Moving the goalposts on purpose: almost sure hypothesis testing and what it would do to a physiotherapy trial

Jürgen Degenfellner*[1]

[1]ZHAW Zurich University of Applied Sciences, Institute of Physiotherapy, Winterthur, Switzerland

Article type: Research article • Word count (main text): ∼3200 • 2 figures, 2 tables, 19 references

Abstract

Background. Almost every clinical trial, physiotherapy trials among them, still declares its result at a single fixed threshold, a p-value below 0.05. The line does not move whether the sample is small and fragile or enormous, and the American Statistical Association has cautioned against this reliance and urged the field to move beyond it. A classical result shows such a fixed line can never be consistent: under a true null the standardised statistic crosses 1.96 infinitely often, however much data is collected. I examine almost sure hypothesis testing, a proposal in which the significance level shrinks with the sample size ($\alpha_n = n^{-p}$, $p > 1$), and ask what it does to the interpretation of a real physiotherapy trial.

Methods. I set out the method without proofs and establish, by simulation under a true null, that a fixed-α rule accumulates false rejections without bound while the shrinking level does not. I then tabulate the rejection threshold the method implies across the sample sizes physiotherapy trials actually reach, and apply it in a fully reproducible secondary reanalysis of the reported summary statistics of a registered, open-access randomised trial (NCT04808141;

*Correspondence: degn@zhaw.ch; ORCID 0000-0001-6886-4583.

$n = 140$) comparing a digital care programme with conventional physiotherapy for chronic low back pain. Every quantity is recomputed from the published summary statistics.

Results. The almost sure test makes only finitely many errors of *either* kind with probability one and dissolves the Jeffreys–Lindley paradox. At the sizes physiotherapy trials reach ($n \approx 40$–100) the rejection bar sits at $|z| \approx 2.5$ to 3.9 ($p \approx 0.0001$ to 0.012), not 0.05. In the reanalysis the trial's main "no difference" primary-outcome conclusion stands, carried by a tight confidence interval rather than by the non-rejection itself; its single borderline "significant" finding (lower dropout in the digital arm, $P = 0.019$, $z = 2.54$) survives at no $p > 1$.

Conclusions. Almost sure testing does not abolish arbitrary thresholds but ties the threshold to the accumulated evidence, disciplining over-claiming from small underpowered trials. Trials could report the verdict across a small range of p beside the effect estimate and its confidence interval. The proposal is a modest but genuine improvement on a fixed bright line, and its limitations, chiefly that the guarantee is asymptotic, are made explicit.

Keywords: statistical significance; p-values; hypothesis testing; Jeffreys–Lindley paradox; physiotherapy; research methods.

Background

Most clinical trials end with a ritual: a number is computed, compared to 0.05, and a verdict of "significant" or "non-significant" is pronounced. That reliance on a single bright line is one the American Statistical Association has cautioned against,[1] and has since urged the field to move beyond altogether.[2] It is so familiar that it is easy to forget how strange it is. The same threshold serves a pilot of thirty patients and a registry of thirty thousand; it does not flex with how much prior belief there was, how much a difference would matter, or how much data is in hand. The irony is old: Fisher, who introduced the significance test, rejected any fixed threshold, holding that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas".[3]

This would be forgivable if the line behaved sensibly as evidence accumulates. It does not. A fixed threshold rejects a true null a fixed fraction of the time, 5% at $\alpha = 0.05$, whatever the sample size; that rate never falls. A field that runs many trials therefore keeps producing

false positives, 5% of its true-null studies, without end. The same rigidity takes a sharper form within a single dataset examined as it grows: under an exactly true null the standardised statistic does not settle but returns above 1.96 infinitely often, a consequence of the law of the iterated logarithm.[4] Either way a fixed line never stops generating false positives, and this owes nothing to how carefully the studies are run; it is built into a threshold that takes no account of the sample size. It is a separate matter from a real effect, however small, which a large enough sample will eventually detect and then keep detecting, the mirror-image rigidity a fixed line shows against a true effect.

A single very large study brings a different failure, the Jeffreys–Lindley paradox:[5] a frequentist may reject the null at $P = 0.01$ on the very data that lead a Bayesian, weighing the same evidence, to grow more convinced the null is true. The two reach opposite conclusions on one dataset, and collecting more data does not settle the disagreement but deepens it.

Naaman offered a way out.[6] If the significance level is allowed to shrink as the sample grows, $\alpha_n = n^{-p}$ with $p > 1$, the test becomes *almost sure*: with probability one it makes only finitely many errors of either kind, so beyond some sample size its verdicts are simply right, and the Jeffreys–Lindley paradox disappears. Equivalently, the fixed significance threshold is replaced by a stricter one as the sample grows.

Physiotherapy is a good place to see what is at stake, though not for the reason one might first expect. That stricter threshold falls furthest below the fixed 0.05 in large studies, so the Jeffreys–Lindley side of the problem is really a large-sample affair. What makes physiotherapy telling is different: its trials are small and typically underpowered for the small-to-moderate effects that are realistic,[7] so their results tend to sit close to the 0.05 line rather than far beyond it; its primary outcomes are pain and function scores, and whatever test is used, each trial's verdict reduces to a single p-value weighed against 0.05, exactly what a shifting threshold acts on; and over half of trials nonetheless report a positive result,[8] so a large share of published "significant" findings clear the 0.05 line only narrowly. Such results, only just "significant", are common across the sciences, where published p-values cluster just below 0.05,[9] and physiotherapy is no exception; they are exactly what a modestly raised threshold reclassifies. I explain the idea without proofs, tabulate what it does at realistic physiotherapy size, apply it to a real, registered, open-access trial, and ask whether it is an improvement or merely a relocation of the same arbitrariness.

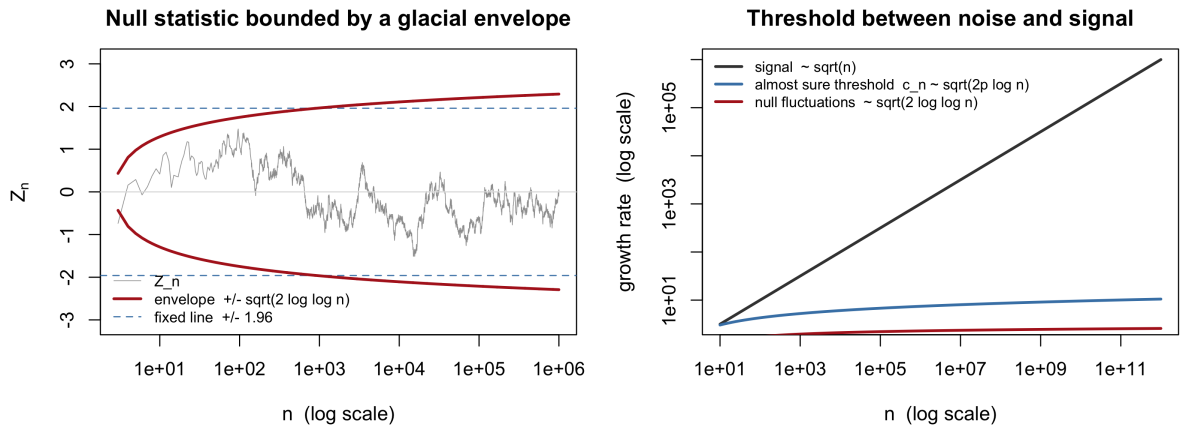

Figure 1: Why a fixed line fails, and what fixes it. *Left:* a simulated path of the standardised statistic Z_n under a true null ($\mu = 0$) stays within the envelope $\pm\sqrt{2\log\log n}$ (red), which reaches only about 3.3 even at $n = 10^{100}$; a fixed line at ± 1.96 (dashed) is therefore crossed again and again. *Right:* on log–log axes the almost sure threshold $c_n \approx \sqrt{2p\log n}$ (here $p = 2$) lies between the null fluctuations, of order $\sqrt{2\log\log n}$, and the statistic under a real effect, of order $\sqrt{n}$, so it eventually rises above the null peaks yet stays below a genuine signal. Code: `asht_lil.R`.

Methods

The fixed threshold and why it cannot be consistent

Take the simplest problem: is a mean different from zero? One computes the usual standardised statistic Z_n and rejects the null when $|Z_n| > 1.96$.

The trouble is what Z_n does over increasing n when the null is *true*: it converges to nothing. The law of the iterated logarithm pins down its largest excursions,

$$\limsup_{n\to\infty} \frac{|Z_n|}{\sqrt{2\log\log n}} = 1 \qquad \text{almost surely.} \tag{1}$$

The peaks of $|Z_n|$ grow, slowly, like $\sqrt{2\log\log n}$, without end. A line at 1.96 is eventually below those peaks and so is crossed infinitely often; a fixed-α test never stops producing false positives (Figure 1, left).

The almost sure test

Naaman's repair raises the line as the sample grows.[6] Set the significance level to

$$\alpha_n = n^{-p}, \qquad p > 1, \tag{2}$$

so the rejection threshold becomes

$$c_n = \Phi^{-1}(1 - \tfrac{\alpha_n}{2}) \approx \sqrt{2p \log n}. \tag{3}$$

Both equations are Naaman's (his Theorem 2.1 takes critical values $\Phi^{-1}(1-\alpha_n)$ with $\alpha_n \propto n^{-p}$, $p > 1$). The equality in (3) is just the two-sided critical value at level α_n (as $1.96 = \Phi^{-1}(0.975)$ is the critical value at $\alpha = 0.05$); the approximation is the standard Gaussian-tail asymptotic $\Phi^{-1}(1-u) \sim \sqrt{-2\log u}$ for small u, which gives $c_n^2 \approx -2\log\alpha_n = 2p\log n$. This grows like $\sqrt{\log n}$, faster than the $\sqrt{\log\log n}$ peaks of the null statistic but far slower than $\sqrt{n}$ (Figure 1, right). That single fact does the work in both directions:

- **Null true:** the false-positive chance at step n is $\alpha_n = n^{-p}$; since $p > 1$ these sum to a finite total ($\sum n^{-p} < \infty$), so by Borel–Cantelli only *finitely many* false positives occur with probability one. Past some point, the test stops rejecting a true null forever.
- **Null false**, with a real effect δ: the signal grows like $Z_n \approx \sqrt{n}\,\delta/\sigma$, outrunning $c_n \approx \sqrt{2p\log n}$, so only *finitely many* false negatives occur with probability one. Past some point, the test detects the effect forever.

A test with this property, finitely many errors of either kind with probability one, is what Naaman calls an *almost sure* test, one that in his words "performs well, regardless of the validity of the null". The guarantee is symmetric: the test is consistent under the null as well as the alternative (his Theorem 2.1), so, in his words, "the correct decision will be reached with a probability approaching one as the sample size increases". Read those words carefully: this is a property of the *decision* and its long-run error rate, not a measure of evidence for the null in any single sample.

The Jeffreys–Lindley paradox dissolves because the moving threshold recedes at the rate the Bayesian evidence demands: where fixed-α frequentist and Bayesian could disagree forever, the almost sure frequentist and the Bayesian eventually agree. The procedure has one tuning knob, p, setting how aggressively one trades small-sample power for this long-run guarantee.

This is an asymptotic promise: "eventually" does much of the work, and a trial of forty patients is one point on an infinite road, not the destination (a point I return to in the Discussion). At any given n, however, the almost sure test is simply a classical test with a stricter, principled, sample-size-dependent threshold.

Simulation

To make the error behaviour visible I simulated the standardised statistic under a true null ($\mu = 0$): 150 independent paths, each tested at every sample size up to $n = 10^6$, under the fixed rule ($\alpha = 0.05$) and under the almost sure rule ($\alpha_n = n^{-p}$, for $p = 1.2$ and $p = 2$), recording at each step both the rejection rate and its cumulative sum across n.

Application to trials of realistic size

Applying the rule needs only one substitution: replace the fixed bar of 1.96 with $c_n \approx \sqrt{2p \log n}$. I tabulate this bar across the sample sizes physiotherapy trials actually reach ($n = 20$ to 1000), for a cautious decay ($p = 1.2$) and an aggressive one ($p = 2$). I then apply it to a real trial: Cui and colleagues,[10] a registered (NCT04808141), open-access randomised trial comparing a digital care programme ($n = 70$) with conventional in-person physiotherapy ($n = 70$) for chronic low back pain. Every quantity is recomputed from the reported summary statistics: for the primary Oswestry Disability Index (ODI, 0–100) outcome I use the reported median between-group difference and its 95% confidence interval; for the between-arm difference in dropout I compute the difference in proportions and its pooled standard error, form the z statistic, and compare it against both the classical bar of 1.96 and c_n over a range of p.

Results

Finitely many errors, confirmed by simulation

Under the true null the fixed bar rejected 5% of the time at *every* n (Figure 2, left), a rate that never decays, so its false rejections accumulated without bound; the almost sure rate n^{-p} decayed fast enough to sum to a finite total, and its rejections stopped. Because $\sum 0.05 = 0.05\, n$ diverges, the fixed test's cumulative false rejections grew without bound ($\approx 0.05\, n$); because $\sum n^{-p} = \zeta(p)$ (the Riemann zeta function) is finite, the almost sure test committed only finitely many, settling near $\zeta(1.2) \approx 5.3$ and $\zeta(2) \approx 1.6$ (Figure 2, right). Simulation matched theory. This is the finite-error guarantee made visible: with probability one the almost sure test makes only finitely many errors of either kind, so beyond some sample size its verdicts are simply right.

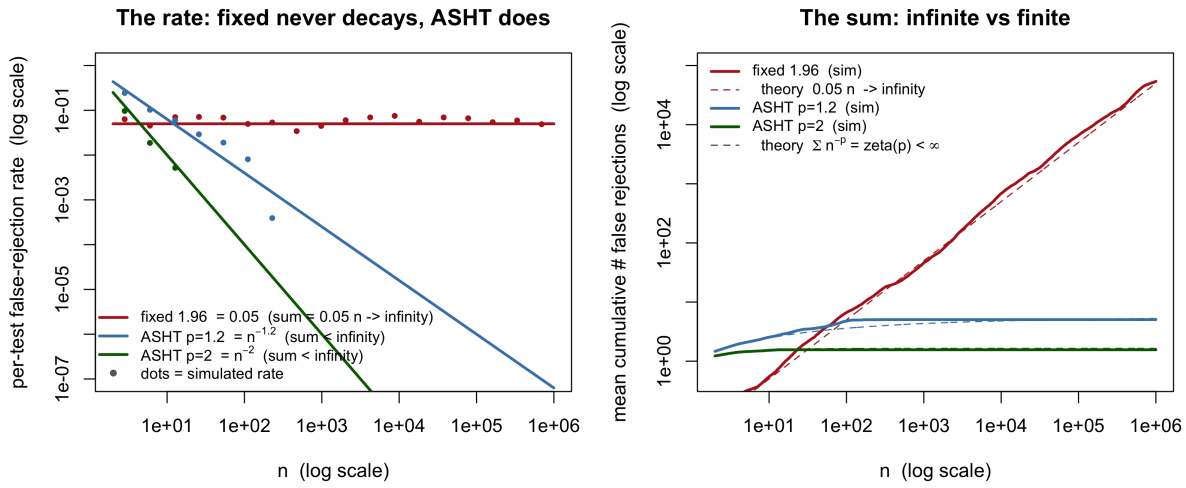

Figure 2: The Borel–Cantelli mechanism, simulated under a true null ($\mu = 0$): 150 independent paths, each tested at every sample size up to $n = 10^6$. *Left, the rate (the summands):* the fixed bar rejects a true null 5% of the time at every n (red), a rate that never decays; the almost sure bar rejects with probability $\alpha_n = n^{-p}$, which does (blue, green). Dots are simulated, lines theoretical. *Right, the sum:* because $\sum 0.05 = 0.05\,n$ diverges, the fixed test's false rejections grow without bound (red, $\approx 0.05\,n$); because $\sum n^{-p} = \zeta(p)$ (the Riemann zeta function) is finite, the almost sure test commits only finitely many, settling near $\zeta(1.2) \approx 5.3$ and $\zeta(2) \approx 1.6$ (blue, green). Simulation (solid) matches theory (dashed). Code: `asht_simulation.R`.

The threshold at realistic sample sizes

Table 1 shows the bar across the physiotherapy range, for a cautious decay ($p = 1.2$) and an aggressive one ($p = 2$).

Table 1: Rejection threshold on $|z|$ under the classical fixed-α rule and under almost sure testing, by total sample size n.

n (total)	classical z	α_n ($p = 1.2$)	z-bar ($p = 1.2$)	α_n ($p = 2$)	z-bar ($p = 2$)
20	1.96	0.027	2.20	2.5×10^{-3}	3.02
40	1.96	0.012	2.51	6.3×10^{-4}	3.42
100	1.96	4.0×10^{-3}	2.88	1.0×10^{-4}	3.89
200	1.96	1.7×10^{-3}	3.13	2.5×10^{-5}	4.21
1000	1.96	2.5×10^{-4}	3.66	1.0×10^{-6}	4.89

The lesson is plain. At the sizes physiotherapists actually work with, roughly 40 to 100, the bar sits between about 2.5 and 3.9 on the z-scale, which implies p-values around 0.0001 to 0.012 rather than 0.05. The familiar result that just scrapes under the line, "$P = 0.04$, therefore it works", does not clear it.

This cuts both ways, part of its appeal. In a very large dataset, where the classical test would crown a clinically meaningless change "highly significant", the bar has risen to 3.7 or 4.9 and

the trivial effect stays where it belongs. The same rule restrains both the over-eager small study and the over-powered large one.

Worked example: digital versus conventional physiotherapy for low back pain

The primary outcome. The primary endpoint was change in the Oswestry Disability Index (ODI, 0–100) from baseline to eight weeks. Both groups improved substantially and similarly; the between-group difference was small and uncertain (median difference -0.55, 95% CI -2.42 to 5.81, $P = 0.412$). Classical and almost sure verdict agree: nothing is rejected. It is tempting to call this evidence that the two programmes are equivalent, but a non-rejection is not, on its own, evidence for the null (absence of evidence is not evidence of absence), and the caution is *sharper* under almost sure testing, whose higher bar carries less power at finite n, so a non-rejection is the more easily a missed effect. What supports the equivalence reading is the confidence interval, not the non-rejection: it confines any difference to a few points on a 100-point scale, well inside what patients would notice. Almost sure testing adds only its long-run consistency under the null, which points the same way as the interval, toward practical equivalence, the authors' own conclusion; but it is the interval, not the non-rejection, that does the work.

The one "significant" finding. Exactly one comparison cleared the conventional bar: dropout was lower in the digital arm (11/70, 15.7%) than the conventional arm (24/70, 34.3%), reported at $P = 0.019$. This is precisely the result the previous section warned about: a single comparison landing just under 0.05 on a sample of 140. So I put it through the almost sure test.

The difference in proportions is 0.186, pooled standard error 0.073, giving $z = 2.54$ (two-sided p of 0.011 without, 0.019 with, the continuity correction the authors used; the two agree). At $n = 140$ the almost sure bar is shown in Table 2.

Table 2: Verdict on the dropout difference ($z = 2.54$, $n = 140$) under each rule.

rule	threshold on $\lvert z\rvert$	reject $z = 2.54$?
classical, $\alpha = 0.05$	1.96	yes
almost sure, $p = 1.1$	2.85	no
almost sure, $p = 1.2$	3.00	no
almost sure, $p = 1.5$	3.43	no
almost sure, $p = 2.0$	4.05	no

The finding survives at *no* $p > 1$, not even the gentlest. The reanalysis therefore leaves the

two programmes with equivalent disability outcomes and only a *suggestion*, not an established fact, that fewer patients abandon the digital programme, a result worth a larger study but not a firm claim.

No correction for multiplicity. A second, more conventional reason to distrust this lone positive: the trial applied no adjustment for multiple comparisons. The Oswestry index was the single pre-specified primary outcome, and dropout is one of a dozen or more secondary and engagement outcomes, each tested two-sided at a fixed 0.05. Under even a lenient Bonferroni correction for ten secondary comparisons the threshold falls to 0.005, and the dropout result ($P = 0.019$) no longer clears it. This is a different objection from the one this paper presses, almost sure testing scales the bar with the sample size, not with the number of tests, but the two cautions point the same way, and the single "significant" finding fails both.

Discussion

Almost sure testing does not abolish the end-of-trial dichotomy, and it does not remove arbitrary constants: it replaces the fixed 0.05 with a significance level that shrinks as $\alpha_n = n^{-p}$, moving the one free choice from the level itself to the exponent p. That is less than some would want, who would move beyond "statistical significance" altogether,[2] but it is not nothing. Unlike a fixed level, the resulting threshold is a function of the evidence gathered and carries a guarantee the old one provably cannot: over the long run the errors of both kinds are finite, and the frequentist and Bayesian readings stop contradicting each other.

Related approaches

The idea that 0.05 should not be a universal constant is not new, though most alternatives reach it by a different route. The "optimal alpha" programme sets the significance level by weighing the two error types against each other: one specifies a critical effect size and how much more serious a Type I error is than a Type II error, and then chooses the α that minimises a weighted sum of the two error rates.[11] The motivation there is the relative cost of the two mistakes, not the sample size as such; the resulting α does depend on n and on the assumed effect size, but balancing the errors is the point. That logic has been carried to whole trial corpora: Habibzadeh reinterpreted thousands of published randomised trials under such a cost-balancing threshold

and, as expected, found fewer "significant" results than at 0.05.[12] A more visible proposal simply lowers the fixed threshold, most prominently to 0.005;[13, 14] that tightens the line everywhere but still holds it fixed, indifferent to n. Closest to Naaman's approach, Pérez and Pericchi let the significance level itself fall as the sample grows, calibrating an adaptive α from a Bayes–frequentist compromise.[15] Almost sure testing begins from the same dissatisfaction with a fixed 0.05 but answers a different question. It weighs no costs and needs no Bayesian calibration; it fixes only the rate at which α shrinks ($\alpha_n = n^{-p}$, $p > 1$), chosen so that the whole sequence of tests makes only finitely many errors of either kind with probability one.[6] It needs neither an effect size nor a cost ratio, only p; its dependence on n is deliberate, and its aim is long-run consistency and the resolution of the Jeffreys–Lindley paradox rather than a per-study cost trade-off. The two schemes can prescribe similar thresholds at a given n while resting on entirely different justifications. These cost-based and sample-size-based thresholds have been developed and applied largely outside clinical medicine, in ecology[11] and in reanalyses of published trials,[12] while the false-positive problem they target is documented in physiotherapy specifically.[8] Almost sure testing has, to my knowledge, never been applied to a physiotherapy trial, or to any clinical trial.

Strengths and limitations

The approach has two clear strengths. Its threshold is principled rather than conventional: it follows from a single interpretable parameter and carries a long-run guarantee a fixed level cannot, reconciling the frequentist decision with the Bayesian reading. And because every quantity is recomputable from the reported summary statistics, the reanalysis is fully reproducible.

The limitations are as important. The central guarantee is asymptotic, and most physiotherapy trials are not: at $n = 140$ that regime, in which "only finitely many errors" would say something about this particular trial, is far off. What almost sure testing offers a single small study is therefore narrower, a stricter and non-arbitrary threshold together with a reconciliation with the Bayesian reading, while its strong-consistency property is really a statement about the accumulating literature rather than any one study. The method must also not be over-read: a non-rejection is not evidence for the null, since absence of evidence is not evidence of absence,[16] and the higher bar in fact lowers power at fixed n. Naaman proves only a property of the decision rule, that it is consistent under the null as well as the alternative (his Theorem 2.1), not that a

non-rejection quantifies evidence for the null. Where the question is genuinely one of equivalence, whether a digital tool is as good as a clinician's hands, evidence *for* that claim, not merely a failure to reject it, calls for a Bayesian analysis: a posterior distribution on the effect lets one accept a null value when its credible interval falls within a region of practical equivalence,[17–19] something a reinterpreted superiority test cannot deliver. Finally, the method does not remove the need to choose a constant: the fixed 0.05 gives way to the exponent p that sets how fast the level shrinks. I would therefore report the decision across a small range of p (say 1.1 to 2) alongside the effect estimate and its confidence interval, treating p as a stated and defensible choice, and read the almost sure threshold as a discipline on over-claiming rather than a replacement oracle.

Conclusions

A fixed significance level is a line that stays put while everything around it (sample size, prior belief, the stakes) moves. The law of the iterated logarithm shows such a line can never be consistent; the Jeffreys–Lindley paradox shows it can put frequentist permanently at odds with Bayesian. Letting the line rise with the data, as Naaman's almost sure test does, repairs both at the cost of one new tuning constant. Applied to a real low back pain trial, it leaves the main "no meaningful difference" conclusion intact, carried by the confidence interval, and quietly retires the trial's one borderline "significant" claim. That is roughly the right behaviour, and physiotherapy, with its small trials, modest effects, and a literature needing fewer and firmer claims, is exactly the kind of field where moving the goalposts, on purpose and by rule, would do more good than harm.

Declarations

Ethics approval and consent to participate. Not applicable. This work is a secondary analysis of aggregate summary statistics from a published, registered randomised controlled trial; no individual-level or identifiable participant data were used.

Consent for publication. Not applicable.

Availability of data and materials. This work uses only the summary statistics reported

in the open-access article by Cui and colleagues[10] (trial registration NCT04808141); the trial's individual participant data were not used. The `R` code reproducing every number, table and figure of the present paper (`asht_demo.R`, `verify_correia.R`, `asht_simulation.R`, `asht_lil.R`) is provided as supplementary material and is openly archived at `https://github.com/jdegenfellner/almost-sure-hypothesis-testing-physio`.

Competing interests. The author declares that he has no competing interests.

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions. JD is the sole author: he conceived the work, performed the analysis, wrote the software, and drafted and revised the manuscript. He acts as guarantor. The author read and approved the final manuscript.

Acknowledgements. During the preparation of this work the author used an AI-assisted large language model (Anthropic Claude) to help with language editing and to check prose and code. After using this tool the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat* 2016;70(2):129–33.

[2] Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$". *Am Stat* 2019;73(sup1):1–19.

[3] Fisher RA. Statistical methods and scientific inference. Edinburgh: Oliver & Boyd; 1956. p. 42.

[4] Hartman P, Wintner A. On the law of the iterated logarithm. *Am J Math* 1941;63(1):169–76.

[5] Lindley DV. A statistical paradox. *Biometrika* 1957;44(1–2):187–92.

[6] Naaman M. Almost sure hypothesis testing and a resolution of the Jeffreys–Lindley paradox. *Electron J Stat* 2016;10(1):1526–50.

[7] Bleakley C, Klempel N, Wagemans J, et al. Statistical power in musculoskeletal research: a meta-review of 266 randomised controlled trials. *Sports Med Open* 2025;11(1):134. doi:10.1186/s40798-025-00908-8.

[8] Bleakley C, Reijgers J, Smoliga JM. Many high-quality randomized controlled trials in sports physical therapy are making false-positive claims of treatment effect: a systematic survey. *J Orthop Sports Phys Ther* 2020;50(2):104–9.

[9] Head ML, Holman L, Lanfear R, et al. The extent and consequences of p-hacking in science. *PLoS Biol* 2015;13(3):e1002106.

[10] Cui D, Janela D, Costa F, et al. Randomized-controlled trial assessing a digital care program versus conventional physiotherapy for chronic low back pain. *npj Digit Med* 2023;6:121.

[11] Mudge JF, Baker LF, Edge CB, et al. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS One* 2012;7(2):e32734.

[12] Habibzadeh F. Reinterpretation of the results of randomized clinical trials. *PLoS One* 2024;19(6):e0305575.

[13] Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav* 2018;2(1):6–10.

[14] Ioannidis JPA. The proposal to lower P value thresholds to .005. *JAMA* 2018;319(14):1429–30.

[15] Pérez ME, Pericchi LR. Changing statistical significance with the amount of information: the adaptive α significance level. *Stat Probab Lett* 2014;85:20–4.

[16] Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311(7003):485.

[17] Kruschke JK. Rejecting or accepting parameter values in Bayesian estimation. *Adv Methods Pract Psychol Sci* 2018;1(2):270–80.

[18] McElreath R. Statistical rethinking: a Bayesian course with examples in R and Stan. 2nd ed. Boca Raton: CRC Press; 2020.

[19] Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. 3rd ed. Boca Raton: CRC Press; 2013.