

PUBLISHED VERSION

Alan H. Welsh, Emma J. Knight

"Magnitude-based inference": a statistical review

Medicine and science in sports and exercise, 2015; 47(4):874-884

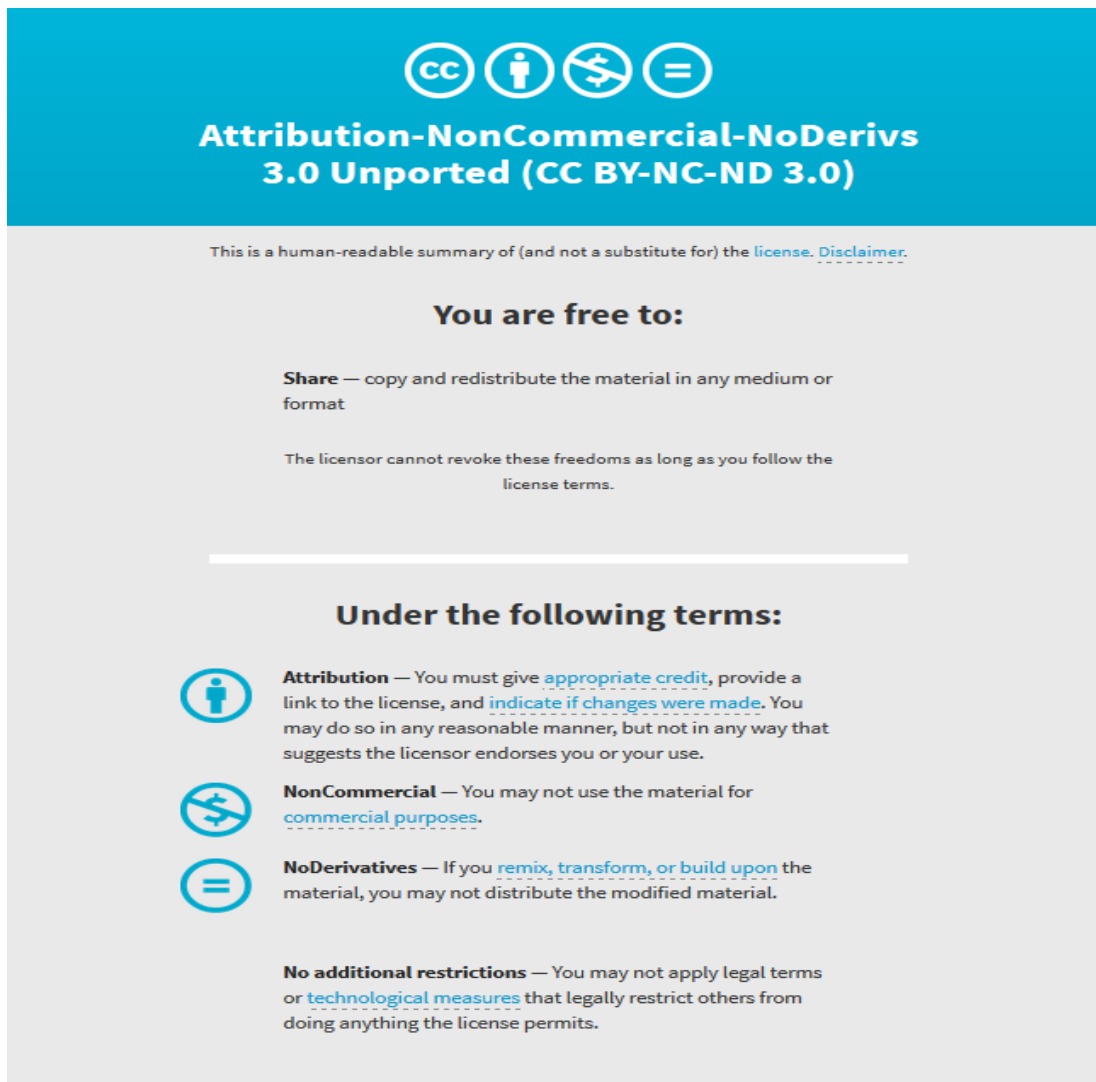
© 2014 by the American College of Sports Medicine. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License, where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

Originally published at:

<http://doi.org/10.1249/mss.0000000000000451>

PERMISSIONS

<http://creativecommons.org/licenses/by-nc-nd/3.0/>



The image shows the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 Unported (CC BY-NC-ND 3.0) license graphic. It features a blue header with the license name and icons for Attribution (person), Non-Commercial (dollar sign with slash), and No Derivatives (equals sign). Below the header, it states: "This is a human-readable summary of (and not a substitute for) the license. [Disclaimer.](#)"

You are free to:

- Share** — copy and redistribute the material in any medium or format

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- NonCommercial** — You may not use the material for [commercial purposes](#).
- NoDerivatives** — If you [remix, transform, or build upon](#) the material, you may not distribute the modified material.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

13th of April 2018

<http://hdl.handle.net/2440/110054>

“Magnitude-based Inference”: A Statistical Review

ALAN H. WELSH¹ and EMMA J. KNIGHT²

¹*Mathematical Sciences Institute, Australian National University, Canberra, Australian Capital Territory, AUSTRALIA; and*
²*Performance Research, Australian Institute of Sport, Belconnen, Australian Capital Territory, AUSTRALIA*

ABSTRACT

WELSH, A. H., and E. J. KNIGHT. “Magnitude-based Inference”: A Statistical Review. *Med. Sci. Sports Exerc.*, Vol. 47, No. 4, pp. 874–884, 2015. **Purpose:** We consider “magnitude-based inference” and its interpretation by examining in detail its use in the problem of comparing two means. **Methods:** We extract from the spreadsheets, which are provided to users of the analysis (<http://www.sportsci.org/>), a precise description of how “magnitude-based inference” is implemented. We compare the implemented version of the method with general descriptions of it and interpret the method in familiar statistical terms. **Results and Conclusions:** We show that “magnitude-based inference” is not a progressive improvement on modern statistics. The additional probabilities introduced are not directly related to the confidence interval but, rather, are interpretable either as *P* values for two different nonstandard tests (for different null hypotheses) or as approximate Bayesian calculations, which also lead to a type of test. We also discuss sample size calculations associated with “magnitude-based inference” and show that the substantial reduction in sample sizes claimed for the method (30% of the sample size obtained from standard frequentist calculations) is not justifiable so the sample size calculations should not be used. Rather than using “magnitude-based inference,” a better solution is to be realistic about the limitations of the data and use either confidence intervals or a fully Bayesian analysis. **Key Words:** BAYESIAN, BEHRENS–FISHER, CONFIDENCE INTERVAL, FREQUENTIST

Over the last decade, “magnitude-based inference” has been developed and promoted in sport science as a new method of analyzing data. Information about the approach is available from Excel spreadsheets, presentations, notes, and articles (4,5,11,12,14), many of which are available from the Web site <http://www.sportsci.org/>. More recently, the approach has been recommended by Wilkinson (25,26). Although “magnitude-based inference” is a statistical approach that is intended to replace other statistical approaches, it has so far attracted minimal scrutiny by statisticians; as far as we know, the only published comments on it by statisticians are those of Barker and Schofield (3) who showed that the approach can be interpreted as an approximate Bayesian procedure. The purpose of this article is

to present a detailed examination of “magnitude-based inference” as a statistical method, examining it both as a frequentist and a Bayesian method.

The development of “magnitude-based inference” seems to have been motivated by 1) some legitimate questions about the use of frequentist significance testing (*P* values) in clinical practice and 2) by the perception that significance tests (at the 5% level) are too conservative when looking for small effects in small samples. In response to such questions about significance testing, a number of researchers advocate the use of confidence intervals instead of *P* values (6,7,9,17,20) but “magnitude-based inference” tries to go further, replacing the confidence interval with probabilities that are supposedly based on the confidence interval.

The first essential step in discussing “magnitude-based inference” is to obtain a clear description of the approach, for which we take the spreadsheets as the definitive implementation of the method. For simplicity, we focus on two specific spreadsheets, describe the “magnitude-based inference” calculations presented in these spreadsheets, and evaluate the method by interpreting the calculations against the explanations given in the published articles (4,5,11,12,14). The spreadsheets we used are *xParallelGroupsTrial.xls* and *xSampleSize.xls* (see spreadsheets, Supplemental Digital Content 1, <http://links.lww.com/MSS/A429>, and Supplemental Digital Content 2, <http://links.lww.com/MSS/A430>, obtained from <http://www.sportsci.org/> on 22 May 2014 under the links “Pre–post parallel groups trial” and “Sample size estimation”), which implement “magnitude-based inference” calculations for what is often loosely described as the problem

Address for correspondence: Emma Knight, Ph.D., Performance Research, Australian Institute of Sport, PO Box 176, Belconnen, Australian Capital Territory 2616, Australia; E-mail: emma.knight@ausport.gov.au.
Submitted for publication March 2014.

Accepted for publication July 2014.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal’s Web site (www.acsm-msse.org).

This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License, where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

0195-9131/15/4704-0874/0

MEDICINE & SCIENCE IN SPORTS & EXERCISE®

Copyright © 2014 by the American College of Sports Medicine

DOI: 10.1249/MSS.0000000000000451

of comparing two means. We reverse-engineered parts of these spreadsheets and rewrote the spreadsheet calculations in R (19) to check that we got the same numerical results and thereby confirm that our transcription of the calculations is correct.

We describe the problem of comparing two means to set the context for using *xParallelGroupsTrial.xls* and *xSampleSize.xls* to describe and discuss “magnitude-based inference” in section 2. We then describe the calculations used in “magnitude-based inference” for the problem of comparing two means in section 3. At the end of the section, we introduce a new graphical representation to illustrate how the approach works. We provide evaluation and comment regarding the calculations in section 4 and describe the “magnitude-based inference” sample size calculation for the problem of comparing two means in section 5. We present further discussion in section 6 and some concluding remarks in section 7. Our conclusion is that “magnitude-based inference” does not get away from using *P* values as it purports to do but actually uses nonstandard *P* values and very high thresholds to increase the probability of finding effects when none are present. Furthermore, the smaller sample size requirements are illusory and should not be used in practice.

THE PROBLEM OF COMPARING TWO MEANS

The problem considered in *xParallelGroupsTrial.xls* (see spreadsheet, Supplemental Digital Content 1, <http://links.lww.com/MSS/A429>) is the problem of comparing two means. More specifically, it is the problem of making inferences about the difference in the means of two normal populations with possibly different variances, on the basis of independent samples from the two populations. This problem is illustrated in *xParallelGroupsTrial.xls* by an example with data from a control group of 20 athletes (in cells E42 to H61) and an experimental group of 20 different athletes (in cells E73 to H92). There are four measurements on each athlete (two before-treatment measurements labeled pre1 and pre2 and two after-treatment measurements labeled post1 and post2). The data are approximately normally distributed, so there is no need to transform the data and the individual treatment effects can be estimated by post1 – pre2 (as is done in cells L42 to L61 and L73 to L92). These individual treatment effects are assumed to be independent, and the problem is to make inferences about the effect of the treatment on a typical (randomly chosen) individual; this effect is summarized by the difference in the means of the separate populations represented by the experimental and control athletes.

The mentioned scenario is a particular example of a general problem in which we have n_1 subjects in a control group and different n_2 subjects in an experimental group, and we have observed individual effects Y_{11}, \dots, Y_{1,n_1} on the control group and observed individual effects Y_{21}, \dots, Y_{2,n_2} on the experimental group. Here, the first subscript represents the

group (“1” identifies the control group, and “2” identifies the experimental group) and the second subscript identifies the subject in the group. The observed effects are conceptualized as realizations of mutually independent normal random variables, such that the n_1 subjects in the control group have mean μ_1 and variance σ_1^2 and the n_2 subjects in the experimental group have mean μ_2 and variance σ_2^2 , and we want to make inferences about the difference in means $\mu_2 - \mu_1$. For simplicity, we assume throughout this article that positive values of $\mu_2 - \mu_1$ represent a positive or beneficial effect. The general problem of making inferences about $\mu_2 - \mu_1$ in this normal model with $\sigma_1^2 \neq \sigma_2^2$ is known as the Behrens–Fisher problem (see for example, Welsh (23)). The Behrens–Fisher problem seems simple on the surface but is in fact a difficult problem that has generated substantial literature. We could simplify to the equal variance problem ($\sigma_1^2 = \sigma_2^2$) but chose to follow the spreadsheets.

“MAGNITUDE-BASED INFERENCE” CALCULATIONS

The calculations for “magnitude-based inference” that we have extracted from the spreadsheet *xParallelGroupsTrial.xls* are expressed in this article in standard mathematical notation rather than as spreadsheet commands. In *xParallelGroupsTrial.xls*, all probabilities *p* are specified as percentages (i.e., 100*p*) and as odds. The definition used in *xParallelGroupsTrial.xls*, 1:(1 – *p*)/*p* if $p < 0.5$ and $p/(1 - p)$ if $p \geq 0.5$, is more complicated than the standard definition $p/(1 - p)$ of odds. Percentages and odds are mathematically equivalent to specifying the probabilities, but we use the probabilities because they are simpler for mathematical work and avoid using the nonstandard definition of odds.

“Magnitude-based inference” is described as being based on a confidence interval for the quantity of interest (here, $\mu_2 - \mu_1$), which is then categorized on the basis of some additional probability calculations. We introduce the notation and the setup by describing the confidence interval used for $\mu_2 - \mu_1$ and then describing the additional probability calculations.

Confidence intervals: approach 1. The first step in “magnitude-based inference” is to compute the approximate 100(1 – α)% Student’s *t* confidence interval (default level 90% entered in E33 or $\alpha = 0.1$) for $\mu_2 - \mu_1$, which does not assume equal population variances and uses Welch’s (22) approximation to the degrees of freedom. Specifically, we estimate $\mu_2 - \mu_1$ by the difference in sample means as $\bar{Y}_2 - \bar{Y}_1$ (in L117), compute the SE of the difference in sample means as $SE(\bar{Y}_2 - \bar{Y}_1)$ (in L123), and then the approximate confidence interval (in L130 and L131) as

$$[\bar{Y}_2 - \bar{Y}_1 - t_\alpha SE(\bar{Y}_2 - \bar{Y}_1), \bar{Y}_2 - \bar{Y}_1 + t_\alpha SE(\bar{Y}_2 - \bar{Y}_1)] \quad [1]$$

where t_α is the critical value (see Appendix, Supplemental Digital Content 3, <http://links.lww.com/MSS/A431>, Background information and formulas for the *P* values and confidence interval for the problem of comparing two means).

TABLE 1. The “qualitative probabilities” used in *xParallelGroupsTrial.xls*.

Range of <i>P</i>	Interpretation
$P < 0.005$	Most unlikely
$0.005 \leq P < 0.05$	Very unlikely
$0.05 \leq P < 0.25$	Unlikely
$0.25 \leq P < 0.75$	Possibly
$0.75 \leq P < 0.95$	Likely
$0.95 \leq P < 0.995$	Very likely
$0.995 < P$	Most likely

The next step is to specify the smallest meaningful positive effect, $\delta > 0$. The smallest negative effect is then set automatically to $-\delta$ (this symmetry is not obligatory, but it is the default in *xParallelGroupsTrial.xls* where entering $-\delta$ in C27 as the “threshold value for smallest important or harmful effect” automatically populates the cells where δ is required). The specified δ defines three regions on the real line, as follows: the “negative or harmful” region $(-\infty, -\delta)$, the “trivial” region $(-\delta, \delta)$ inside which there is no effect, and the “positive or beneficial” region (δ, ∞) . The confidence interval is then classified by the extent of overlap with these three regions into one of the four categories, as follows: “positive,” “trivial,” “negative,” or “unclear,” where this last category is used for confidence intervals that do not belong to any of the other categories. The way this was done is illustrated, for example, in Figure 2 of the articles of Batterham and Hopkins (4,5).

Probability calculations: approach 2. “Magnitude-based inference” as implemented in *xParallelGroupsTrial.xls* does not directly compare the confidence interval (equation 1) with the three regions defined by δ but instead bases the classification on new probabilities supposedly associated with each of these three regions. As we will see in the following sections, these quantities are *P* values (from particular tests) and are not obtained directly from the confidence interval.

The three quantities calculated (described as “chances” or “qualitative probabilities” in I135 in *xParallelGroupsTrial.xls*) are the “substantially positive (+ve) or beneficial” value

$$p_b = 1 - G_v\{[\delta - (\bar{Y}_2 - \bar{Y}_1)]/SE(\bar{Y}_2 - \bar{Y}_1)\} \quad [2]$$

computed in L135, the “substantially negative (-ve) or harmful” value

$$p_h = G_v\{[-\delta - (\bar{Y}_2 - \bar{Y}_1)]/SE(\bar{Y}_2 - \bar{Y}_1)\} \quad [3]$$

computed in L139, and the “trivial” value $1 - p_b - p_h$ computed in L137. In these expressions, G_v is the distribution function of the Student’s *t* distribution with *v* degrees of freedom. The values p_b , p_h , and $1 - p_b - p_h$ are interpreted (in L136, L140, and L138) against a seven-category scale of “most unlikely,” “very unlikely,” “unlikely,” “possibly,” “likely,” “very likely,” and “most likely,” as shown in Table 1. Note that the definitions of the categories are not always the same (0.01 and 0.99 are sometimes used instead of 0.005 and 0.995, see for example Batterham and Hopkins (4)) and the words attached to the interpretation are not always the same (“almost certainly not” and “almost certainly” are sometimes used instead of “most unlikely” and “most likely,” see for example Batterham and Hopkins (4) and Hopkins et al. (14)).

We describe these categories (in L136, L140, and L138) as the status of the value and refer to the descriptions of p_b , $1 - p_b - p_h$, and p_h as the beneficial, trivial, and harmful status, respectively.

The next step requires us to specify threshold values against which to compare p_b and p_h . Hopkins (12) and Hopkins et al. (14) discuss two kinds of “magnitude-based inference,” namely, “clinical inference” and “mechanistic inference.” For “clinical inference,” we have to specify the “minimum chance of benefit” (default $\eta_b = 0.25$ in E37) and the “maximum risk of harm” (default $\eta_h = 0.005$ in E36). For “mechanistic inference,” there is no direct clinical or practical application and positive and negative values represent equally important effects, so a single value is required (default $\alpha/2 = 0.05$, obtained by setting $\eta_b = \eta_h = 0.05$). In practice, the threshold values for the two types of study are used in the same way, so the key practical distinction is between possibly unequal and equal threshold values. In either type of study, we classify the data as supporting one of the four conclusions shown in Table 2. The classifications “beneficial,” “harmful,” and “trivial” are qualified in L141 and L142 by the corresponding classifications of p_b , p_h , and $1 - p_b - p_h$.

To see how the calculations work, we ran them through the spreadsheet *xParallelGroups.xls* and our own R code using the post1 – pre2 example data given in *xParallelGroups.xls*. We report the results for the analysis on the raw scale. The 90% confidence interval for the difference of the means is -0.3 to 14 ; the *P* value for testing the null hypothesis that the difference of the means is zero (so the means are the same) rounds to 0.12 . Both these calculations show that there is only weak evidence of a treatment effect. For $\delta = 4.41$ (which corresponds to 0.2 SD, one of the suggested default values, entered into cell C27 as -4.41) and the default values for $\eta_b = 0.25$ and $\eta_h = 0.005$ in the spreadsheet, the comparison of the post1 – pre2 measurements in the experimental and control groups in the example data produces $p_b \approx 0.72$, $p_h \approx 0.01$, and $1 - p_b - p_h \approx 0.27$ (*xParallelGroups.xls* gives $1 - p_b - p_h \approx 0.28$ because it handles the rounding differently), so the default “mechanistic inference” is “possibly beneficial” and the default “clinical inference” is “unclear, get more data.” We give a brief explanation of how these conclusions are reached from Tables 1 and 2. For default “mechanistic inference,” we have $p_b > 0.05$ and $p_h < 0.05$ so the Table 2 classification is positive. Because p_b is classified already as possibly positive according to Table 1, the

TABLE 2. ‘Clinical inference based on threshold chances of harm and benefit’ as specified in *xParallelGroups.xls*.

Range of p_b	Range of p_h	Report
$\eta_b < p_b$	$\eta_h < p_h$	‘Unclear, get more data’
$\eta_b < p_b$	$p_h < \eta_h$	‘Positive’
$p_b < \eta_b$	$\eta_h < p_h$	‘Negative’
$p_b < \eta_b$	$p_h < \eta_h$	‘Trivial’

η_b is the ‘minimum chance of benefit’ (default $\eta_b = 0.25$), and η_h is the ‘maximum risk of harm’ (default $\eta_h = 0.005$). To carry out ‘mechanistic inference,’ set $\eta_h = \eta_b$ (default = 0.05).

“mechanistic inference” inherits “possibly” and is reported as “possibly positive.” For default “clinical inference,” we have $p_b > \eta_b = 0.25$ and $p_h > \eta_h = 0.005$, so the Table 2 classification is “unclear, get more data.” If we change η_h and/or η_b , we do not change the probabilities p_b , p_h , or $1 - p_b - p_h$, but we may change their classification. For example, if we increase η_h from 0.005 to 0.05 (by changing E36 to 5), the “clinical inference” changes to “possibly beneficial.”

We find it helpful for understanding how the probabilities p_b , p_h , and $1 - p_b - p_h$ are being used to look at a graphical representation of the classification schemes used in “magnitude-based inference.” The three probabilities p_b , p_h , and $1 - p_b - p_h$ add up to one (only two of them are needed to determine the third), so they can be plotted in a triangle (called a ternary plot) together with the regions corresponding to the four possible conclusions presented in Table 2. The solid point in the lower left corner of the triangle represents the values of p_b , p_h , and $1 - p_b - p_h$ computed using the post1 – pre2 data in *xParallelGroups.xls*. As the point lies in the beneficial region, the “clinical inference” for $\eta_h = 0.05$ is “beneficial.” The underlying gray grid (representing the threshold values from Table 1) refines this to “possibly beneficial.” Note that changing η_h from 0.05 to 0.005 changes the regions by moving the edge of the beneficial region closer to the left hand side of the triangle, and, in this case, the point is in the “unclear” region (note that in L142, η_b is hard-coded to 0.25). For “mechanistic inference,” we set $\eta_b = \eta_h = 0.05$, which corresponds to moving the boundary of the harmful region toward the right hand side of the triangle (to the gray $p_b = 0.05$ line) and makes the beneficial and harmful regions symmetric (note that in L141, η_b and η_h are hard-coded to 0.05).

In summary, the confidence interval gives an estimate of the treatment effect and its uncertainty and shows that there is only weak evidence of a beneficial treatment effect. Formally, the confidence interval and the P value show that the treatment effect is not significant. “Magnitude-based inference” produces the more optimistic conclusion that there is evidence of a possibly beneficial treatment effect. Is this “magnitude-based inference” conclusion meaningful, and should we use it?

INTERPRETATION

The confidence interval in equation 1 is a standard confidence interval for the Behrens–Fisher problem (e.g., Snedecor and Cochran (21) and has the usual interpretation, as follows: if we draw a very large number of samples independently from the normal model and we compute a confidence interval like equation 1 for $\mu_2 - \mu_1$ from each sample, then $100(1 - \alpha)\%$ of the confidence intervals will contain $\mu_2 - \mu_1$. This is a frequentist interpretation because the level $[100(1 - \alpha)\%]$ is derived from the sampling distribution of $\bar{Y}_2 - \bar{Y}_1$ and interpreted in terms of repeated samples. Precision and care are needed in the definition of a confidence interval, and attempts to give “informal” or

“friendly” working definitions are almost inevitably not correct.

The graphical classification based on the confidence interval as showing evidence of “negative,” “trivial,” “positive,” or “unclear” effects according to its relation to regions defined in the parameter space is used for “explaining” “magnitude based inference,” but Batterham and Hopkins (5) describe it as “crude,” do not recommend using it, and do not implement it in *xParallelGroupsTrial.xls*, instead preferring to base the conclusion on the values p_b , p_h , and $1 - p_b - p_h$.

The interpretation of the values p_b , p_h , and $1 - p_b - p_h$ is quite complicated. Different interpretations of these values are given, sometimes in the same article. For example, Hopkins (12) states that

“The calculations are based on the same assumption of a normal or t sampling distribution that underlies the calculation of the P value for these statistics.”

and

“Alan Batterham and I have already presented an intuitively appealing vaguely Bayesian approach to using the confidence interval to make what we call magnitude-based inferences.”

The first statement claims that the values have a frequentist sampling theory interpretation (it is interesting that it refers to P values rather than to confidence intervals), whereas the second claims that they have a “vaguely Bayesian” interpretation. These statements both need careful analysis.

From our calculation presented in the Appendix (see Appendix, Supplemental Digital Content 3, <http://links.lww.com/MSS/A431>), when $\delta = 0$, p_h is the one-sided P value for testing the null hypothesis that $\mu_2 - \mu_1 = 0$ against the alternative that $\mu_2 - \mu_1 > 0$ and $p_b = 1 - p_h$, so the third probability $1 - p_b - p_h = 0$. Similarly, p_b is the one-sided P value for testing the null hypothesis that $\mu_2 - \mu_1 = 0$ against the alternative that $\mu_2 - \mu_1 < 0$. Thus, if we let p be the two-sided P value, when $\delta = 0$, we have $p_b = 1 - p/2$ and $p_h = p/2$. The switch from two-sided to one-sided P values and the relation to “magnitude-based inference” terminology are important; the small p case corresponds to both a small “risk of harm” p_h and a large “chance of benefit” p_b . If $\delta > 0$, we can interpret p_h as the one-sided P value for testing the null hypothesis that $\mu_2 - \mu_1 = -\delta$ against the alternative that $\mu_2 - \mu_1 > -\delta$. Similarly, p_b can be interpreted as the one-sided P value for testing the null hypothesis that $\mu_2 - \mu_1 = \delta$ against the alternative that $\mu_2 - \mu_1 < \delta$. Starting from P values leads to an interpretation in terms of tests and shows that “magnitude-based inference” has not replaced tests by confidence intervals but is actually based on tests and can itself be regarded as a test. As we increase δ , the effect is to increase $1 - p_b - p_h$ and eventually decrease both p_b and p_h . For a P value in the range 0.05–0.15, this shifts the analysis toward a positive conclusion; we decrease the “risk of harm,” p_h , at the cost of also decreasing the “chance of success,” p_b , but usually not by enough to lose the “evidence” for a positive

effect (given that η_b is kept small). Because $\eta_b = 0.25$ is relatively small (compared with, say, 0.95), the important threshold for obtaining a positive result is actually η_h . This is shown by the curve drawn along the left hand side of the triangle to the base in Figure 1 to show how p_b , p_h , and $1 - p_b - p_h$ change as δ changes. The cross on the base of the triangle corresponds to $\delta = 0$ when p_h equals half the usual P value and represents the weakest evidence of a positive effect; increasing δ initially strengthens the evidence of a beneficial effect but eventually makes the evidence trivial. If the threshold values are not well calibrated, we can also strengthen the evidence by changing the threshold values (particularly by increasing η_h).

Alternatively, we can try to interpret p_b and p_h as quantities derived from the confidence interval, as shown in equation 1. Starting from a confidence interval actually leads naturally to a Bayesian rather than a frequentist interpretation for p_b and p_h . In the Bayesian framework, we need to make $\mu_2 - \mu_1$ a (nondegenerate) random variable with a (prior) distribution specified before collecting the data. The data are combined (using the laws of probability) with the prior distribution to produce the conditional distribution of $\mu_2 - \mu_1$ given the data, which is called posterior distribution. If we adopt the improper prior distribution with

probability density function, $g(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto 1/\sigma_1^2 \sigma_2^2$, then, given the data, $[\mu_2 - \mu_1 - (\bar{Y}_2 - \bar{Y}_1)]/SE(\bar{Y}_2 - \bar{Y}_1)$ has the Behrens–Fisher distribution. The prior distribution is improper because its integral is not finite so it cannot be standardized (like a proper probability density function) to have integral one; the Behrens–Fisher posterior distribution is a proper distribution and hence can be used in the usual way to compute posterior probabilities. In fact, the Behrens–Fisher distribution is not particularly tractable and it is often approximated by simpler distributions. If we approximate the Behrens–Fisher distribution by the Student’s t distribution with ν degrees of freedom, the expressions equations 2 and 3 can be rearranged as

$$p_b = \Pr\{[\mu_2 - \mu_1 - (\bar{Y}_2 - \bar{Y}_1)]/SE(\bar{Y}_2 - \bar{Y}_1) \geq [\delta - (\bar{Y}_2 - \bar{Y}_1)]/SE(\bar{Y}_2 - \bar{Y}_1) | \text{data}\} \\ = \Pr(\mu_2 - \mu_1 \geq \delta | \text{data})$$

and, similarly,

$$p_h = \Pr(\mu_2 - \mu_1 \leq -\delta | \text{data}).$$

That is, p_b and p_h can be interpreted as approximate posterior probabilities under a specific choice of prior distribution that the difference in population means is greater/less than $\delta/-\delta$, respectively. Both choices, the specific prior distribution and the approximation to the Behrens–Fisher distribution, can be replaced by other choices.

Batterham and Hopkins (5) state that

“The approach we have presented here is essentially Bayesian but with a ‘flat prior’; that is, we make no prior assumption about the true value.”

The improper prior used in the analysis is an example of a vague prior. A vague prior does not impose strong assumptions about the unknown parameters on the analysis. This does not mean that it imposes no assumptions because, in fact, it imposes a quite definite assumption. Moreover, as Barker and Schofield (3) carefully explained, the appearance of imposing only vague information is dependent on the scale on which we look at the parameters because a prior on one scale actually imposes strong information on some functions of the parameters. If we take “flat” to mean “vague,” the prior with probability density function $g(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto 1/\sigma_1^2 \sigma_2^2$ is literally flat or constant if we transform the variances to log variances but it is not flat on the variance scale. This explicitly shows that the information depends on the scale of the parameters.

In response to Barker and Schofield (3), Hopkins and Batterham (13) dismissed what they refer to as “an imaginary Bayesian monster.” However, what Barker and Schofield (3) wrote is correct. It is not possible to squash a prior flat on the real line while maintaining an area of unity. Both the mean and the variance are infinite, so it is not correct to write that “the mean of a flat prior may as well be zero.” It is also not correct to claim that “All values of the statistic from minus infinity to plus infinity are therefore equally infinitesimally likely—hence the notion of no assumption about the true value.” The

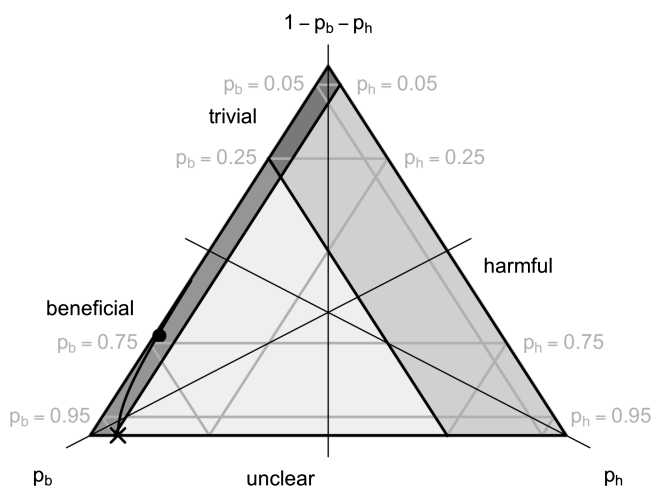


FIGURE 1—Ternary plot of the probabilities p_b , p_h , and $1 - p_b - p_h$ showing the four regions corresponding to the different possible conclusions “beneficial,” “trivial,” harmful,” and “unclear” when $\eta_b = 0.25$ and $\eta_h = 0.05$. The threshold values from Table 1 are represented by gray lines. Note that the 0.005 and 0.995 lines are not actually visible because they are very close to the side of the triangle and the vertex of the triangle, respectively; the lines we can see represent the probabilities 0.05, 0.25, 0.75, and 0.95. The gray p_b labels on the left hand edge of the triangle are for the lines running parallel to the right hand side, and the gray p_h labels on the right hand edge of the triangle are for the lines running parallel to the left hand side. The horizontal lines for $1 - p_b - p_h$ are drawn in but not labeled to reduce clutter. We have also partitioned the triangle into the regions specified in Table 2 using the threshold values $\eta_b = 0.25$ and $\eta_h = 0.05$ (we use $\eta_h = 0.05$ rather than the default 0.005 to make the region visible.) The regions are shaded to make them easier to distinguish. The region labels are written outside the triangle adjacent to the region. The black point represents values of p_b , p_h , and $1 - p_b - p_h$ (from the example in the spreadsheet), which lead to the conclusion “possibly beneficial.” The cross on the base represents the values of p_b , p_h , and $1 - p_b - p_h$ when $\delta = 0$, and the curve through the cross and black point shows the effect of changing δ on p_b , p_h , and $1 - p_b - p_h$.

distribution is actually that of a parameter rather than a statistic as claimed, and the flatness is not equivalent to making no assumption about the true value. Taking the limit of a posterior on the basis of a proper prior as the prior becomes improper does not correspond to using “no prior real information about the true value” any more than using the corresponding improper prior does. The “empirical evidence” based on bootstrapping presented by Hopkins and Batterham (13) is not relevant to the argument.

As we noted, the calculations implemented in the spreadsheet can be interpreted as approximating the Behrens–Fisher distribution by the Student’s t distribution with ν degrees of freedom. Patil (18) states that this is not a satisfactory approximation; other approximations have been provided by Cochran (8), Patil (18), and Molenaar (16). Although the different approximations often give similar results, this means that the approximation used in the spreadsheet is not the one that you would choose to use for a Bayesian analysis. This suggests that the Bayesian interpretation was not intended in the original formulation.

One of the consequences of the fact that p_b and p_h are not directly related to the confidence interval, equation 1, is that conclusions based on p_b and p_h can sometimes seem unsatisfactory when compared with the confidence interval. The conclusion is determined solely by which region the point $(p_b, p_h, 1 - p_b - p_h)$ falls into. This shows that the conclusion is based on a type of hypothesis test. It is not a standard frequentist significance or hypothesis test (this would treat fewer hypotheses and allow fewer outcomes) or a standard Bayesian hypothesis test (this would choose between the hypotheses that the difference in population means is positive, trivial, or negative by adopting the hypothesis with the largest posterior probability) because of the additional requirements imposed by the fixed threshold probabilities η_b and η_h . Nonetheless, it has much more to do with hypothesis testing than interval estimation, showing that hypothesis testing has been replaced by a different kind of test rather than been avoided. This is inevitable when, as in “magnitude-based inference,” the outcome of the analysis is one of a simple set of possible categories.

One advantage of recognizing that “magnitude-based inference” is a type of test is that we can evaluate its properties as a test. In particular, we can compute the probability of reaching beneficial, harmful, or trivial conclusions by simulating 10,000 samples from the model, computing p_b , p_h , and $1 - p_b - p_h$ for each sample and calculating the proportion of times these probabilities lead to each of the conclusions of interest. The results of performing this simulation for the case $\mu_2 - \mu_1 = 0$ (so $\mu_2 = \mu_1$ and there is no effect) for samples with similar other characteristics to the example data ($n_1 = n_2 = 20$, $\sigma_1^2 = 15^2$, $\sigma_2^2 = 11^2$), for a fine grid of δ values and for choices of η_b and η_h , are shown in Figure 2. The probability of finding a beneficial effect equals η_b when $\delta = 0$ increases as we increase δ until it starts decreasing and eventually decays to zero. The probability of finding a harmful effect behaves similarly but starts at η_h when $\delta = 0$. The probability of finding a trivial effect (the correct answer because the simulation is for the case $\mu_2 - \mu_1 = 0$) equals zero for small δ and then increases to one as δ increases. Other than at $\delta = 0$, the probabilities of the different conclusions are not simply related to η_b or η_h because they also depend on δ . The vertical dashed gray line corresponds to $\delta = 4.418$, the value used in our analysis. Using this δ for default “mechanistic inference” ($\eta_b = \eta_h = 0.05$), we find that the probability of finding an effect (beneficial or harmful) when there is no effect is 0.54. That is, the probability of a Type I error is more than 10 times the standard value of 0.05. This increase in the probability of a Type I error explains why “magnitude-based inference” is less conservative than a standard test; it is equivalent to using the usual P value with a 0.5 threshold, an increase that is unlikely to be acceptable. Similarly, for default “clinical inference” ($\eta_b = 0.25$, $\eta_h = 0.005$), we find that the probability of finding a beneficial effect is 0.057 and the probability of finding a harmful effect is 0.657 when there is no effect. That is, the probability of a Type I error is 0.714. Increasing η_h to 0.05 increases the probability of finding a beneficial effect to 0.255 and slightly decreases the probability of finding a harmful effect to 0.647 when there is no effect (so the probability of a Type I error is 0.902). These results may

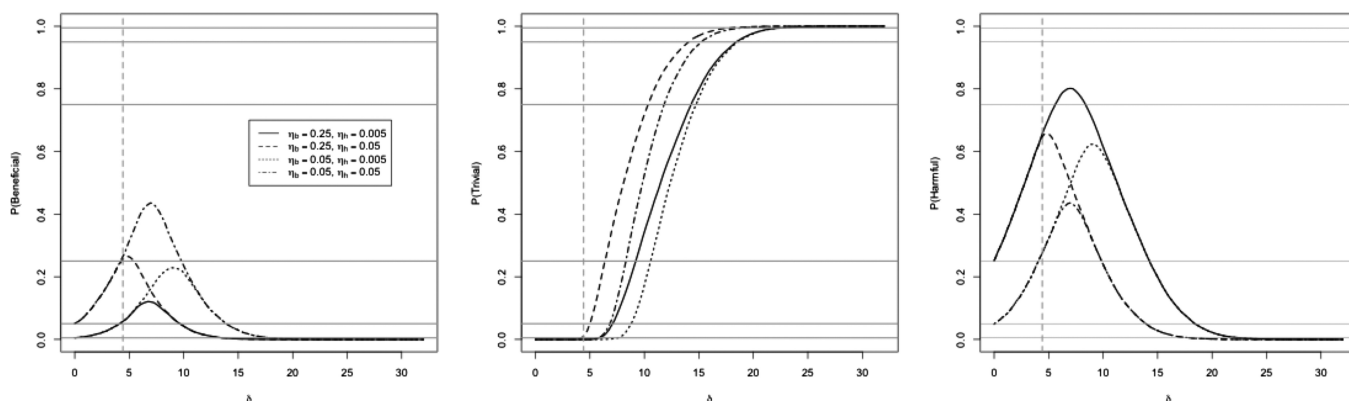


FIGURE 2—Plots of the probabilities of finding beneficial, trivial, or harmful effects as functions of δ for four values of (η_b, η_h) when there is no effect. The 10,000 data sets were simulated to have $\mu_2 - \mu_1 = 0$, with similar other characteristics to the example data ($n_1 = n_2 = 20$, $\sigma_1^2 = 15^2$, $\sigma_2^2 = 11^2$). The vertical dashed gray line corresponds to $\delta = 4.418$, the value used in our analysis.

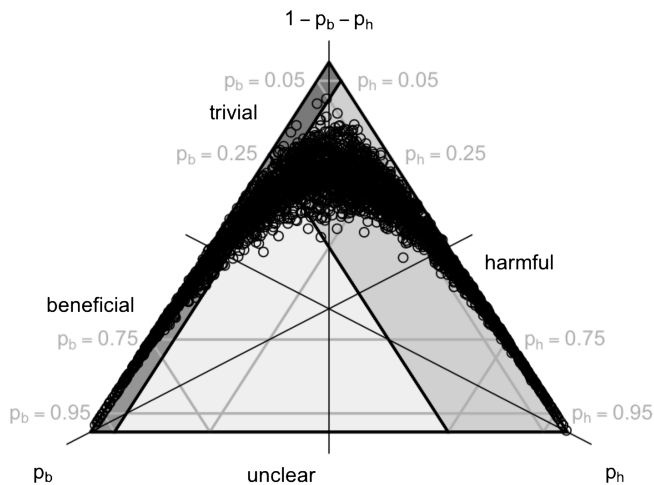


FIGURE 3—Ternary plot showing the distribution of 3000 realizations of the triple p_b , p_h , and $1 - p_b - p_h$ when $\delta = 4.418$. The data were generated in the same way as the data used in Figure 2.

be easier to understand if we plot a random sample of p_b , p_h , and $1 - p_b - p_h$ triples on a ternary diagram. When $\delta = 0$, the points are distributed uniformly along the base of the triangle. As δ increases, the points are distributed along and around a curve; Figure 3 shows the distribution for $\delta = 4.418$. If we continue to increase δ , the curve moves up the triangle until, eventually, all the points lie on the $1 - p_b - p_h$ vertex. The message from the simulation is that 1) we cannot simply interpret η_b and η_h as frequentist thresholds that directly describe standard properties of the test and 2) the probability of a Type I error (finding an effect that is not there) is surprisingly high.

In summary, “magnitude-based inference” is based on testing rather than interval estimation. It does not fit neatly into either the standard frequentist or the standard Bayesian testing frameworks. Using confidence intervals or moving to a full explicit Bayesian analysis would resolve the difficulties of justifying “magnitude-based inference.” However, for a convincing Bayesian analysis, the prior distribution needs to be well justified (for example, being based on solid empirical evidence).

SAMPLE SIZE CALCULATIONS

The second spreadsheet *xSampleSize.xls* (see spreadsheet, Supplemental Digital Content 2, <http://links.lww.com/MSS/A430>) provides various sample size calculations; we discuss only the calculations presented for the problem of comparing two means described previously. The sample size calculations implemented in *xSampleSize.xls* are actually for the simpler equal variance case $\sigma_1^2 = \sigma_2^2 = \sigma^2$ rather than the full Behrens–Fisher problem. In the equal variance case, the distribution theory is much simpler; it is based on the Student’s t distribution, which is simpler than the Behrens–Fisher distribution (so no approximation is required), and the degrees of freedom is a function of n_1 and n_2 but not of σ^2 and hence does not have to incorporate estimates

of σ^2 (see Appendix, Supplemental Digital Content 3, <http://links.lww.com/MSS/A431>).

Let r_2 be the proportion of observations in the second group. Then, we can write $n_2 = nr_2$ and $n_1 = n(1 - r_2)$ (so the sample size is n) and the variance of $\bar{Y}_2 - \bar{Y}_1$ is

$$\text{Var}(\bar{Y}_2 - \bar{Y}_1) = \sigma^2 \left(\frac{1}{n_2} + \frac{1}{n_1} \right) = \sigma^2 \left(\frac{1}{nr_2} + \frac{1}{n(1-r_2)} \right) = \frac{\sigma^2}{nr_2(1-r_2)}.$$

The standard frequentist sample size calculation for this problem (e.g., Snedecor and Cochran (21)) is derived by working out the value of n that we require to carry out a two-sided hypothesis test with the probability of a Type I error (that we reject the null hypothesis when it is correct) equal to the level α and the probability of a Type II error (that we accept the null hypothesis when it is false) equal to β (so the power is $1 - \beta$) when the true difference between the means is $\mu_1 - \mu_2 = \delta$. The value of δ is usually taken to be the smallest meaningful difference between μ_2 and μ_1 . The smallest meaningful difference is the minimum size of the difference that is scientifically or clinically important; this is the reason we have used δ as before and not introduced a new symbol. Standard calculations lead to the equation

$$n = \frac{\sigma^2 [G_{n-2}^{-1}(1 - \alpha/2) + G_{n-2}^{-1}(1 - \beta)]^2}{r_2(1-r_2)\delta^2}. \quad [4]$$

(For a one-tailed test, replace $G_{n-2}^{-1}(1 - \alpha/2)$ by $G_{n-2}^{-1}(1 - \alpha)$.) Because n appears on both sides of this equation, we need to solve it by successive approximation. That is, we start with an initial value n_0 , substitute it into the right hand side of the equation to compute n , replace n_0 by n , and repeat the process a few times or until it converges, meaning that the value of n stops changing between iterations.

In fact, the spreadsheet *xSampleSize.xls* implements a different calculation for “sample size for statistical significance.” It starts with $n_0 = 22$ and makes three iterations to calculate n in I103 from

$$n = \frac{2\sigma^2 [G_{n-2}^{-1}(1 - \alpha/2) + G_{n-2}^{-1}(1 - \beta)]^2}{r_2(1-r_2)\delta^2}. \quad [5]$$

The factor 2 in the numerator is not needed because it is incorporated into $r_2(1 - r_2)$; its effect is to make the sample size twice as large as it should be (as calculated from equation 4).

Batterham and Hopkins (5) state that “Studies designed for magnitude-based inferences will need a new approach to sample size estimation based on acceptable uncertainty,” but they do not derive a new approach. As before, we treat what is implemented in the spreadsheet *xSampleSize.xls* as definitive. The calculation starts with $n_0 = 12$ and makes four iterations to calculate n in I34 from

$$n = \frac{2\sigma^2 [G_{n-2}^{-1}(1 - \eta_h) + G_{n-2}^{-1}(1 - \eta_b)]^2}{r_2(1-r_2)(2\delta)^2}. \quad [6]$$

No derivation for this formula is given, but its similarity to equation 5 is striking, and we think that it has been adapted from equation 5. This belief is strengthened by the

identification of η_h and η_b with acceptable Type I and Type II “clinical error rates” (11). A “Type I clinical error” is using an effect that is harmful, and a “Type II clinical error” is not using an effect that is beneficial. This identification is incorrect, and we cannot equate Type I and Type II errors to “clinical errors.” As we have seen, η_h and η_b are the levels of two different tests and not the level and power of a single test. They affect the performance of the test but are not simply summaries of the performance of the test (because this also depends on δ). We cannot justify taking a formula, making up the quantities it is applied to, and pretending that the result means something.

“Magnitude-based inference” claims to require smaller sample sizes. For example, applying equation 5 with the default values in *xSampleSize.xls* (namely, the “proportion in the second group” $r_2 = 1/2$ the “smallest change” $\delta = 0.05$, Type I error $\alpha = 0.05$, Type II error $\beta = 0.2$, and “within-subject SD (typical error)” $\sigma^2 = 1$), the sample size is $n/2 = 127$ in each group (from equation 4, it should actually be 64). Applying equation 6 with the same settings but with “Type 1 clinical errors,” $\eta_h = 0.005$, and “Type II clinical errors,” $\eta_b = 0.25$, instead of the Type I and Type II errors, the sample size is $n/2 = 44$ in each group. This represents a reduction in the required sample size of $44/127 = 0.35$, which is substantial.

Ignoring the critical issue of whether equation 6 is a valid formula in “magnitude-based inference,” we compare the numerical values equation 6 produces with those obtained from equation 5 and show that the reduction in sample size occurs simply because it is looking for larger effects rather than obtaining a true advantage. The comparison entails comparing both the numerators and denominators in equations 5 and 6, noting that the main important difference occurs in the denominators. The only differences in the numerators on the right hand sides of equations 5 and 6 are in the replacement of $\alpha/2 = 0.025$ by $\eta_h = 0.005$ and $\beta = 0.2$ by $\eta_b = 0.25$. The effect is to increase the numerical value of the numerator; using the default values, the increase in the numerator of equation 6 over the numerator of equation 5 is by the ratio $10.581/7.857 = 1.34 \approx 4/3$. The only difference in the denominators on the right hand sides of equations 5 and 6 is in the replacement of δ by 2δ . This means that equation 6 is effectively dividing the numerator in equation 6 by 4 times the denominator used in equation 5. Combining the increases in the numerator and denominator of equation 6, we see that the overall change in the computed sample size is by a factor $4/3 \div 4 = 1/3$, explaining the claimed reduction of 30% in the sample size (for example, Batterham and Hopkins (5) state that “Sample sizes are approximately one-third of those based on hypothesis testing, for what seems to be reasonably accepted uncertainty”). The sample size for “magnitude-based inference” is 1/3 of the size of that for equation 5 simply because, for no stated reason, the conventional frequentist sample size has been divided by three. That is, it is smaller because it is defined to be smaller. If we are meant to treat 2δ in equation 6 as the minimum

important difference, then this has been made twice as large as that in equation 5. This means that, in its own terms, “magnitude-based inference” is solving easier (rather than the popularly perceived more difficult) problems. It requires smaller sampler sizes not because it is less conservative as hoped but because it is effectively looking for larger effects and, hence, in sample size calculations, is less ambitious about what it is trying to achieve.

Hopkins (11) also discusses the use of a confidence interval method (that he calls method II to distinguish it from method I based on equation 6) to choose the sample size, although this is not implemented in the spreadsheet. The method is to choose the sample size to make the length of the 90% confidence interval less than the length of the trivial region 2δ . For the equal variance case with $n_1 = n_2 = n/2$, this leads to the equation $n = 4\sigma^2 G_{n-2}^{-1}(1-\alpha/2)^2/\delta^2$, which again has to be solved by successive approximation. Hopkins (11) claims that with $\alpha = 0.1$, this gives almost identical results to those of equation 6. For the suggested default values, the numerator is smaller than that in equation 6 by roughly a factor of 2 $(1/2)^2$, so the sample size from method II is roughly half the sample size from the method described in the previous paragraphs (method I). However, because the extra 2 in the numerator of equation 6 should not be there (to the extent that one can say what should and should not be included in a formula adopted without derivation), the sample size from method II should actually be the same as that from method I. Looked at on its own merits, method II is a standard calculation with a proper frequentist interpretation whereas method I is not.

DISCUSSION

The distinction between nonclinical or “mechanistic inference” and “clinical inference” is emphasized in “magnitude-based inference” but not clearly explained, and no specific examples or context for using the one or the other are given. Generally, the clinical and nonclinical descriptions are used informally to distinguish between studies done on people and studies (including animal studies) that are not. This is not what is meant by “mechanistic” and “clinical” in “magnitude-based inference” because both can be applied to studies on people. As far as we can tell, “mechanistic inference” is applied to studies carried out without a specific context or end use in mind so that no distinction is made between positive and negative effects and these are treated as equally important. “Clinical inference” is used in studies wherein there is enough context or a sufficiently clear end use to identify an effect in one direction as beneficial and an effect in the other as harmful and therefore to allow these two directions to be treated asymmetrically. In a sense, “mechanistic inference” is like a two-sided hypothesis test that simply looks for an effect and “clinical inference” is like two one-sided hypothesis tests at different levels looking for an effect in one direction and the absence of an effect in the other. In fact, as we have seen, in “magnitude-based inference,” both problems are treated like

two one-sided hypothesis tests and this is one reason weak (classical) evidence appears as a stronger (“magnitude-based inference”) evidence.

A key criticism explicitly mentioned as motivating “magnitude-based inference” is that “the null hypothesis of no relation or no difference is always false—there are no truly zero effects in nature” (4,5). This is not strictly true and, as pointed out by Barker and Schofield (3), the null hypothesis of no effect can be a useful idealization. However, we believe that the most important overriding motivation is the perception that significance testing with a 5% threshold is too conservative, particularly when looking for small effects in small samples.

Many statisticians choose to report confidence intervals rather than P values not because the 5% threshold for P values is too conservative but because confidence intervals present more information more directly about effects of interest. Confidence intervals and tests are linked in the sense that we can carry out a test either by computing a P value or a confidence interval; whether a confidence interval set at the usual 95% level contains zero or not tells us whether the P value is below or above 5%, effectively enabling us to carry out the same test. The level (like the P value threshold) can be varied, but journals and readers (reasonably) tend to prefer to see the standard values being used.

Although confidence intervals are used as an explicit starting point for “magnitude-based inference,” Batterham and Hopkins (5) argued that confidence intervals are themselves flawed and used this to motivate the development of the additional probabilities p_b , p_h , and $1 - p_b - p_h$:

“We then show that confidence limits also fail and then outline our own approach and other approaches to making inferences based on meaningful magnitudes.”

Batterham and Hopkins (4,5) believe that they fail because the correct interpretation of confidence intervals is too complicated. The interpretation is complicated, but this cannot be avoided in a frequentist analysis. Confidence intervals may or may not be exactly what we want from an analysis, but in frequentist inference, they represent what we can legitimately obtain and no amount of wishful thinking can get around this. Other answers suggested by Batterham and Hopkins (4) revolve around the choice of a standard 95% level:

“We also believe that the 95% level is too conservative for the confidence interval; the 90% level is a better default, because the chances that the true value lies below the lower limit or above the upper limit are both 5%, which we interpret as very unlikely (Hopkins, 2002). A 90% level also makes it more difficult for readers to re-interpret a study in terms of statistical significance.”

That is, like tests, confidence intervals are also seen as too conservative. However, the level of a confidence interval seems to be perceived as somewhat easier to change than the

level of a test. A further concern made explicit in the final sentence is that confidence intervals are too closely related to tests; the fact that confidence intervals can be used to carry out tests is seen as a weakness, regardless of the fact that confidence intervals contain other useful information. Lowering the level of the confidence interval to 90% breaks the link to the usual 5% threshold used in testing and makes the analysis less conservative (i.e., apparently strengthens weak effects) by effectively increasing the P value threshold for significance from 0.05 to 0.10.

In the way “magnitude-based inference” implements “mechanistic inference,” weak evidence for an effect is strengthened by replacing the P value by half the P value (p_h with $\delta = 0$) and then decreasing the P value further by changing the null hypothesis from $\mu_2 - \mu_1 = 0$ to $\mu_2 - \mu_1 = -\delta$, with $\delta > 0$. The standard threshold $\eta_b = \eta_h$ could also be increased to further strengthen the evidence, but this is a more obviously doubtful change for less gain and the default value does not do this. Partly, it is unnecessary to change the threshold because the standard threshold is large enough after the other changes (which are less easy to track) and it is comforting that the thresholds have not apparently changed. “Mechanistic inference” in “magnitude-based inference” does not abandon the use of P values but promotes a complicated and confusing way of bringing about changes that have the same effect as simply increasing the usual P value threshold. If changing a threshold is not acceptable, then redefining the P value to achieve the same effect should not be either. Although we can sympathize with the frustration of the researcher finding that the evidence they have for an effect is weaker than they would like, we have to recognize the limitations of the data and be careful about trying to strengthen weak evidence just because it suits us to do so.

“Clinical inference” in “magnitude-based inference” can be seen as a response to another motivating issue for “magnitude-based inference,” namely, the question of how to use P values for “clinical inference.” Batterham and Hopkins (5) referred to the approach of Guyatt et al. (10) who introduced a threshold for a clinically significant effect and suggested that a trial is “positive and definite” if the lower boundary of the confidence interval is above the threshold and “positive” but needing studies with larger samples if the lower boundary is somewhat below the threshold. This is more stringent than ordinary statistical significance (unless the threshold is zero), so it is not surprising that Batterham and Hopkins (5) wrote, “This position is understandable for expensive treatments in health care settings, but in general we believe it is too conservative.” The other approaches to clinical inference in the literature (see Man-Son-Hing et al. (15) and Altman et al. (1) for reviews) also tend to require more than simple statistical significance so would likely also be deemed too conservative for “magnitude-based inference.”

“Magnitude-based inference” achieves a less conservative “clinical inference” by making the same steps as in “mechanistic inference” and, in addition, changing the thresholds.

The increase in η_b looks spectacular, but this is misleading because η_b is not actually important when the P value is in the range 0.05–0.15 and, although the decrease in η_b works against the other changes (in the P value and δ), the gains from the other two changes have larger effects and outweigh the decrease in η_b . The key question is, if other researchers feel that clinical conclusions should be more conservative than mere statistical significance, should we use a method for clinical inference that is explicitly designed to be less conservative?

Throughout this article, we have proceeded as if the normal model holds exactly and the Behrens–Fisher problem of comparing two sample means is the appropriate problem to treat. This is standard in theoretical discussion of statistical methods and therefore also appropriate for our analysis of “magnitude-based inference,” but it ignores other important statistical issues that arise in practice when we choose an analysis. These include the following: 1) using the structure of the data properly (e.g., in many sport science studies, including in the example data included in the spreadsheet, repeated observations are taken on each subject and there are often more than two groups of subjects so a more flexible analysis than that offered in *xParallelGroupsTrial.xls* is appropriate), 2) including covariates in a more flexible analysis (*xParallelGroupsTrial.xls* allows only a single covariate), 3) choosing the right distribution (binary and count data are best analyzed using more appropriate models), 4) choosing the scale for analysis (data transformation is often but not always needed, and it is largely an empirical question when it is), and 5) choosing an appropriate effect size to present results (when they are available, Cohen standardized effect sizes can be useful for comparing results across studies that have measured different variables or have used different scales of measurement, but, in general, direct unstandardized effect sizes are more meaningful in practice and are easier to interpret (2,24)). These issues are common to all statistical analysis, but the limitations of the spreadsheets may make it easier to overlook them in “magnitude-based inference.”

REFERENCES

- Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. 2nd ed. New York (NY): Wiley; 2000. pp. 1–254.
- Baguley T. Standardized or simple effect size: what should be reported? *Br J Psychol*. 2009;100:603–17.
- Barker RJ, Schofield MR. Inferences about magnitudes of effects. *Int J Sports Physiol Perform* 2008;3:547–57.
- Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Sportsmedicine*. 2005;9:6–13.
- Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 2006;1:50–7.
- Berry G. Statistical significance and confidence intervals. *Med J Aust*. 1986;144:618–9.
- Bulpitt C. Confidence intervals. *Lancet*. 1987;1:494–7.
- Cochran WG. Approximate significance levels of the Behrens–Fisher test. *Biometrics*. 1964;20:191–5.
- Gardner M, Altman D. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J*. 1986;292:746–50.
- Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. *Can Med Assoc J*. 1995;152:169–73.
- Hopkins WG. Estimating sample size for magnitude based inferences. *Sportsmedicine*. 2006;10:63–70.
- Hopkins WG. A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p-value. *Sportsmedicine*. 2007;11:16–20.
- Hopkins WG, Batterham AM. Letter to the editor: an imaginary Bayesian monster. *Int J Sports Physiol Perform*. 2008;3:411–2.
- Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc*. 2009;41(1):3–13.
- Man-Son-Hing M, Laupacis A, O'Rourke K, et al. Determination of the clinical importance of study results. *J Gen Intern Med* 2002;17:469–76.
- Molenaar IW. Simple approximations to the Behrens-Fisher distribution. *J Stat Comput Simulation*. 1979;9:283–8.

CONCLUSIONS

We have given a precise description of “magnitude-based inference” for the problem of comparing two means and discussed its interpretation in detail. “Magnitude-based inference” begins with the computation of a confidence interval (that has the usual frequentist interpretation in terms of repeated samples). We show that the calculations can be interpreted either as P values for particular tests or as approximate Bayesian calculations, which lead to a type of test. In the former case, this means that 1) the “magnitude-based inference” calculations are not derived directly from the confidence interval but from P values for particular tests and 2) “magnitude-based inference” is less conservative than standard inference because it changes the null hypothesis and uses one-sided instead of two-sided P values. The inflated level of the test means that it should not be used. Finally, the sample size calculations should not be used. Rather than use “magnitude-based inference,” a better solution is to be realistic about the limitations of the data and use either confidence intervals or a fully Bayesian analysis.

We thank participants at an informal workshop held at the Australian Institute of Sport for contributing experiences, thoughts, and ideas on using “magnitude-based inference.” We thank Nic West for discussions about how “magnitude-based inference” is used by sports scientists. We are grateful for comments on earlier drafts by Robert Clark, Peter Lane, and Jeff Wood. We thank Lingbing Feng for his assistance in converting our manuscript from a LaTeX .tex file to a MS Word .doc file. We thank the editor and four reviewers for their detailed comments on an earlier version of this article.

This research was supported by funding from the Australian Institute of Sport.

The authors declare that there are no conflicts of interest in undertaking this study.

The results of the present study do not constitute endorsement by the American College of Sports Medicine.

17. Northridge ME, Levin B, Feinleib M, Susser MW. Editorial: statistics in the journal—significance, confidence, and all that. *Am J Public Health*. 1997;87:1092–5.
18. Patil VH. Approximations to the Behrens-Fisher distributions. *Biometrika*. 1965;52:267–71.
19. R Core Team. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing, ISBN 3-900051-07-0. Available from: <http://www.R-project.org/> [cited 2012 Oct 21].
20. Rothman K. A show of confidence. *N Engl J Med*. 1978;299:1362–3.
21. Snedecor GW, Cochran WG. *Statistical Methods*. 7th ed. Ames (IA): Iowa State University Press; 1980. pp. 98–105.
22. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika*. 1937;29:350–62.
23. Welsh AH. *Aspects of Statistical Inference*. New York (NY): Wiley; 1996. pp. 150–2.
24. Wilkinson L. Statistical methods in psychology journals: guidelines and explanations. *Am Psychol*. 1999;54:594–604.
25. Wilkinson M. Testing the null hypothesis: the forgotten legacy of Karl Popper? *J Sports Sci*. 2013;31:919–20.
26. Wilkinson M. Distinguishing between statistical significance and practical/clinical meaningfulness using statistical inference. *Sports Med*. 2014;44:295–301.