Chance level performance in expert diagnoses with applied kinesiology: A case study in bespoke inference

Jennifer Wilson, Beth Baribault, and Joachim Vandekerckhove

4/11/2018

Abstract

We test the diagnostic validity of manual muscle testing (MMT). The manual muscle test is an alternative-medicine technique used to assess the suitability of natural remedies for minor medical ailments. The technique involves a trained tester gauging a patient's muscle resistance while the patient holds a candidate supplement close to the body, and supplements are then recommended on the basis of between-trial changes in muscle resistance. MMT is widely used despite the lack of a known underlying physical mechanism. In a pre-registered study, we evaluate the ability of practitioners to reliably rank supplements in order of suitability (i.e., in terms of their positive effect on perceived muscle resistance). We provide details of a custom analysis in which we quantify the evidence for competing accounts. The data are overwhelmingly more consistent with the supposition that the rankings are random than with any competing account. All our data and methods are available on the Open Science Framework (osf.io/8d4wy).

Introduction

Manual muscle testing (MMT) is a diagnostic technique used in alternative medicine to assess the suitability of natural food supplements (e.g., tea, bee pollen. . .) for minor medical conditions (e.g., chronic fatigue). Several schools of applied kinesiology @eden2008energy involve a belief that the *mere proximity* of such remedial supplements has a measurable effect on a patient's muscle tone, so that the suitability of a remedy can be predicted by the effect it has on patient muscle strength. According to a recent survey @jensen2015, some 200,000 practitioners adhere to one of these schools.

The manual muscle test

The use of MMT as a diagnostic procedure involves exposing the patient to various substances in order to detect either a remedy or toxin. A weakening of muscle tone is interpreted as a signal that the substance is a toxin to the patient whereas a tonifying (i.e., strengthening) response is taken to signify a potential remedy. Substances can be variously tested through ingestion, insalivation, or non-local proximity (NLP;). Non-local proximity testing relies on the assumption that the mere presence of a substance can cause a change in muscle strength. Procedures that rely on NLP have the advantage of avoiding the need for ingestion or topical application (i.e., they are entirely noninvasive). NLP testing is commonly performed using a standard "arm pull down" muscle test (see Figure 1).

To perform a standard muscle test using the arm-pull-down technique, a patient is asked to stand and hold a substance (a potential remedy ortoxin) next to the body with one arm, while extending the opposite arm, palm down. The muscle tester will then apply pressure to the back of the extended hand and judge whether the muscle response was "strong" or "weak" relative to trials in which other substances are held.

Previous research

While there is some evidence to support the notion that experts are able to detect trial-to-trial changes in muscle tone @florence1984clinical [@pollard2005interexaminer; @schmitt1998], previous studies on the efficacy of the non-local manual muscle test have yielded conflicting results, with some supporting the validity of the test @radin1984possible and others finding inconclusive results @arnett1999double [@keating2004evaluation; @ludtke2001test; @quintanar1988sugar].

Unfortunately, while Radin's finding of a nonlocal effect has failed to replicate, these failures to replicate are weakened by their reliance on classical null hypothesis significance testing (NHST). NHST allows researchers to reject, but never confirm, null hypotheses: The procedure starts with the assumption that there is no effect (the null hypothesis) and then uses that assumption to compute the probability p that data as extreme as, or more extreme than, the real data would be observed upon repeated execution of the exact same study. But, since the null hypothesis is assumed, it cannot logically be concluded. In contrast, we will use a comparative procedure where we quantify the evidence for (or against) the null hypothesis.

Experiment

Participants

We recruited two groups of participants: "patient" participants who reported suffering from fatigue and "muscle tester" participants who were trained in manual muscle testing techniques.

"Patient" participants responded to fliers posted in the Orange County community and on the campus of the University of California, Irvine. After responding, participants completed a Fatigue Assessment Scale (FAS;). Participation was contingent on the participant's FAS score or level of fatigue in addition to their consent to participate in the experiment, which was described as involving light pressure applied to the arm. Participants also read a description about muscle testing and stated whether or not they believed in its efficacy. A total of 18 participants with ages ranging from 18 to 38 continued to the testing phase.

Muscle testers responded to fliers posted in the community or emails requesting experienced muscle testers. Email addresses were obtained by searching for practitioners trained in applied kinesiology in Orange County, CA. Once muscle testers expressed interest, they read a description of the type of technique being studied and the test they would be asked to perform. Nine muscle testers chose to participate in the study. The muscle tester group consisted of 4 men and 5 women whose ages ranged from 44 to 72. The typical muscle tester in the sample had approximately 10 years of experience. However, experience ranged from two to thirty-seven years.

Materials

For the to-be-tested substances, we selected an assortment of five herbal remedies commonly used to treat fatigue: ashwagandha, bee pollen, yerba mate, eleuthero, and green tea. These herbs are commonly used in an attempt to increase energy and were selected because of their placement in adaptogenic energy sections in local health food stores in Orange County, CA. Two grams of each herb and two cotton balls were placed into opaque bottles and sealed. The bottles were randomized and labeled A through E. Indistinguishability of the different bottles was confirmed by the senior author (JV) who was blinded to the assembly of the bottles.

"Patient" participants were provided a description of the muscle testing procedure and asked whether they believed that muscle testing has the ability to determine if a supplement is needed by the body. The description and question are available via the Open Science Framework (OSF; https://osf.io/qe9p4/).

Procedure

Testing was divided into two separate two-hour sessions. In the first session six muscle testers performed muscle tests on each of ten patients. In the second session, three muscle testers performed tests on each of eight patients. Prior to testing, patient and muscle testers were led into separate rooms to discuss the procedure. Patients were debriefed on how the test is done. Muscle testers were briefed on the technique

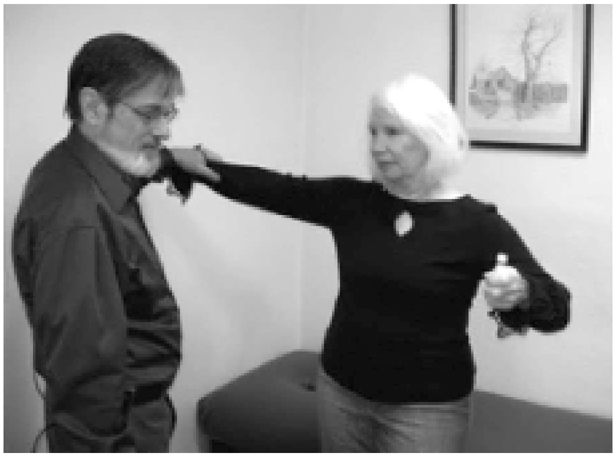

Figure 1: Figure 1: A muscle testing trial. The participant is standing with one arm outstretched and holding a vial while the muscle tester applies gentle downward pressure on the outstretched arm.

being used, the materials to use during testing, and how to record results. At this point, two muscle testers (both in the second session) were no longer comfortable with the “arm pull down” technique as described in the recruitment email. Their data were incomplete at the end of the session and were not used for the analyses. When testing began, all muscle testers were in separate rooms and participants moved between rooms to interact with each muscle tester.

Each muscle tester received a set of five herbs in opaque bottles, randomly labeled $A-E$. The testers were asked to test each patient in their session using these five bottles and then rank the five bottles in order of their suitability for that patient. After testing a patient with all five bottles, the testers were asked to indicate with a simple Yes/No question whether they were confident in their judgment. The response sheet is available on OSF (https://osf.io/ysv92/).

Once the testers finished testing every patient in their session, the bottles were re-randomized (i.e., the testers were given a different set of bottles with the same substances but a different random set of labels $A-E$) and the testers re-tested all patients.

The experimental protocol was preregistered on OSF (https://osf.io/wymjr/) and approved by the Institutional Review Board of the University of California, Irvine (HS# 2015-1926).

Model-based data analysis

In this section, we provide full detail regarding our custom model-based data analysis. The section concludes with an informal conceptual summary that avoids most of the technical detail.

To avoid issues with null hypothesis significance testing (i.e., the inability to confirm null hypotheses;), we opted for a model-comparison approach to analyze the consistency of muscle test results. If the muscle testing procedure is sensitive to a true signal, rank orders given by two muscle testers for the same patient should agree better than chance. Similarly, rank orders provided by a single muscle tester on two occasions should agree better than chance.

In order to quantify agreement between and within muscle testers, we used Kendall’s @Kendall1938 scoring rule for ordered lists. One version of the rule involves determining the score Kendall’s τ, which is the total number of adjacent pairwise swaps required to move from a given order to a target order.[1]

[1] This is one of several current definitions of Kendall’s τ. Other definitions are linearly rescaled versions of this definition.

A family of distributions of Kendall's τ

In the absence of biases, and assuming the complete absence of agreement, all rank orderings of k items—of which there are $k!$ possible—are equally likely. This implies an expected probability distribution for τ. The expected probabilities of all scores τ, by list length, are given in Table 1. The probability of observing a given τ with a certain list length k can now be determined via this lookup table. We denote the probability as $F_k(\tau)$, so that $F_4(3) = 6/24 = 0.25$ (because 6 of the 24 possible orderings of a four-item sequence have a Kendall's τ of 3).

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$\tau = 0$	1	1/2	1/6	1/24	1/120
$\tau = 1$	.	1/2	2/6	3/24	4/120
$\tau = 2$	.	.	2/6	5/24	9/120
$\tau = 3$	.	.	1/6	6/24	15/120
$\tau = 4$	.	.	.	5/24	20/120
$\tau = 5$	.	.	.	3/24	22/120
$\tau = 6$	.	.	.	1/24	20/120
$\tau = 7$	.	.	.	.	15/120
$\tau = 8$	.	.	.	.	9/120
$\tau = 9$	.	.	.	.	4/120
$\tau = 10$	.	.	.	.	1/120
	1	2	6	24	120

Table 1: Probability distributions of Kendall's τ, by sequence length k. Each column gives the probabilities for different values of τ assuming a sequence of length k. The bottom row gives $n!$, the total number of possible orderings for a list of length n.

These distributions of τ in the absence of agreement and bias can serve to construct a new family of distributions of τ. First, the multinomial distribution $F_k(\tau)$ expresses the probability of observing a particular value of τ in a list of length k: $P(\tau \mid k) = F_k(\tau)$.

In order to perform model comparison, it will be useful to define alternative distributions of τ in which the assumption of no agreement is relaxed. We define a *noncentral* version of the $F_k(\tau)$ distribution family whose noncentrality parameter σ indicates *the number of extreme items on which a pair of raters agree* (i.e., any pair of muscle testers will agree on the position of the top or bottom σ items). σ is a non-negative integer and $\sigma < k$. The distributions satisfy

$$F_k(\tau \mid \sigma) = F_{k-\sigma}(\tau \mid 0) = F_{k-\sigma}(\tau),$$

meaning that for each agreed-upon item, the probability distribution of σ shifts one column to the left (in Table 1).

If $\sigma = 0$, there is no agreement and the expected distribution reduces to the central $F_k(\tau)$ distribution. If $\sigma = k - 1$, there is complete agreement—because the n^{th} item is then determined—and τ must be 0.

Inferring the agreement σ

The new parameter σ has a straightforward interpretation as the least number of extreme items on which a pair of participants agree. Given a set of k observed scores $T_k = (\tau_1, \tau_2, \ldots, \tau_k)$, the joint likelihood is

$$F_k(T_k \mid \sigma) = \prod_{i=1}^{k} F_k(\tau_i \mid \sigma),$$

and with a conjugate multinomial prior distribution for σ, the posterior distribution of σ assuming the model $F_k(\cdot)$ is

$$P(\sigma \mid F_k(\cdot), T_k) = \frac{P_\sigma(\sigma) \prod_{i=1}^{k} F_k(\tau_i \mid \sigma)}{\sum_{s=0}^{k-1} P_\sigma(s) \prod_{i=1}^{k} F_k(\tau_i \mid s)}.$$

A lapse rate λ

The multinomial likelihood function $F_k(\cdot \mid \sigma)$ has the property that it is zero for values $\tau > \frac{1}{2}(k-\sigma)(k-\sigma-1)$. While it is true that these values are impossible if participants agree on σ or more items and make no reporting errors, we can allow for reporting errors by introducing a *lapse rate* λ, leading to the following likelihood:

$$G_k(T_k \mid \sigma, \lambda) = \prod_{i=1}^{k} \left[(1-\lambda)\, F_{k-\sigma}(\tau_k) + \lambda F_k(\tau_k)\right].$$

That is, with probability $(1-\lambda)$ the response comes from the regular process, so the likelihood is $F_{k-\sigma}(\tau)$, but with probability λ the response comes from the completely random process, so the likelihood is $F_k(\tau)$. The lapse rate λ could be estimated, integrated away as a nuisance variable, or can be set to a convenient small value such as .05. We choose to treat λ as a nuisance variable and assign it a uniform distribution from 0 to $\frac{1}{2}$; that is, we do not believe that more than 50% of data points will be the result of a lapse. The density is then $p_\lambda(\cdot) = 2$, and the nuisance variable is treated by taking a weighted average of the likelihood using this prior as weight. Since λ is continuous, the weighted average is an integral over λ:

$$G_k(T_k \mid \sigma) = \int_0^{\frac{1}{2}} p_\lambda(\lambda) \prod_{i=1}^{k} \left[(1-\lambda)\, F_{k-\sigma}(\tau_k) + \lambda F_k(\tau_k)\right] d\lambda.$$

Combining this marginalized likelihood with the prior for σ and normalizing yields the posterior probabilities of ultimate interest:

$$P(\sigma \mid G_k(\cdot), T_k) = \frac{P(\sigma) G_k(T_k \mid \sigma, \lambda)}{\sum_{s=0}^{k-1} P(s) G_k(T_k \mid \sigma, \lambda)}.$$

Prior probability and Bayes factors

In order to obtain a posterior probability, we need to define an a-priori probability for each value of σ. These prior probabilities simply indicate the strength of our belief in each possible σ-value—but they are necessarily subjective. We address this subjective element in two ways.

First, we defined three reasonable prior distributions: a *flat prior* under which each value for σ is equally likely (20%); a *skeptical prior* under which the model of no correspondence ($\sigma = 0$) is most likely (50%) and the other four models are equally likely (12.5%); and an *adherent prior* under which the model with complete correspondence ($\sigma = 4$) is most likely (50%) and the four other models are equally likely (12.5%). To assess the influence of prior beliefs in σ, we compare the outcome of our analysis under each scenario. It will turn out that our conclusions are robust to the choice of prior.

Second, we will compute a *Bayes factor* @Kass1995. While our ultimate interest is in the posterior probability or the *posterior ratio*—the relative probability that $\sigma = 0$ versus $\sigma > 0$, where the latter's probability is the sum of the posterior probabilities for all nonzero values of σ—this ratio can be written as a product of the *prior ratio* and some factor B_{01}:

$$\frac{P(\sigma = 0 \mid G_k, T_k)}{P(\sigma > 0 \mid G_k, T_k)} = \frac{P(\sigma = 0)}{P(\sigma > 0)} \times B_{01}$$

In this equation, B_{01} is known as the *Bayes factor* and here indicates the degree to which the data support the “random-assignment hypothesis” ($\sigma = 0$) over the alternative ($\sigma > 0$). Importantly, B_{01} is independent of the prior probabilities of σ and is therefore an attractive quantification of evidence in the data. The Bayes factor is also symmetric, in that it can simply be inverted to obtain the degree of support for the alternative hypothesis ($B_{10} = B_{01}^{-1}$). Conventionally, a Bayes factor of 10 or more indicates strong evidence.

Conceptual summary of the analysis technique

We developed a novel procedure to quantify the consistency in muscle test results both across muscle testers and across testing occasions. We consider two broad scenarios: a “random-response model” under which rank orders agree only by chance and an “alternative model” under which rank orders agree more than would be expected by chance. We will calculate a *Bayes factor* that expresses the relative evidence for the random-response model over the alternative model. A Bayes factor of 10 or more indicates strong evidence.

Results

The Bayes factor favoring the random-response model is approximately 1.6 trillion ($B_{01} \approx 1.6205 \times 10^{12}$).

indicating overwhelmingly strong evidence. Accordingly, the posterior probability that $\sigma = 0$ approaches 1 for all prior distributions we consider.

For a graphical illustration of this model selection result, Figure 2 depicts with lines the expected probability distribution of τ for all possible values of σ. The distributions widen and flatten as agreement σ goes to 0, and they peak at complete agreement ($\tau = 0$) as σ goes to $k - 1$. The bars indicate the observed distribution in our sample. Visual inspection reveals a striking correspondence between the data and the random-response model of no correspondence.

For this analysis, we excluded trial pairs in which one tester did not indicate strong confidence in their judgment (10 pairs) and trial pairs in which the patient indicated that they did not believe in the effectiveness of muscle testing (42 pairs). This left a total of 83 trial pairs (out of 129 initially). All changes to our censoring rule (e.g., also including trials with low tester confidence or trials with skeptical patients) led to Bayes factors that *more strongly* supported the random-response model (with $B_{01} \approx 2.131 \times 10^{21}$ if no censoring is performed at all), so we do not elaborate on these decisions.

A similar test can be applied to the within-tester consistency (between the two occasions on which the same patient was tested by the same tester). For this scenario, $B_{01} \approx 2.983 \times 10^{3}$ with censoring and $B_{01} \approx 5.626 \times 10^{6}$ without.

Discussion

The practice of manual muscle testing as a diagnostic tool implies a perceptual challenge for the practitioner. The reliability of the test can be evaluated via the correspondence between repeated tests (the same tester with the same patient) and via the correspondence between testers (with the same patient). On both measures, the degree of correspondence was overwhelmingly more consistent with a random-response model than with any model that assumed the detection of a signal. The assignment of supplements by our trained testers was random despite the tester’s reported confidence in the outcomes and the participant’s reported faith in the procedure.

Our experiment delivers strong evidence that manual muscle testing with non-local proximity fails as a diagnostic procedure.

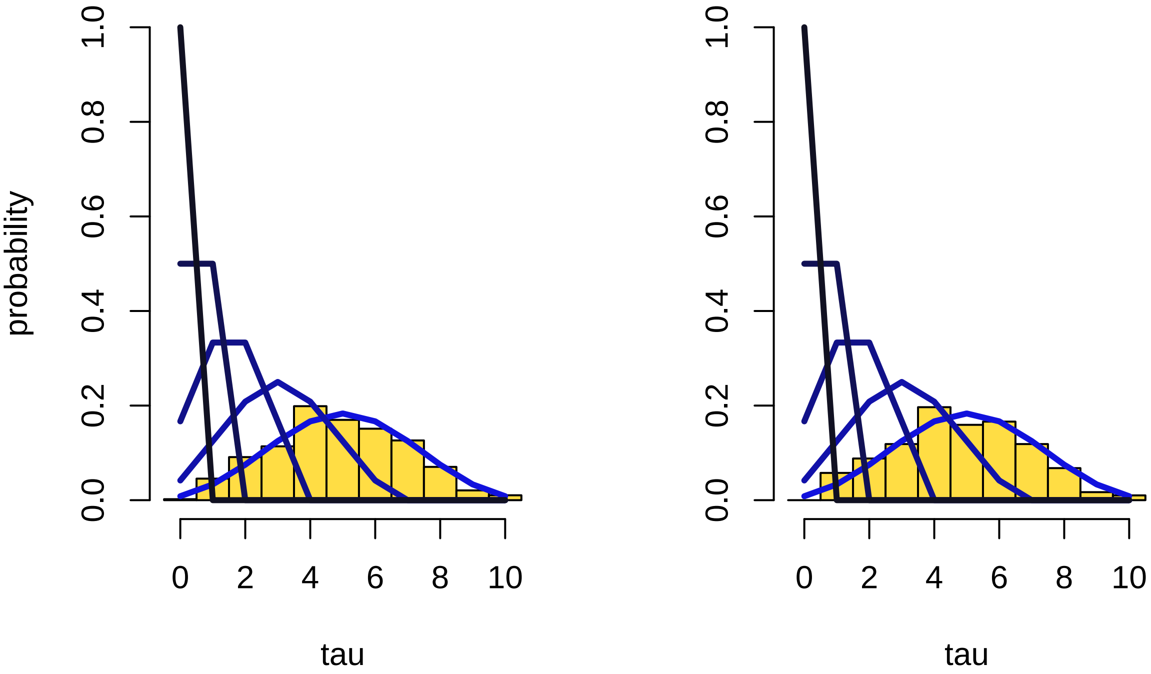

Figure 2: A muscle testing trial. Distributions of tau as predicted by each of the five models under consideration (lines), overlaying the observed distribution of tau in the experiment (bars). The lines are graphical representations of the model-predicted expected probabilities, which were computed by combining the probabilities in Table 1 with a lapse process that's given by the rightmost column in that table. The data appear to correspond most closely to the model assuming no agreement (s=0) both for the global correspondence analysis (left) and the internal consistency analysis (right).