professional credibility to be stolen so blatantly.

[1] See: www.homeopetpro.com.

Acknowledgment

Dr. Soraya V. Juarbe-Diaz helpfully commented on an earlier version of this manuscript.

References

Batt, L., Batt, M., McGreevy, P., 2007. Two tests for motor laterality in dogs. Journal of Veterinary Behavior: Clinical Applications and Research 2, 47–51.

Bausell, R.B., 2007. Snake Oil Science. The Truth about Complementary and Alternative Medicine. Oxford University Press, New York, 324 pp.

Branson, N.J., Rogers, L.J., 2006. Relationship between paw preference strength and noise phobia in *Canis familiaris*. Journal of Comparative Psychology 120, 176–183.

Brown, D.C., 2006. Control of selection bias in parallel-group controlled clinical trials in dogs and cats: 87 trials (2000–2005). Journal of the American Veterinary Medical Association 229, 990–993.

Cracknell, N.R., Mills, D.S., 2008. A double-blind placebo-controlled study into the efficacy of a homeopathic remedy for fear of firework noises in the dog (*Canis familiaris*). The Veterinary Journal 177, 80–88.

Crowell-Davis, S.L., Seibert, L.M., Sung, W., Parthasarathy, V., Curtis, T.M., 2003. Use of clomipramine, alprazolam, and behavior modification for treatment of storm phobia in dogs. Journal of the American Veterinary Medical Association 222, 744–748.

Davenas, E., Beauvais, F., Amara, J., Oberbaum, M., Robinzon, B., Miadonnai, A., Tedeschi, A., Pomeranz, B., Fortner, P., Belon, P., Sainte-Laudy, J., Potevin, B., Benveniste, J., 1988. Human basophil degranulation triggered by very dilute antiserum against IgE. Nature 333, 816–818.

DeAngelis, C.D., Fontanarosa, P.B., 2008. Impugning the integrity of medical science. The adverse effects of industry influence. Journal of the American Medical Association 299, 1833–1835.

Ernst, E., 2002. A systematic review of systematic reviews of homeopathy. British Journal of Clinical Pharmacology 54, 577–582.

Hribjartsson, A., Gotzche, P.C., 2001. Is the placebo powerless? – An analysis of clinical trials comparing placebos with no treatment. New England Journal of Medicine 344, 1594–1602.

King, J.N., Overall, K.L., Appleby, D., Simpson, B.S., Beata, C., Chaurand, C.J.P., Heath, S.E., Ross, C., Weiss, A.B., Muller, G., Bataille, B.G., Paris, T., Pageat, P., Brovedani, F., Garden, C., Petit, S., 2004. Results of a follow-up investigation to a clinical trial testing the efficacy of clomipramine in the treatment of separation anxiety in dogs. Applied Animal Behaviour Science 89, 233–242.

King, J., Simpson, B., Overall, K.L., Appleby, D., Pageat, P., Ross, C., Chaurand, J.P., Heath, S., Beata, C., Weiss, A.B., Muller, G., Paris, T., Bataille, B.G., Parker, J., Petit, S., Wren, J., . Treatment of separation anxiety in dogs with clomipramine. Results from a prospective, randomized, double-blinded, placebo-controlled clinical trial. Applied Animal Behaviour Science 67, 255–275.

Linde, K., Scholtz, M., Ramirez, G., Clausius, N., Melchart, D., Jonas, W.B., 1999. Impact of study quality on outcome in placebo controlled trials of homeopathy. Journal of Clinical Epidemiology 52, 631–636.

Linde, K., Melchart, D., 1999. Randomized controlled trials of individualized homeopathy: a state-of-the-art review. Journal of Alternative and Complementary Medicine 4, 388.

Maddox, J., Randi, J., Stewart, W.W., 1988. 'High-dilution' experiments a delusion. Nature 334, 287–290.

Martin, P., Bateson, P., 1986. Measuring Behaviour: An Introductory Guide. Cambridge University Press, NY, 200 pp.

OED online. <http://proxy.library.upenn.edu:2340/entrance.dtl?side=S> (accessed 6.10.2008).

Overall, K.L., 2005. Veterinary behavioural medicine: a roadmap for the 21st century. The Veterinary Journal 169, 130–143.

Overall, K.L., Agulnick, L., Kapes, M., Dunham, A.E., 1999. Sonographic analysis of dog vocalization: a pilot study involving distressed and unaffected dogs (poster/abstract). American Veterinary Society of Animal Behavior (AVSAB) Meeting, New Orleans, LA, July 1999.

Overall, K.L., Dunham, A.E., Acland, G., 1999b. Responses of genetically fearful dogs to the lactate test: assessment of the test as a provocative index and application in mechanistic diagnoses (poster/abstract), World Congress on Psychiatric Genetics, Monterey, CA. Molecular Psychiatry 4, S125.

Overall, K.L., Dunham, A.E., Frank, D., 2001. Frequency of nonspecific clinical signs in dogs with separation anxiety, thunderstorm phobia, and noise phobia, alone or in combination. Journal of the American Veterinary Medical Association 219, 467–473.

Overall, K.L., Hamilton, S.P., Chang, M.L., 2006. In Brief: understanding the genetic basis of canine anxiety: phenotyping dogs for behavioral, neurochemical, and genetic assessment. Journal of Veterinary Behavior: Clinical Applications and Research 1, 124–141.

Park, R., 2000. Voodoo Science. The Road from Foolishness to Fraud. Oxford University Press, New York, 230 pp.

Ross, J.S., Hill, K.P., Egilman, D.S., Krumholz, H.M., 2008. Guest authorship and ghostwriting in publication related to rofecoxib: a case study of industry documents from rofecoxib litigation. Journal of the American Medical Association 299, 1800–1812.

Saper, R.B., Phillips, R.S., Sehgal, A., Khouri, N., Davis, R.B., Paquin, J., Thuppil, V., Kales, S.N., 2008. Lead, mercury, and arsenic in US- and Indian-manufactured Ayurvedic medicines sold on the internet. Journal of the American Medical Association 300, 915–923.

Serpell, J.A., Hsu, Y., 2001. Development and validation of a novel method for evaluating behavior and temperament in guide dogs. Applied Animal Behaviour Science 72, 347–364.

Sheppard, G., Mills, D.S., 2003. Evaluation of dog-appeasing pheromone as a potential treatment for dogs fearful of fireworks. Veterinary Record 152, 432–436.

Whitehead, Anne, 2002. Meta-Analysis of Controlled Clinical Trials. John Wiley and Sons Ltd., London, 336 pp.