Sports Med (2016) 46:1563–1573
DOI 10.1007/s40279-016-0517-x

ORIGINAL RESEARCH ARTICLE

Error Rates, Decisive Outcomes and Publication Bias with Several Inferential Methods

Will G. Hopkins[1] · **Alan M. Batterham**[2]

Published online: 12 March 2016

Abstract

Background Statistical methods for inferring the true magnitude of an effect from a sample should have acceptable error rates when the true effect is trivial (type I rates) or substantial (type II rates).

Objective The objective of this study was to quantify the error rates, rates of decisive (publishable) outcomes and publication bias of five inferential methods commonly used in sports medicine and science. The methods were conventional null-hypothesis significance testing [NHST] (significant and non-significant imply substantial and trivial true effects, respectively); conservative NHST (the observed magnitude is interpreted as the true magnitude only for significant effects); non-clinical magnitude-based inference [MBI] (the true magnitude is interpreted as the magnitude range of the 90 % confidence interval only for intervals not spanning substantial values of the opposite sign); clinical MBI (a possibly beneficial effect is recommended for implementation only if it is most unlikely to be harmful); and odds-ratio clinical MBI (implementation is also recommended when the odds of benefit outweigh the odds of harm, with an odds ratio >66).

Methods Simulation was used to quantify standardized mean effects in 500,000 randomized, controlled trials each for true standardized magnitudes ranging from null through marginally moderate with three sample sizes: suboptimal (10 + 10), optimal for MBI (50 + 50) and optimal for NHST (144 + 144).

Results Type I rates for non-clinical MBI were always lower than for NHST. When type I rates for clinical MBI were higher, most errors were debatable, given the probabilistic qualification of those inferences (unlikely or possibly beneficial). NHST often had unacceptable rates for either type II errors or decisive outcomes, and it had substantial publication bias with the smallest sample size, whereas MBI had no such problems.

Conclusion MBI is a trustworthy, nuanced alternative to NHST, which it outperforms in terms of the sample size, error rates, decision rates and publication bias.

Key Points

Null-hypothesis significance testing (NHST) is increasingly criticized for its failure to deal adequately with conclusions about the true magnitude of effects in research on samples.

A relatively new approach—magnitude-based inference (MBI)—provides up-front, comprehensible, nuanced uncertainty in effect magnitudes.

In simulations of randomized, controlled trials, MBI outperforms NHST in respect of inferential error rates, rates of publishable outcomes with suboptimal sample sizes and publication bias with such samples.

Electronic supplementary material The online version of this article (doi:10.1007/s40279-016-0517-x) contains supplementary material, which is available to authorized users.

✉ Will G. Hopkins
will@clear.net.nz

[1] Institute of Sport Exercise and Active Living, Victoria University, Melbourne, VIC, Australia

[2] Health and Social Care Institute, Teesside University, Middlesbrough, UK

1 Introduction

Biomedical researchers study effects on health, performance or other measures of interest in a sample drawn from a population. Statistical inference is the process by which researchers use data from the sample to draw a conclusion about the effect in the population—a conclusion that will be useful or applicable to practitioners and other researchers working with other individuals or samples drawn from that population. In plain language, statistical inference tells us something about the real or true effect, not just the sample effect. The real or true effect is the value that a researcher would expect to get from a very large sample, assuming no biases in the methods of sampling, measurement and analysis.

The traditional approach to inference is the null-hypothesis significance test (NHST), which is aimed at claiming whether the population effect could be null or zero. Generations of researchers have been critical of NHST [1–5], and problems with its use appear regularly even in top journals [6, 7]. Our own dissatisfaction with NHST led us to propose an alternative—magnitude-based inference (MBI)—which is aimed at drawing conclusions about the probability that the population effect is substantial or trivial rather than null [8, 9]. MBI is a simple variety of Bayesian inference, which has been independently proposed by others as a solution to the problems of NHST [10, 11]. The last decade has seen an upsurge in the use of MBI by the community of researchers in sports medicine and science, judging by the nearly 2000 citations of articles promoting MBI in the Google Scholar database. However, in a recent critique published in one of our major journals, the authors advised researchers against using MBI [12]. In this article, we provide evidence to dismiss the critique and to reassure researchers in our disciplines that MBI is superior to NHST.

No sample represents a population exactly, so any inference about the population value of an effect based on a sample can be wrong. An inference that the population value is substantial, clinically important, real or otherwise non-trivial—when, in reality, it is trivial—represents a false-positive or so-called type I error, whereas a false-negative or type II error occurs when a trivial true value is inferred to be non-trivial. A good inferential method should have a low type I rate if the population value is trivial and a low type II rate if the population value is substantial. The error rates with MBI have been explained and quantified to a limited extent [13, 14], but the authors of the recent critique of MBI claimed that the approach suffered from apparently high rates of type I error. They and others [15] also asserted that MBI had a questionable theoretical foundation. In this article, we explain the error rates in detail and extensively quantify the error rates in NHST and MBI. We show that the type I error rates in the non-clinical version of MBI are much lower than those in NHST, and that the rates of other errors in the non-clinical and clinical versions of MBI are generally lower and otherwise acceptable. We also provide published evidence of the sound theoretical basis of MBI [10, 11, 16] and show that MBI has important advantages over NHST: more intuitive interpretation, smaller required sample sizes, higher rates of publication-worthy findings and less publication bias.

2 Methods

2.1 Inferences and Inferential Errors with NHST

For a valid head-to-head comparison of NHST and MBI, we need definitions of type I (false-positive) and type II (false-negative) error rates that can be applied to both approaches. First, we revisited the two major frequentist schools of inference: Fisher and Neyman–Pearson. Fisher devised the *P* value—calculated from the observed data in a single experiment—as an index of the strength of evidence against the null hypothesis, but he ridiculed concepts of false-positive and false-negative errors as "absurdly academic" [17]. Neyman and Pearson's framework requires the specification of a precise alternative hypothesis, and they defined type I and type II error rates as the probabilities of rejecting a true null hypothesis and rejecting a true alternative hypothesis, respectively. However, these definitions relate to 'long-run' error rates, specified in advance and designed to limit the number of incorrect decisions made over ongoing repeated experiments [18]. The Neyman–Pearson definitions of error rates are therefore relevant—for example, to quality control in industrial settings—but not to any single study. Indeed, Schneider argued that "scientific settings suitable for Neyman–Pearson's model seem restricted" (see page 428 in reference [18]). Since the major frequentist schools of inference are unhelpful in this context, below we present and justify definitions of these errors that, in our experience, reflect contemporary custom and practice in biomedical research. The definitions permit valid, pragmatically relevant comparison with the error rates in MBI.

An inference in NHST is a conclusion about whether or not the effect is substantial. In support of this assertion, consider that the sample size in NHST is determined by the desire to have an 80 % chance of obtaining statistical significance when the true effect has the smallest important value. Statistical significance with this arguably optimal sample size therefore implies that the effect is substantial or, in popular parlance, that "there is a real effect".

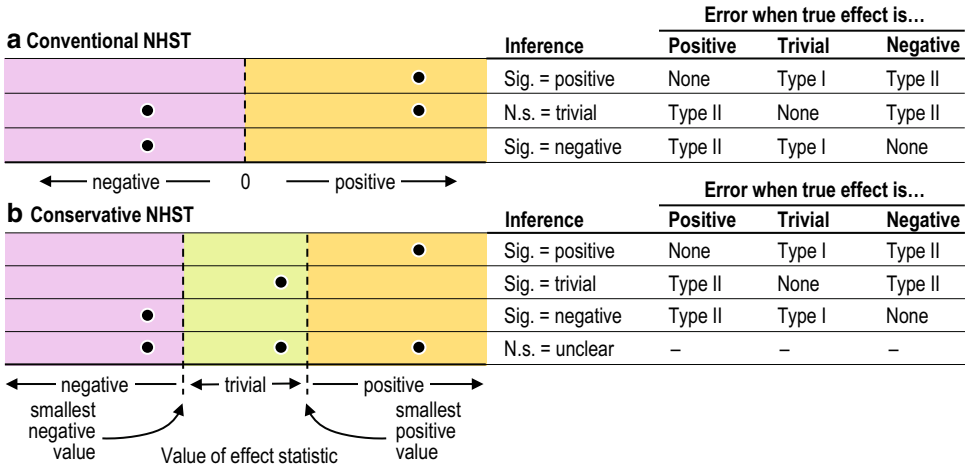
Statistical non-significance implies that the effect is not substantial and is therefore presumably trivial, although it is often reported as "no effect". A type I error occurs when a true null effect is declared significant. Any trivial true effect declared significant must also be a kind of false-positive error and is therefore logically a type I error. The type II error occurs when non-significance is obtained for a substantial true effect. Significance for an observed effect of a sign opposite to that of the true effect is also a false-negative finding about the true effect, so we have also labelled such errors as type II. These errors are illustrated in Fig. 1 for what we call the conventional approach to inference with NHST.

Conventional NHST appears to be a reasonable decision-making process for studies performed with the optimal sample size, but, as we will see, the interpretation of non-significant as insubstantial leads to high type II error rates with suboptimal sample sizes, while significant interpreted as substantial leads to high type I error rates with supra-optimal sample sizes. In an attempt to mitigate these problems, some researchers declare the magnitude of an observed significant effect to be the magnitude of the true effect, while non-significant effects are left undecided or unclear. This more conservative approach and the resulting inferential errors (defined as for conventional NHST) are shown in Fig. 1.

2.2 Inferences and Inferential Errors with MBI

MBI was developed from the intuitively obvious interpretation of the confidence interval of an effect statistic, such as the difference between two mean values, as the uncertainty in the true value of the effect. In simple terms, the upper and lower confidence limits are interpreted as how big, in a positive or negative sense, the true effect could be, where "could" is defined by the level of confidence. From a frequentist (Neymanian) perspective, of course, the level of confidence refers strictly to the proportion of confidence intervals that contain the true value [19]. Indeed, the confidence interval may be interpreted as a statement of posterior probability only within a Bayesian system of inference [20]. In MBI, our intuitive interpretation of the conventional confidence interval as the likely range for the true value of the effect requires a "least informative" prior (a uniform, flat distribution), and, as such, MBI is regarded as an objective or "reference" Bayesian method [21]. The intuitive interpretation of the standard confidence interval is therefore valid from a Bayesian perspective, notwithstanding claims to the contrary [12, 15]. Inferences and inferential errors in MBI follow directly from this interpretation.

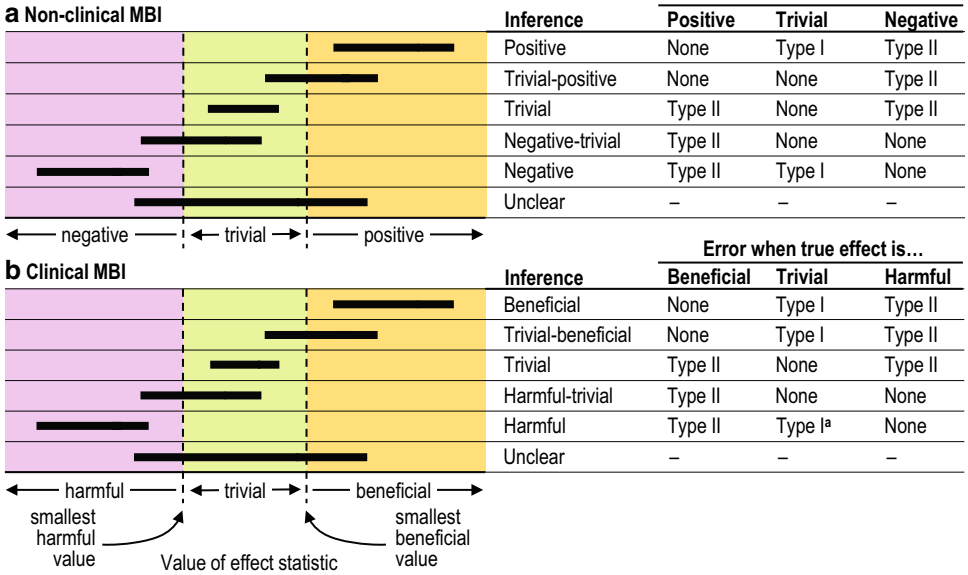
In our view, it is self-evident that the outcome in a study of a sample is acceptable, publishable and—in the case of outcomes with clinical or practical application—implementable, if the uncertainty in the true effect is acceptable (or, equivalently, if the precision of the estimate of the true effect is adequate). So how should the researcher decide what represents acceptable uncertainty? The first approach in MBI is to choose a confidence interval with an appropriate level of confidence. If the upper confidence limit represents a substantial value while the lower confidence limit represents something substantial in the opposite sense, the effect should be reported as "unclear"; you do not move the field forwards by much to say that, on the basis of your data, the effect could be anything between substantially negative and substantially positive. Effects become clear when the confidence interval no longer includes substantial effects of the opposite sign. An informative way to report the magnitude of a clear effect consistent with the use of the confidence interval is simply as the qualitative range represented by the lower and upper confidence limits: both negative, negative-trivial, trivial-



a Conventional NHST	Inference	Error when true effect is… Positive	Trivial	Negative
●	Sig. = positive	None	Type I	Type II
● ●	N.s. = trivial	Type II	None	Type II
●	Sig. = negative	Type II	Type I	None

b Conservative NHST	Inference	Error when true effect is… Positive	Trivial	Negative
●	Sig. = positive	None	Type I	Type II
●	Sig. = trivial	Type II	None	Type II
●	Sig. = negative	Type II	Type I	None
● ●	N.s. = unclear	–	–	–

Fig. 1 Inferences and inferential errors in **a** conventional null-hypothesis significance testing (NHST) and **b** conservative NHST. The *black circles* indicate observed values of an effect in a sample, and the inference is the outcome from the analysis. The *coloured zones* indicate negative or substantially negative effects (*purple*), trivial effects (*green*) and positive or substantially positive effects (*orange*). *Sig.* statistically significant, *N.s.* statistically non-significant



a Non-clinical MBI		Error when true effect is…		
	Inference	Positive	Trivial	Negative
	Positive	None	Type I	Type II
	Trivial-positive	None	None	Type II
	Trivial	Type II	None	Type II
	Negative-trivial	Type II	None	None
	Negative	Type II	Type I	None
	Unclear	–	–	–

b Clinical MBI		Error when true effect is…		
	Inference	Beneficial	Trivial	Harmful
	Beneficial	None	Type I	Type II
	Trivial-beneficial	None	Type I	Type II
	Trivial	Type II	None	Type II
	Harmful-trivial	Type II	None	None
	Harmful	Type II	Type I[a]	None
	Unclear	–	–	–

Fig. 2 Inferences and inferential errors in **a** non-clinical magnitude-based inference (MBI) and **b** clinical MBI. The *coloured zones* are as defined in Fig. 1, except that for clinical MBI, substantial is harmful or beneficial. The *horizontal black bars* represent the confidence intervals: symmetric 90 % for non-clinical MBI, asymmetric 99 % on the harm side and 50 % on the beneficial side of the observed effect for clinical MBI. The observed values of the effect are not shown. [a]This error was determined with a 90 % confidence interval

positive or both positive. Figure 2 illustrates these outcomes and associated confidence intervals for clear and unclear effects, along with the inferential errors.

A false-negative or type II error occurs when the true value of an effect is substantial and the confidence interval does not include substantial values of the same sign. Sample-size estimation in MBI is determined by specifying the two type II error rates for true positive and true negative effects without any need to define and specify a type I error rate. By analogy with NHST, it is intuitively reasonable to define a type I error in MBI as the error when the true effect is trivial but is inferred to be non-trivial—that is, the confidence interval does not include trivial effects, as shown in the figure. This definition differs from that in a recent critique of MBI, where the authors apparently decided that any overlap of a confidence interval with substantial effects would incur a type I error when the true effect was null [12].

A problem with inferences based on confidence intervals is the arbitrary level of confidence: should it be 90, 95, 99 % or something else? The solution to this problem provides the second approach to MBI: calculation and interpretation of the probabilities that the true effect is substantial in a negative sense, substantial in a positive sense and trivial. The calculations are based on the same sampling theory that underlies the confidence interval and the traditional *P* value. This practical approach to deriving posterior probabilities, based on an intuitive Bayesian interpretation of the standard confidence interval, has been advanced by others and might be especially valuable in assessing the clinical relevance of effects [10, 20, 22]. Once calculated, the probabilities can be interpreted in qualitative terms to communicate the uncertainty in plain language. We use the following scale: <0.5 %, most unlikely; 0.5–5 %, very unlikely; 5–25 %, unlikely; 25–75 %, possible; 75–95 %, likely; 95–99.5 %, very likely; >99.5 %, most likely [9]. The true value of an effect deemed clear with a 90 % confidence interval has a <5 % chance of being substantially negative (say) and is therefore very unlikely to be negative, because the confidence interval does not include substantial negative values. Clearly, non-negative effects could have various combinations of probabilities of being trivial and positive but need be reported only as the magnitude with the largest qualitative probability (e.g. likely positive). Those who use MBI usually also state a qualitative magnitude for the observed effect (trivial, small, moderate, large, very large or extremely large), but we will not address error rates for attribution of such magnitudes here.

Consideration of the probabilities for the magnitude of the true effect led to a different version of MBI for effects where substantial means beneficial and harmful [9, 14]. Here, harmful refers to an effect on the dependent (outcome) variable that is in the opposite direction to benefit, rather than to adverse side effects. For such clinically or practically relevant effects, implementation of a harmful effect represents a more serious error than failure to implement a beneficial effect. Although these two kinds of error are both false-negative type II errors, they are analogous to the statistical type I and II errors of NHST, so they are denoted as clinical type I and type II errors, respectively. The default maximum acceptable rates for these errors are 0.5 and 25 %; in plain language, an effect is implementable if it is possibly beneficial and most unlikely to be harmful. Any possibly beneficial effect with a higher risk of harm is unclear, and all other effects are

clear and not implementable. The different inferential outcomes and different kinds of inferential error can be visualized with confidence intervals consisting of a 50 % level on the benefit side of the observed effect and a 99 % level on the harm side, as shown in Fig. 2 for what we call clinical MBI.

As in non-clinical MBI, the type I error of NHST can be appropriated for errors when the true effect is trivial, but there is an important difference between the type I errors in non-clinical and clinical MBI. In non-clinical MBI, a clear effect deemed possibly trivial and possibly positive does not incur a type I error when the true value is trivial, because the inference allows for the true value to be trivial. Equivalently, a type I error occurs in MBI only when a trivial effect is declared very unlikely to be trivial. In clinical MBI, a decision has to be made about whether to implement an effect, and a possibly beneficial effect is a candidate for implementation. Hence, in clinical MBI, a type I error occurs if the effect is deemed clear and possibly beneficial when the true effect is trivial. A type I error also occurs when a trivial true effect is deemed wholly harmful (very unlikely to be trivial).

The decision about implementation is marginal when the chance of benefit is 25 % and the risk of harm is 0.5 %, which corresponds to a benefit/harm odds ratio of 66. This ratio is used as the minimum value for declaring unclear effects beneficial in a less conservative approach to clinical MBI, in which we allow for an increased chance of benefit to outweigh an otherwise unacceptable risk of harm in under-powered studies. The same threshold can be used to justify implementation of effects deemed unlikely to be beneficial in over-powered studies, when the risk of harm is sufficiently low. The inferences and errors shown in Fig. 2 apply to this odds-ratio version of clinical MBI, except that some unclear effects and some wholly trivial effects become beneficial and incur errors accordingly.

2.3 Derivation of Error Rates

NHST and MBI can be used with bootstrapping when the sampling distribution of the effect statistic cannot be quantified; however, for convenience, we have limited the estimation of error rates to designs where the original data, and therefore the effect statistic, are guaranteed to be normally or T distributed. The design is irrelevant, so we opted for a randomized, parallel-group, pre–post, controlled trial, with the difference in the change scores (post minus pre) as the dependent variable. A test–retest intraclass correlation of 0.818 was chosen to give an optimum sample size of 50 in each of the two groups for MBI, with a smallest important effect defined by standardization (see below) as 0.2 of the baseline between-subject standard deviation. The estimated sample sizes are 50.2 and 49.7 in each group, respectively, with the default error rates of 5 % type II (type I or II) for non-clinical MBI and 0.5 % type I and 25 % type II for clinical MBI [9]. The optimum sample size for NHST with the usual 5 % type I error and 20 % type II error is 144 in each group. These sample sizes were estimated with a spreadsheet at the Sportscience site [23]. Error rates for sample sizes of 10 + 10 (representing a grossly under-powered study not uncommon in our literature; power for NHST = 12 %), 50 + 50 (compromise optimum for non-clinical and clinical MBI; power = 38 %) and 144 + 144 (optimum for NHST; power = 80 %) were determined by simulation, in which 500,000 randomly generated samples were analysed for each of a range of true values of the effect. The resulting sampling uncertainty in error rates of 0.01 % (the lowest shown in the figures), expressed as a standard error via the binomial distribution, is ±0.0014 %, which produced practically negligible deviations in such values in the figures. Standardization with the between-subject baseline sample standard deviation was used to define trivial and important differences in the changes in the means in the two groups, and the standardized true effects chosen for the simulations were ±0.6 (borderline moderate effects), ±0.5, ±0.4, ±0.3 (small effects), ±0.2 (borderline small effects), ±0.199 (borderline trivial effects), ±0.1 (trivial effects) and 0.0 (null effects) [9]. The standardized effect was corrected for small-sample bias [24]. Positive standardized effects were chosen as beneficial for clinical inferences.

The samples were generated and analysed with the Statistical Analysis System (Version 9.4; SAS Institute, Cary, NC, USA). The SAS program is available in the Electronic Supplementary Material (Online Resource 1). The analyses were performed with the general linear mixed model (Proc Mixed), allowing for a different error variance in the two groups. *P* values, confidence limits and probabilities of magnitudes of the true effect were estimated under the assumption of the central T distribution for the effect statistic, although—strictly speaking—the non-central T distribution is required for standardized effects to account for uncertainty in the between-subject standard deviation. The error rates therefore represent those obtained by researchers who routinely use the central T distribution to analyse standardized differences in means.

2.4 Derivation of Decision Rates and Publication Bias

The proportion of effects leading to a decision in conventional NHST is, by definition, 100 %; the decision rates in conservative NHST and MBI are the proportions of significant and clear effects, respectively. Mean values of the decisive effects were calculated to address the issue of

publication bias. Mean values of decisive effects in conventional NHST are problematic: all effects are decisive, but non-significant effects have to be pronounced as "not real" or even null, regardless of the observed value. However, perhaps only radical proponents of NHST would seriously suggest treating the effects as null in a meta-analysis. If all values are published, their mean value will be the true value, and there will be no publication bias. If only the significant effects are published, the mean will be the same as that for conservative NHST. We have therefore shown the mean of significant positive effects, to demonstrate the publication bias that ensues when significant effects that apparently go "against the tide" fail to get into print [25].

3 Results

3.1 Error Rates

The inferential error rates with the five methods of inference for each of the three sample sizes are shown in Fig. 3. The observed error rates for true effects that are used to define the inference and/or sample size are those predicted by theory. For conventional NHST, the type I rate was 5 % with a null true effect and any sample size, the type II rate was 20 % for a smallest effect of ±0.20 and an optimum sample size of 144 + 144, and the corresponding type I rate was 80 % for a true effect of ±0.199. For non-clinical MBI, the type I and II rates for marginally substantial effects of ±0.2 were both 5 %, while for clinical MBI, the type I and type II rates (type II for true harm and true benefit, respectively) were 0.5 and 25 %, all regardless of sample size. The type I error rate for clinical MBI for a true effect of 0.199 and an optimum sample size of 50 + 50 was 71 % (the difference between the predicted value of 75 % being made up by 4 % of unclear effects).

The error rates shown in Fig. 3 can be interpreted as the expected number of erroneous decisions in every 100 study inferences. The number of such errors for a true null effect in grossly under-powered studies (here, a sample size of 10 + 10) are of particular interest. These numbers and the precise meaning of the errors are as follows: for conventional NHST, 5.0 where the effect was declared significant and therefore substantial; for conservative NHST, 5.0 where the effect was declared significant and the observed value was substantial; for non-clinical MBI, 1.8 where the 90 % confidence interval spanned entirely substantial values (equivalently, the chance of the effect being trivial was <5 %, or very unlikely); for clinical MBI, 2.9 where the effect was most unlikely harmful (risk <0.5 %) and at least possibly beneficial (chance >25 %), and 0.9 where the effect was very likely harmful; and for odds-ratio MBI, an additional 8.8 (12.5 in total) where the risk of harm was higher than 0.5 % but the chance of benefit was high enough to make the odds ratio of benefit/harm exceed 66. Of the 12.5 errors in odds-ratio MBI, 6.8 and 0.9 were declared likely and very likely beneficial, respectively.

Non-clinical MBI had the lowest type I error rates across all trivial true values and sample sizes, reaching a maximum of 5 % for marginally trivial values and falling below 0.1 % for trivial values in the range of ±0.1 with the largest sample size. Type I rates for NHST exceeded those for clinical MBI for negative–trivial values across all sample sizes (~7–90 % versus ~0.1–5 %). For null and positive trivial values, the type I rates for clinical MBI exceeded those for NHST for a sample size of 50 + 50 (~15–70 % versus ~5–40 %), while for the largest

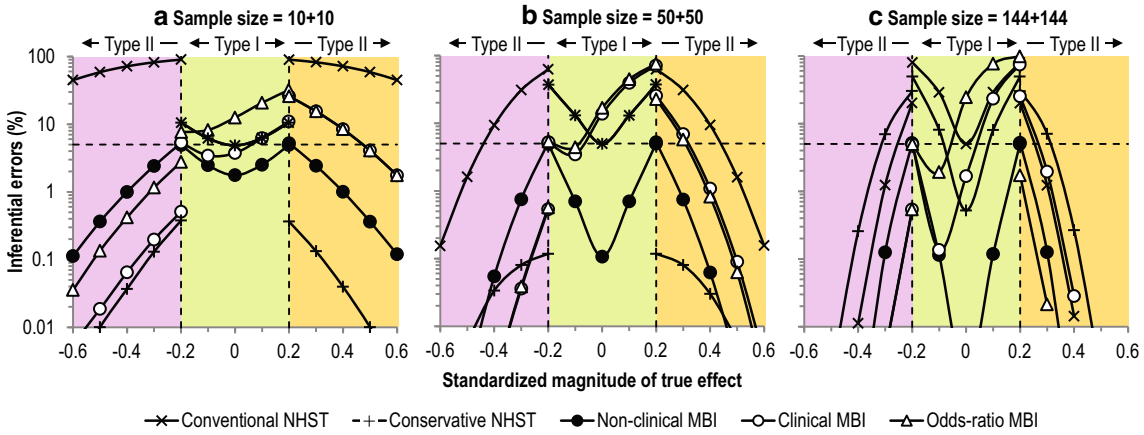


Fig. 3 Inferential error rates with the five methods of inference with sample sizes of **a** 10 + 10, **b** 50 + 50 and **c** 144 + 144. The *coloured zones* are as defined in Figs. 1 and 2. The *dashed horizontal lines* indicate an error rate of 5 %. *MBI* magnitude-based inference, *NHST* null-hypothesis significance testing

sample size, the type I rates for clinical MBI (~2–75 %) were intermediate between those of conservative NHST (~0.5–50 %) and conventional NHST (5–80 %). Odds-ratio MBI produced type I rates similar to those for clinical MBI across all trivial values at the MBI-optimum sample size of 50 + 50, but the rates exceeded those for clinical MBI for the suboptimal sample size (~8–20 versus 5–10 %) and the supra-optimal sample size (~2–98 versus ~0.1–75 %).

For further understanding of the type I errors with odds-ratio MBI, which are the highest of all methods for true effects in the range of null to marginally trivial–beneficial, the qualitative probabilities of benefit are shown in Table 1. Errors where the effect was deemed likely beneficial predominated with the smallest sample size, whereas for optimal and supra-optimal sample sizes, the majority of errors were possibly beneficial.

Conventional and conservative NHST enjoyed the highest and lowest type II rates for sample sizes of 10 + 10 (maxima of ~90 and ~0.4 %, respectively, for the smallest important effects, versus 5–0.5 % for MBI) and 50 + 50 (~60 and ~0.1 %, respectively, versus 5–0.5 %). For the NHST optimum sample size of 144 + 144, both forms of NHST had the highest type II rates for negative true values (maxima of 20 and 30 %, respectively, for conventional and conservative NHST, versus 0.5–5 % for MBI), while for positive true values, the rate for clinical MBI (maximum 25 %) was intermediate between the two NHST values, and the rates for odds-ratio MBI and non-clinical MBI were lower (maxima of ~2 and 5 %, respectively).

3.2 Rates of Decisive Effects

Figure 4 shows the rates (proportions) of effects leading to decisions about the magnitude of the true effect. For conventional NHST, the rate was 100 % regardless of the sample size and effect magnitude. Conservative NHST had the lowest rates, bottoming out at 5 % for true null effects across all sample sizes. Between 35 and 95 % of trivial–small effects were decisive with MBI for the smallest sample sizes (at least 60 % for odds-ratio MBI), rising to at least 95 % for the MBI-optimum sample size and 100 % for the largest sample size.

3.3 Mean Values of Publishable Effects

Figure 4 shows that substantial bias in publishable effects (an absolute difference between the true and mean publishable values of at least 0.2 units) did not occur with MBI for any true value with the smallest sample size, whereas conservative NHST produced substantial bias for all but the null value, and the mean value of statistically significant positive effects had the greatest bias for trivial and negative true values. For the larger sample sizes, MBI and conservative NHST did not produce substantial bias, but

Table 1 Qualitative probabilistic terms accompanying type I errors for inferred beneficial effects with odds-ratio magnitude-based inference (MBI) for sample sizes that are suboptimal (10 + 10), optimal (50 + 50) and supra-optimal (144 + 144) for MBI

Sample size	True effect[a]	Rate (%) of inferred probability of a beneficial effect					
		Very unlikely[b]	Unlikely	Possible	Likely	Very likely[c]	Total
10 + 10	−0.199	0	0	1.2	1.5	0.09	2.8
	−0.1	0	0	2.4	3.3	0.3	6.0
	0	0	0	4.0	6.8	0.9	11
	0.1	0	0	5.9	12	2.2	20
	0.199	0	0	7.6	19	4.8	32
50 + 50	−0.199	0	0.1	0.3	0	0	0.5
	−0.1	0	0.7	2.9	0.08	0	3.7
	0	0	2.2	14	1.0	0.05	17
	0.1	0	3.5	34	6.0	0.7	44
	0.199	0	2.9	48	20	4.5	75
144 + 144	−0.199	0.02	0.01	0	0	0	0.03
	−0.1	1.3	0.5	0.02	0	0	1.8
	0	12	11	1.6	0.03	0	24
	0.1	17	36	21	1.8	0.1	76
	0.199	3.5	20	50	20	5.0	98

[a] In standardized units

[b] Including most unlikely

[c] Including most likely

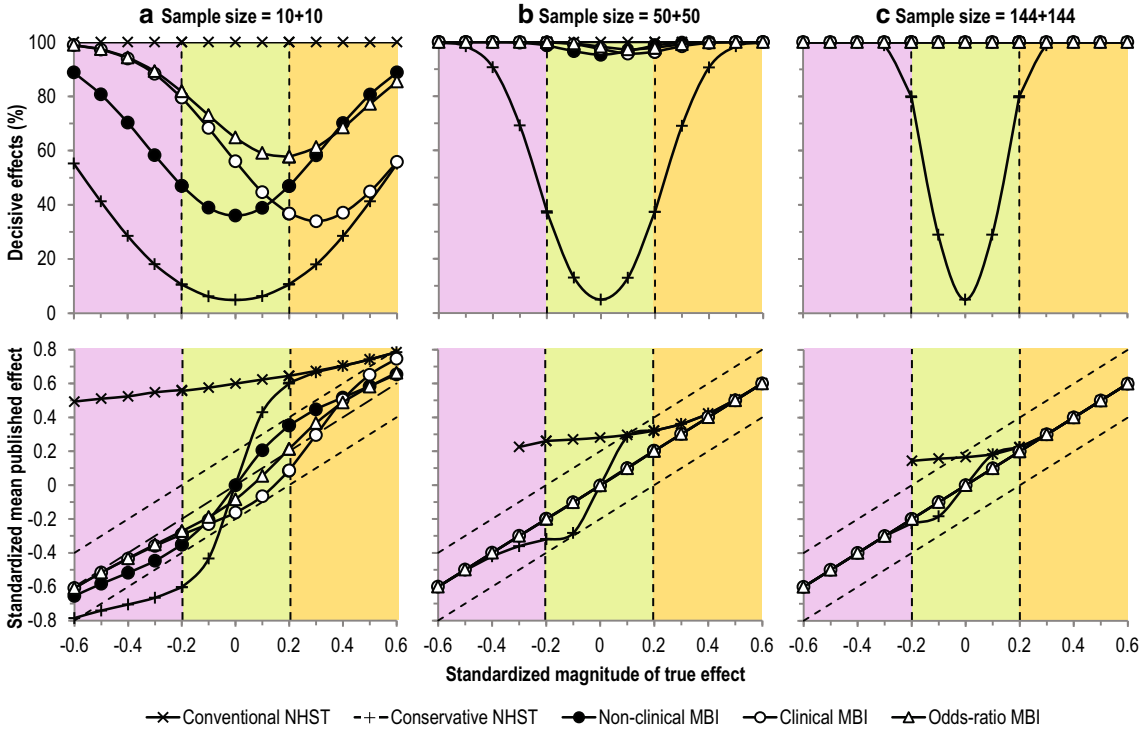


Fig. 4 Rates of decisive effects and mean standardized magnitudes of publishable effects with the five methods of inference. The *coloured zones* are as defined in Figs. 1 and 2. The *dashed oblique lines* delineate the zones of trivial bias (the mean decisive effect falls within the true standardized effect of ±0.2). The mean decisive effect for conventional null-hypothesis significance testing (NHST) is that of significant positive effects. The mean publishable effects are for significant (NHST) or clear (magnitude-based inference [MBI]) effects and sample sizes of **a** 10+10, **b** 50+50 and **c** 144+144

the mean value of statistically significant positive effects still showed substantial bias for some trivial and negative true values.

4 Discussion and Conclusion

The agreement between theoretical and observed error rates for the special cases used to define sample size gives reassurance that the simulations provided trustworthy estimates of error rates, along with the related statistics for rates of decisive effects and bias in such effects, across all values of true effects and all sample sizes. The agreement also provides evidence that the spreadsheet for sample-size estimation at Sportscience gives correct estimates for MBI and NHST, refuting a recent claim to the contrary [12]. The slight underestimate of the type I error rate for a marginally trivial–beneficial effect with clinical MBI was due to a small proportion of unclear effects with the optimum sample size, a phenomenon arising from sampling variation that was noted in the article on sample-size estimation at Sportscience [23].

Possibly the most controversial finding in the present study is the low rate of type I errors with non-clinical MBI, in contrast to the high rate claimed in a recent critique published in this journal [12]. As argued previously in the article on sample-size estimation [23], in a letter to the editor of this journal [13] and in this article, a type I error is not committed when a true trivial effect is inferred to be possibly trivial and possibly substantial—or, indeed, unlikely trivial and likely substantial. It is only when the confidence interval does not include trivial effects, and therefore the effect is deemed very unlikely to be trivial, that a type I error is incurred, and the maximum rate of such errors is 5 %. In contrast, the type I error rate for conventional NHST is at least 5 % for all sample sizes, while for conservative NHST, it drops below 5 % only for sample sizes greater than those optimal for MBI and only for true effects close to the null.

High type I error rates are inevitable for NHST and the two versions of clinical MBI as the sample size increases and as the magnitude of trivial effects approaches the smallest important value (the smallest important beneficial value for clinical MBI). For conventional NHST, the theoretical maximum value of 80 % was reached for the NHST optimum sample size (144 + 144), and the rates would be even higher for supra-optimal sample sizes. For conservative NHST, the type I rate approached its

theoretical maximum of 50 % with a sample size of 50 + 50; for larger sample sizes, the type I rate was anchored at 50 % for marginally trivial effects but fell markedly for smaller trivial effects. The type I rate for clinical MBI showed behaviour similar to that of conservative NHST for trivial effects approaching the marginally beneficial value, where the upper limit was anchored at 75 %. However, for trivial effects approaching the marginally harmful value, the upper limit was anchored at only 5 %.

The type I rates for clinical MBI were substantially higher than those for NHST for null and positive true values with a sample size of 50 + 50. The probabilistic inferences for the majority of these errors were only possibly beneficial, so a clinician would make the decision to use a treatment on the basis of the effect, knowing that there was not a high probability of benefit. Type I error rates for odds-ratio MBI were the largest of all of the inferential methods for null and positive trivial effects, but, for the most part, these rates were due to outcomes where the chance of benefit was rated unlikely or very unlikely but the risk of harm was so much lower that the odds ratio was >66. Inspection of the confidence intervals for such effects would leave the clinician with little expectation of benefit if the effect were implemented, so the high type I error rates should not be regarded as a failing of this approach.

The extreme difference in the type II error rates for the two versions of NHST with suboptimal sample sizes highlights the dilemma facing those who use NHST to make decisions about effects: the rates are unacceptable if non-significant effects are deemed trivial, but taking the conservative approach of interpreting only significant effects results in low rates of decisive effects and substantial bias in effect magnitudes. Ours is the first study to quantify the rates and the bias in terms of standardized magnitudes. The bias with conservative NHST was substantial for grossly under-powered studies (10 + 10) but surprisingly negligible for studies with approximately one third of the NHST optimum size (50 + 50) or more. The bias arising from publishing significant effects of only one sign was considerably greater and was substantial with the largest sample size for true trivial effects of the opposite sign. This kind of bias can be eliminated only if all significant effects that are seemingly contradictory to theory or current evidence are submitted and published.

The type II rates for the various forms of MBI fell between those of the two versions of NHST, except for the largest sample size, when the MBI rates were all less than those of NHST. In any case, the maximum type II rate for clinical MBI is the default 25 %, corresponding to failure to observe a possibly beneficial effect. Any researcher uncomfortable with this rate can reduce it to the 20 % of NHST but will have to accept that there is a higher type I rate for marginally trivial effects, a larger optimal sample size and increased bias in publishable effects.

The rates of decisive effects were lowest with conservative NHST—the version that effectively operates when only significant effects in under-powered studies end up in print. In contrast, MBI had higher rates of decisive effects across the whole range of trivial and small true effects, and the resulting effects showed only trivial bias. These two findings are arguably as important as any considerations of what defines the meaning and acceptable rates of type I and type II errors. If researchers using MBI can publish more of their under-powered studies, if the uncertainty in the effects is explicit as confidence intervals of acceptable width and if the resulting meta-analysed effects are not biased, then the underlying inferential error rates must also be acceptable.

It is important to address a theoretical issue that some Bayesian statisticians may regard as a limitation of MBI. As stated, the basis of MBI is the requirement for a least informative prior belief in the true value of the effect. This prior distribution is uniform (flat) on the scale of the outcome variable [20]. Such uniformity indicates that the posterior distribution will have the same shape as the likelihood function and therefore the standard confidence interval will be equivalent to the Bayesian "credible interval" and may be interpreted as such [21]. Importantly, a prior cannot be uniform on two different scales—for example, raw and logarithmic—so common data transformations might lead to interpretational problems [10]. However, we believe that this threat is largely theoretical, as with any reasonable sample-size alterations of scale and any consequent non-uniformity of the prior have a negligible effect on the effect estimates [26]. Related to the non-uniformity of the prior distribution across different scales is the fact that a least informative prior is not a formal mathematical representation of the lack of prior information [10], but the MBI approach reflects a preference for letting the data speak for themselves, with the inference driven by the data at hand [20].

The probabilities of benefit and harm in MBI are defined in relation to a value for the smallest important effect. Researchers should justify a value within a published protocol in advance of data collection, to show they have not simply chosen a value that gives a clear outcome with the data. Users of NHST are not divested of this responsibility, as the smallest important effect informs sample-size estimation. Standardization—the approach used here—is one of several methods [27, 28].

Although the principles of MBI have been advanced by others [10, 11, 20, 22], we are the first to give the principles a practical application by providing criteria to decide whether an effect has acceptable uncertainty. These

criteria, based on 90 % confidence limits for non-clinical effects and probabilities of benefit and harm for clinical effects, may seem as arbitrary as NHST's 80 % power for a 5 % level of significance. With the current default criteria, MBI generally outperforms NHST on issues of sample size, type I error rate, type II error rate, decision rate and publication bias. Further debate on the criteria should focus on whether a better balance can be achieved between these crucial issues. Table 2 summarizes the five inferential methods and their relative strengths and weaknesses.

In conclusion, MBI posits that effects inferred not to be substantial in some sense, such as negative or harmful, should be considered clear and publishable, even though the inferred magnitude of some effects ranges from trivial through positive or beneficial. NHST works in a similar manner by positing that effects inferred not to be null are deemed statistically significant and publishable. A simple consideration of the confidence interval for some significant effects shows that the true value could also range from trivial through substantial. The crucial difference is that MBI overtly includes the uncertainty with its inferences (the qualitative magnitudes of the lower and upper confidence limits, or the outcome expressed as likely trivial, possibly beneficial and so on), whereas NHST leads to an assertion only about whether or not the effect is substantial.

Those who would argue that a researcher using NHST should take into account the confidence interval in assessing the magnitude of significant effects are effectively arguing for MBI with the added burden of either low decision rates and substantial bias in under-powered studies or low decision rates for marginally null effects in over-powered studies. Those who would still argue that NHST, by virtue of its requirement for larger sample sizes, somehow achieves more trustworthy inferences will have to reconcile their argument with the publication bias in under-powered studies arising from the low decision rates. They will also have to acknowledge that the higher decision rates of MBI will promote progress away from the existing culture of manuscript rejection, which makes a career in science unattractive for some young researchers and frustrating for the more experienced.

Recently, it has been emphasized that there is no universal method of inference; rather, what researchers need is a statistical toolbox of robust methods [29]. We have provided evidence that MBI is superior to NHST and therefore deserves a place in this toolbox. Researchers may

Table 2 Summary of null-hypothesis significance test (NHST) and magnitude-based inference (MBI) methods, with their main strengths and weaknesses

	Method	Main strengths	Main weaknesses
Conventional NHST	Significant effects are substantial; non-significant effects are null	Requires consideration only of the *P* value	High type I with large samples; high type II with small samples; low publication rate and substantial publication bias[d] with small samples
Conservative NHST	Significant effects are assessed for magnitude; non-significant effects are unresolved	Adds magnitude to conventional NHST	High type I for marginally trivial effects and type II for marginally small effects with large samples; low publication rate and substantial publication bias[d] with small samples
Non-clinical MBI	Confidence limits are assessed for magnitude; unclear effects have substantial negative and positive limits	Explicit uncertainty reduces misinterpretation; lowest type I rate[b] and low type II rate; trivial publication bias	Unacceptable to some reviewers
Clinical MBI	Unclear effects have a reasonable chance of benefit and an unacceptable risk of harm; all other effects are clear[a]	Explicit assessment of probability of benefit and harm[c] best for clinical or practical settings; high publication rate and trivial publication bias	Unacceptable to some reviewers; high type I for null to marginally trivial–beneficial effects with moderate–large samples[e]
Odds-ratio MBI	As for clinical MBI, but unclear effects with sufficiently high odds of benefit relative to odds of harm are deemed beneficial	Highest decision rates for small samples; lowest trivial publication bias	Unacceptable to some reviewers; highest type I for null to marginally trivial–beneficial effects with moderate-large samples[e]

Type I rate of false-positive error, *type II* rate of false-negative error

[a] Clear possibly beneficial effects are deemed implementable; unlikely beneficial effects are not implementable

[b] *Possibly substantial* and *likely substantial* are not considered errors for a trivial true effect

[c] *Harm* refers to a direct adverse effect, not adverse side effects

[d] The rate and bias are based on the assumption that only significant effects are accepted for publication

[e] Most such errors are for effects inferred to be *possibly* or *likely beneficial*

cite this article as justification for choosing a version of MBI from the toolbox and for not making inferences with the *P* value.

Acknowledgments The authors thank Kenneth Quarrie for his valuable feedback on drafts of this article.

Compliance with Ethical Standards

Conflict of interest Will G. Hopkins and Alan M. Batterham have no conflicts of interest to declare with regard to this publication. No funding was received for the conduct of this study and/or the preparation of this manuscript.

References

1. Carver R. The case against statistical significance testing. Harv Educ Rev. 1978;48:378–99.
2. Cohen J. The earth is round ($p < .05$). Am Psychol. 1994;49:997–1003.
3. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010;25:225–30.
4. Ziliak ST, McCloskey DN. The cult of statistical significance. Ann Arbor: University of Michigan Press; 2008.
5. Cumming G. The new statistics: why and how. Psychol Sci. 2014;25:7–29.
6. Halsey LG, Curran-Everett D, Vowler SL, et al. The fickle *P* value generates irreproducible results. Nature Methods. 2015;12:179–85.
7. Nuzzo R. Scientific method: statistical errors. Nature. 2014;506:150–2.
8. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. Int J Sports Physiol Perform. 2006;1:50–7.
9. Hopkins WG, Marshall SW, Batterham AM, et al. Progressive statistics for studies in sports medicine and exercise science. Med Sci Sports Exerc. 2009;41:3–12.
10. Gurrin LC, Kurinczuk JJ, Burton PR. Bayesian statistics in medical research: an intuitive alternative to conventional data analysis. J Eval Clin Pract. 2000;6:193–204.
11. Shakespeare TP, Gebski VJ, Veness MJ, et al. Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk–benefit contours. Lancet. 2001;357:1349–53.
12. Welsh AH, Knight EJ. "Magnitude-based inference": a statistical review. Med Sci Sports Exerc. 2015;47:874–84.
13. Batterham AM, Hopkins WG. The case for magnitude-based inference. Med Sci Sports Exerc. 2015;47:885.
14. Hopkins WG. A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a *P* value. Sportscience. 2007;11:16–20.
15. Barker RJ, Schofield MR. Inference about magnitudes of effects. Int J Sports Physiol Perform. 2008;3:547–57.
16. Hopkins WG, Batterham AM. An imaginary Bayesian monster. Int J Sports Physiol Perform. 2010;3:411–2.
17. Gigerenzer G. Mindless statistics. J Socio Econ. 2004;33:587–606.
18. Schneider JW. Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. Scientometrics. 2015;102:411–32.
19. Armitage P, Berry G. Statistical methods in medical research. 2nd ed. Oxford: Blackwell Scientific; 1994.
20. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to *P* values. J Epidemiol Community Health. 1998;52:318–23.
21. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: Wiley; 2004.
22. Burton PR. Helping doctors to draw appropriate inferences from the analysis of medical studies. Stat Med. 1994;13:1699–713.
23. Hopkins WG. Estimating sample size for magnitude-based inferences. Sportscience. 2006;10:63–70.
24. Becker BJ. Synthesizing standardized mean-change measures. Br J Math Stat Psychol. 1988;41:257–78.
25. Hopewell S, Loudon K, Clarke MJ, et al. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009;(1):MR000006.
26. Hughes MD. Reporting Bayesian analyses of clinical trials. Stat Med. 1993;12:1651–63.
27. George K, Batterham AM. So what does this all mean? Phys Ther Sport. 2015;16:1–2.
28. Cook JA, Hislop J, Adewuyi TE, et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference Elicitation in Trials) review. Health Technol Assess. 2014;18(28):v–vi, 1–175.
29. Gigerenzer G, Marewski JN. Surrogate science: the idol of a universal method for scientific inference. J Manage. 2015;41:421–40.