International Epidemiological Association

International Journal of Epidemiology, 2019, 1–9
doi: 10.1093/ije/dyz241
Original article

Original article

Most recommended medical interventions reach *P*<0.005 for their primary outcomes in meta-analyses

Despina Koletsi,[1,2] Marco Solmi,[3,4] Nikolaos Pandis,[5] Padhraig S Fleming,[6] Christoph U Correll[7,8,9,10] and John P A Ioannidis[11,12,13,14,15]*

[1]Department of Orthodontics, School of Dentistry, National and Kapodistrian University of Athens, Athens, Greece, [2]Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Zurich, Switzerland, [3]Department of Neuroscience, University of Padua, Padua, Italy, [4]Padua Neuroscience Center, University of Padua, Padua, Italy, [5]Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine, Medical Faculty, University of Bern, Bern, Switzerland, [6]Department of Oral Bioengineering, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK, [7]Department of Psychiatry, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY, USA, [8]Department of Psychiatry and Molecular Medicine, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA, [9]The Feinstein Institute for Medical Research, Center for Psychiatric Neuroscience, Manhasset, NY, USA, [10]Department of Child and Adolescent Psychiatry, Charité Universitätsmedizin Berlin, Berlin, Germany, [11]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA, [12]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA, [13]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA, [14]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA and [15]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA

*Corresponding author. Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA. E-mail: jioannid@stanford.edu

Editorial decision 24 October 2019; Accepted 1 November 2019

Abstract

Background: It has been proposed that the threshold of statistical significance should shift from *P*-value < 0.05 to *P*-value < 0.005, but there is concern that this move may dismiss effective, useful interventions. We aimed to assess how often medical interventions are recommended although their evidence in meta-analyses of randomized trials lies between *P*-value $= 0.05$ and *P*-value $= 0.005$.

Methods: We included Cochrane systematic reviews (SRs) published from 1 January 2013 to 30 June 2014 that had at least one meta-analysis with GRADE (Grading of Recommendations Assessment, Development and Evaluation) assessment and at least one primary outcome having favourable results for efficacy at *P*-value < 0.05. Only comparisons of randomized trials between active versus no treatment/placebo were included. We then assessed the respective UpToDate recommendations for clinical

practice from 22 May 2018 to 5 October 2018 and recorded how many treatments were recommended and what were the *P*-values in their meta-analysis evidence. The primary analysis was based on the first-listed outcomes.

Results: Of 608 screened SRs with GRADE assessment, 113 SRs were eligible, including 143 comparisons of which 128 comparisons had first-listed primary outcomes with UpToDate coverage. Altogether, 60% (58/97) of interventions with *P*-values < 0.005 for their evidence were recommended versus 32% (10/31) of those with *P*-value 0.005–0.05. Therefore, most (58/68, 85.2%) of the recommended interventions had *P*-values < 0.005 for the first-listed primary outcome. Of the 10 exceptions, 4 had other primary outcomes with *P*-values < 0.005 and another 4 had additional extensive evidence for similar indications that would allow extrapolation for practice recommendations.

Conclusions: Few interventions are recommended without their evidence from meta-analyses of randomized trials reaching *P*-value < 0.005.

Key words: *P*-value, statistical threshold, recommendation, meta-analysis, Cochrane, UpToDate

Key Messages

- Among treatments that have evidence summarized in meta-analyses of randomized trials and that are also recommended, the large majority reach $P < 0.005$ in primary outcomes for their evidence.
- In most of the exceptions landing between $P = 0.05$ and 0.005, there is additional extensive evidence that would allow for recommendations.
- Clinical and statistical significance are different concepts and statistical significance may not suffice for recommending an intervention, but requiring at least $P < 0.005$ for statistical significance seems reasonable, as it will not result in many recommended interventions being discarded.

Introduction

Thresholds are widely used by researchers and authors to claim statistical significance and to interpret research findings.[1,2] The most commonly reported threshold is that of '*P*-value < 0.05'. However, the use and misuse of *P*-values to determine the effectiveness of an intervention has received a great amount of criticism. *P*-values alone do not prove the clinical relevance of the effect.[3] Use of thresholds that claim statistical significance of the results do not fully substantiate effectiveness of a treatment for clinical practice or vice versa, although statistical significance is used to inform clinical significance. For example, a non-significant finding for an intervention of interest does not necessarily mean that the effect of this treatment is not clinically meaningful, as non- optimal sample sizes or low event rates may overshadow the true effect.[4–6]

Following the 2016 statement on *P*-values by the American Statistical Association (ASA),[7] discussions have been rekindled about how to improve the use of statistical inference tools. One possibility is to just abandon *P*-value thresholds entirely,[8] since they are so widely misused and misinterpreted. However, simply maintaining *P*-values without thresholds may cause statistical anarchy in the largely statistically-undertrained community of researchers.[9] Current lack of training impedes the widespread adoption of possibly better alternatives e.g. Bayesian or false-discovery rate approaches. Moreover, *P*-value thresholds have already been widely, almost ubiquitously, used in past published papers, including randomized trials and meta-analyses thereof. One proposal is to lower the routinely used *P*-value threshold from 0.05 to 0.005 to claim statistical significance with *P*-values 0.005–0.05 being merely 'suggestive'.[10,11] Endorsement of lower thresholds would reduce 'positive' results and may thus decrease false-positives, since $P = 0.005$ confers almost a 10-fold larger Bayes factor against the null than $P = 0.05$. The shift to the more conservative *P*-value threshold may also encourage the conducting of fewer, well-designed and larger studies with increased power to satisfy these thresholds. The effects of the currently used statistical standards on the credibility of research claims are not confined to biomedical science.[10] Empirical evidence shows that research from disciplines other than medicine or genetic epidemiology may achieve substantial gains if they adopt more

stringent levels of statistical significance to substantiate their claims. Some reports that come from studies on psychology[12] or experimental economics[13] reveal a nearly double rate of replication of these studies given the adoption of $P < 0.005$, compared with P- value between 0.05 and 0.005.

However, counter-arguments also exist.[14,15] A more stringent threshold may discard true-positive results, cultivate more extreme selective outcome reporting and promote use of surrogate endpoints (that more easily reach lower P-values).

For clinical studies that are already completed and reported and for the assessment of their evidence, the choice of threshold pertains mostly to the trade-off between false-positives and true-positives. Results of individual trials are typically included in meta-analyses and these are used eventually for recommendations for practice. Ideally, one wants to recommend effective, useful interventions and avoid recommending those that are not effective or useful. Statistical significance of the evidence is one among many considerations influencing treatment recommendations, and careful assessment of the quality/strength of the evidence [e.g. as appraised by Grading of Recommendation, Assessment, Development, and Evaluation (GRADE)][16,17] is pivotal. Regardless, although clinical significance is a much broader construct than statistical significance, it would be useful to understand empirically what levels of statistical significance are currently linked to recommendations for clinical use.

Here, we identified the P-value of the evidence for primary efficacy outcomes for various treatments that had reached the traditional statistical significance (P-value < 0.05) in Cochrane meta-analyses in various fields of medicine. We aimed to assess whether lowering the P-value threshold by one decimal point to 0.005 would discard many treatments that are clearly recommended.

Methods

Protocol registration

The protocol for this meta-epidemiological study was developed and *a priori* registered in the Center for Open Science, Open Science Framework (osf: https://osf.io/4phzy/).

Study selection and eligibility criteria

We considered all Cochrane Database of Systematic Reviews (CDSR) systematic reviews (SRs) published from 1 January 2013 to 30 June 2014 from the CDSR, a set that we had also used in previous work.[18] We searched for newer updated versions of these reviews published until May 2018; whenever such versions existed, we used the most updated version.

We used SRs including at least one comparison (meta-analysis) with GRADE assessment and at least one primary outcome having favourable statistically significant results at P-value < 0.05 for efficacy/effectiveness, or equivalently excluding the null in the 95% confidence interval (CI). We only kept comparisons between treatment (active) versus no treatment/placebo, rather than head-to-head comparisons between different active treatments (head-to-head compared treatments may both be very effective and recommended, but their differences may be negligible and non-significant). When different active treatments versus no treatment/placebo comparisons were included in the same SR, they were considered separately.

SRs and comparisons within SRs were excluded when they did not report any primary outcome with P-value < 0.05 and when they had primary outcomes with P-value < 0.05 favouring the no treatment/placebo arm. Comparisons including primary studies other than randomized controlled trials (RCTs) were excluded. Primary outcomes that pertained to toxicity/harms from therapy were also excluded.

Data extraction

Data were extracted on SR level, on comparison (i.e. intervention) level as well as on outcome level, forming a three-level hierarchical structure for the dataset. Our aim was to embrace all levels of information, as there might be >1 primary outcome per comparison recorded, or >1 comparison for the same SR.

At the SR level, we extracted: author, year, country, and region of publication, whether the review was new or an update, Cochrane group involved (as an indicator of medical domain), SR type (interventional, diagnostic). At the comparison level, every comparison comprising of treatment versus no treatment/placebo within the same SR was considered eligible, and information on the category of the intervention (including surgical, pharmacologic, behavioural or medical treatments, and diet or exercise interventions) was obtained. On the outcome level, we recorded: type of outcome [objective (mortality or outcomes assessed with an instrument or preset measurable criteria) or subjective], subtype of outcome (mortality, pain, quality of life, other), quality of the evidence according to GRADE (very low, low, moderate, high), number of studies included, effect size metric (odds ratio, risk ratio, risk difference, mean difference, standardized mean difference, other), point estimate (the summary effect as per SR authors' decision to analyse using either random or fixed effects), 95% CI and exact P-value (when not reported, we calculated it from the effect size and 95% CI). Additionally, and on an exploratory basis, we extracted data on between-study heterogeneity I^2 and between-study variance.

Recommendations for practice were extracted from UpToDate chapters between 22 May 2018 and 5 October 2018[19] on recorded treatments/interventions from the included SRs. We recorded whether a treatment modality was clearly recommended, described as an option in some circumstances, not recommended, or recommended against. Although we originally anticipated that it would have been useful to also record the relative prioritization of recommended treatments (e.g. first-line, second-line, etc.), in-depth examination of UpToDate showed that this gradation is rarely stated, whereas it is typically discernible whether a treatment is clearly recommended or recommended as an option under some circumstances.

Main outcomes and statistical analyses

The proportion of *P*-values ranging from 0.05 to 0.005 were recorded from the distribution of *P*-values (*P*-curve) for all eligible comparisons, using the first-listed primary outcome in the GRADE summary of findings (SoF) tables in the SRs when multiple eligible primary outcomes existed. In each case, we assessed which interventions were in the border zone (*P*-value 0.05–0.005) and were clearly recommended for clinical use. We examined whether the proportion of recommended treatments was different in *P*-value strata (<0.005 versus 0.005–0.05). Further, we also assessed separately the more granular categories (clearly recommended, option in some circumstances, not recommended, recommended against) in terms of their distributions of *P*-values. The primary analysis was based on the first-listed outcomes across all comparisons within the SRs. Whenever we identified outcomes in the opposite direction to the primary (first-listed) one for the same comparison, these were assigned a *P*-value = 1.00.

Cross-tabulations were conducted for the association between the *P*-value category (0.005–0.05 or <0.005) and recommendation category, intervention, outcome type/subtype, GRADE category, amount of heterogeneity (I^2<50% or ≥50%), and tertiles of sample size. Pearson chi-squared and Fisher's exact tests were conducted as appropriate. On an exploratory basis, we undertook an ordinal logistic regression for the effect of *P*-value (0.005–0.05 or <0.005) and GRADE on recommendation.

Sensitivity analyses

We conducted sensitivity analyses using only the highest, only the lowest, and only the geometric mean *P*-value from each eligible comparison, when multiple eligible primary outcomes were identified.

Furthermore, analyses were conducted where all comparisons were re-analysed using random effects models.

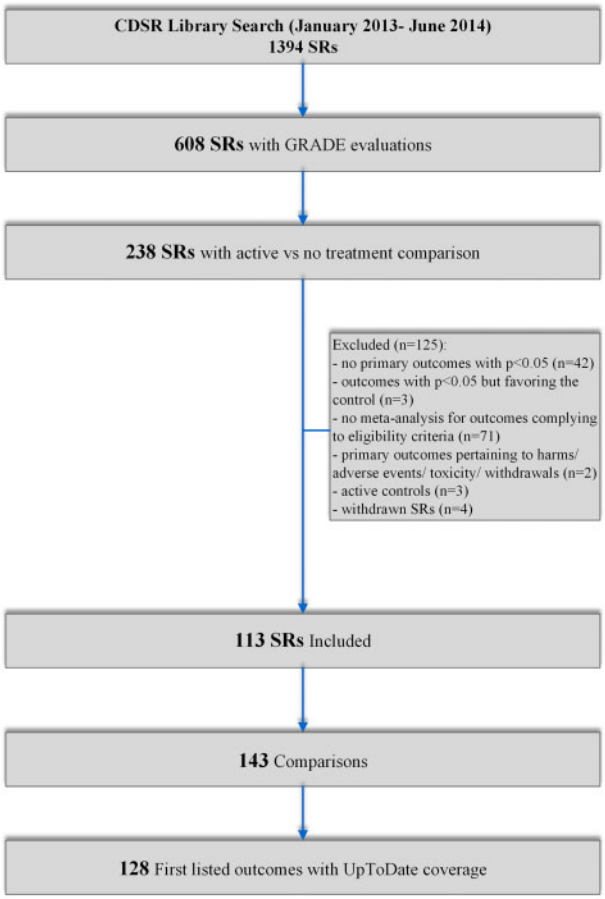

Figure 1. Flow chart of systematic review (SR) selection.

We accessed statistical data of each SR through the Cochrane Library and the respective RevMan files.

Analyses were performed with STATA version 15.1 software (Stata Corporation, College Station, Tex, USA).

Results

From an initial pool of 1394 SRs across all medical domains and covering 18 months, we had previously identified 608 SRs with included GRADE evaluations, routinely within SoF Tables. Of these, 113 fit the predetermined inclusion criteria, encompassing 143 eligible comparisons of active versus no treatment/placebo, and 128 of those treatments/indications had been covered by UpToDate (Figure 1). Characteristics of topics and outcomes are presented for those with *P*-value 0.005–0.05 and those with $P < 0.005$ (Table 1).

Primary analysis

For the 128 first-listed primary outcomes of treatments that were covered by UpToDate, only 31 (24.2%) reported a *P*-value between 0.05 and 0.005. Altogether, 60% (58/

Table 1. Characteristics according to *P*-value dichotomized categories (i.e. 0.005–0.05 and <0.005) for the evidence on the first-listed primary outcome ($n = 128$ treatments also covered by UpToDate)

	P-value				Total		*P*-value
	0.005–0.05		<0.005				
	No.	%	No.	%	No.	%	
Intervention							0.99[a]
Behavioural/medical device	6	19.4	18	18.6	24	18.8	
Diet/exercise	3	9.7	10	10.3	13	10.2	
Pharmacological	22	70.9	69	71.1	91	71.0	
Type of outcome							0.36[b]
Objective	17	54.8	44	45.4	61	47.7	
Subjective	14	45.2	53	54.6	67	52.3	
Subtype of outcome							0.11[a]
Morbidity	15	48.4	28	28.9	43	33.6	
Mortality	4	12.9	7	7.2	11	8.6	
Pain	4	12.9	31	32.0	35	27.3	
Quality of life	1	3.2	4	4.1	5	3.9	
Other	7	22.6	27	27.8	34	26.6	
I^2 (%)							0.18[b]
0–49	18	58.1	69	71.1	87	68.0	
50–100	13	41.9	28	28.9	41	32.0	
Sample size (tertiles)							0.09[b]
60–533	15	48.4	27	27.8	42	32.8	
537–1378	7	22.6	36	37.1	43	33.6	
1401–99 797	9	29.0	34	35.1	43	33.6	
Total	31	100.0	97	100.0	128	100.0	

[a]Fisher's exact.
[b]Pearson chi-square.

97) of the interventions with *P*-values < 0.005 for their evidence were recommended versus 32% (10/31) of those with *P*-value 0.005–0.05 (odds ratio = 3.12; 95% CI 1.34, 7.24; $p = 0.008$). Most (58/68, 85.2%) of the recommended interventions had *P*-values < 0.005 in the first-listed primary outcome. The respective proportions were 16/22 (72.7%) for optional interventions, 17/25 (68.0%) for the not recommended interventions and 6/13 (46.2%) for interventions recommended against. A high-level of evidence according to GRADE was seen in 23/97 (23.7%) of *P*-values < 0.005 and 3/31 (9.7%) of *P*-values 0.005–0.05 (odds ratio = 2.90; 95% CI 0.85, 9.73; $P = 0.09$).

Since both *P*-value and GRADE might be taken into consideration when making recommendations for practice, we considered both of them in a logistic regression. The odds of a treatment being recommended increased 2.75-fold with *P*-value < 0.005 (95% CI 1.26, 6.02) after adjusting for GRADE (Table 2).

When multiple outcomes were available ($n = 65$ topics), for only one topic was one of them statistically significant in the opposite direction than the first listed primary outcome. This pertained to the risk for pneumonia after the use of combined inhalers for chronic obstructive pulmonary disease.

Table 2. Ordinal logistic regression for the effect of *P*-value (treated as a binary variable, i.e. 0.005–0.05 or <0.005) and GRADE (very low, low, moderate, high) on recommendations for clinical use

	Odds ratio	95%CIs	*P*- value
P-value			
0.005–0.05	Reference		
<0.005	2.75	1.26, 6.02	0.01
GRADE			0.03[a]
Very low	Reference		
Low	3.13	0.98, 10.01	0.05
Moderate	4.39	1.51, 12.75	0.007
High	5.74	1.67, 19.69	0.005

[a]Wald test for GRADE.

Sensitivity analysis

The association of *P*-value <0.005, rather than 0.005–0.05, with recommended interventions remained similar when we examined the lowest *P*-value across multiple eligible outcomes for each eligible comparison, the highest *P*-value or the geometric mean of the *P*-values (Supplementary Table 1, available as Supplementary data at *IJE* online).

When we calculated the *P*-values for all eligible first-listed outcomes by random effects estimates of the summary measures, the overall picture was very similar. The number of treatments that were recommended based on the standard *P*-value threshold of <0.05 (but not <0.005) increased by only 5 (Supplementary Table 2, available as Supplementary data at *IJE* online). Results remained similar for the lowest, the highest and the geometric means of *P*-values according to random effects (Supplementary Table 3, available as Supplementary data at *IJE* online).

Recommendations despite P≥0.005

Based on our primary analysis, we recorded 10 first-listed outcomes where the active treatments were recommended for clinical practice and where the reported summary of the evidence had a *P*-value in the zone of 0.05 and 0.005, thus not conforming to the proposed rule of having *P*-value < 0.005 (Table 3). Of those, four comparisons reported a *P*-value for another examined efficacy outcome (apart from the first-listed) that was <0.005. These treatments pertained to vaccines (oral Ty21a) for preventing typhoid fever,[20] supplementation with folic or folinic acid for methotrexate-receiving rheumatoid arthritis patients,[21] oral antibiotics for chronic obstructive

Table 3. List of the recorded 10 treatments that were recommended by UpToDate and which had *P*-value 0.005–0.05 for the first-listed outcomes in reviews in the Cochrane Database of Systematic Reviews

Disease	Intervention	First-listed outcome	No. of outcomes	*P*-primary (first-listed)
Post solid organ transplant tuberculosis (TB)	Isoniazid	Risk of developing TB post-transplant	1	0.027
Typhoid fever in adults and children aged ≥5 years of age	Oral Ty21a (3 doses)	Cases of typhoid fever, year 1	4	0.0093
Adults with chronic obstructive pulmonary disease (COPD)	Administration of an oral prophylactic antibiotic continuously or intermittently	Number of people with one or more exacerbations	3	0.01
Pregnant women at increased risk of fetal complications	Fetal and umbilical Doppler ultrasound	Any perinatal death	1	0.037
Adults with cluster headache	Intranasal zolmitriptan 5 mg	Pain free at 30 min	4	0.015
Patients receiving methotrexate for rheumatoid arthritis	Supplementation with either folic or folinic acid	Incidence of nausea, vomiting or abdominal pain (gastrointestinal effects)	4	0.0083
Gastro-oesophageal reflux disease-like symptoms	H2-receptor antagonists	Heartburn	1	0.04
Endoscopy negative reflux disease	H2-receptor antagonists	Heartburn	1	0.0055
People with *S. mansoni* infection	Praziquantel 40 mg/kg	Parasitological failure	1	0.044
Cytomegalovirus (CMV) disease in solid organ transplant recipients	Pre-emptive medication for CMV viraemia	CMV disease	4	0.017

pulmonary disease[22] and zolmitriptan for cluster headache.[23] The remaining 6 were isoniazid for post-solid organ transplant tuberculosis,[24] Doppler ultrasound for pregnant women,[25] H2-receptor antagonists for gastro-oesophagyal reflux or endoscopy,[26] praziquantel for *Schistosoma mansoni* infection[27] and pre-emptive treatment for cytomegalovirus (CMV) viraemia in solid organ transplant recipients[28] (Table 3). With the exception of Doppler ultrasound in pregnancy and praziquantel for *S. mansoni*, the interventions had been extensively tested for other similar indications: isoniazid has been tested for many prophylaxis settings; H2-receptor antagonists have been used extensively to relieve symptoms in various settings; and pre-emptive CMV treatment has been used in many immunosuppressed and transplant groups.

Using random effects calculations, 4 of the 5 additional interventions with *P*-value 0.005–0.05 were related to smoking cessation assessed in specialized settings (post-depression, pre-operatively, workplace).[29–31] Smoking cessation has far more evidence across diverse settings. The last intervention was maintenance with all-*trans* retinoic acid/arsenic trioxide for acute promyelocytic leukaemia,[32] which is associated with a very large effect and is a classic, highly-cited intervention in hematology.

In all, the vast majority of treatments bearing evidence summarized in meta-analyses of RCTs and being recommended for use, reached $P < 0.005$ in their primary outcomes. The few exceptions were easily explained.

Discussion

We found that among treatments that have evidence summarized in meta-analyses of randomized trials in the CDSR and reach *P*-value < 0.05 for their first-listed primary outcome, only about one-quarter do not also reach *P*-value < 0.005. Furthermore, few of those that are recommended by UpToDate do not reach *P*-value < 0.005. Those that are recommended, but do not reach that more conservative level of statistical significance almost always have some other primary outcome (other than the first-listed) that reaches *P*-value < 0.005, or they have additional extensive evidence for similar indications. In our series of examined treatments, only two interventions had no *P*-values < 0.005 for any primary efficacy outcome and the examined outcome and had also no or little other favourable evidence for similar indications. Praziquantel for *S. mansoni*[27] is not so highly statistically significant, however it has a large treatment effect (risk ratio = 3.13). Moreover, if indications were to be extended to other parasites besides *S. mansoni*, praziquantel does show effectiveness also against other types of schistosomiasis, clonorchiasis, opisthorchiasis, tapeworm infections, cysticercosis, hydatid disease and other fluke infections. Finally, ultrasound during

pregnancy is a safe and simple intervention recommended for use, even if the *P*-value for the evidence on reducing perinatal deaths is not that low.[25] In all, moving the *P*-value threshold to 0.005 would not result in many recommended treatments being 'discarded'. In addition, nearly two-thirds of the treatments where outcomes that 'surpassed' the *P*-value < 0.005 threshold, were actually recommended for clinical practice. Thus, in terms of a cost–benefit evaluation of the proposed threshold, the results of this empirical study show evidence of matching of the new threshold to clinical decision making, when statistical assessment of the findings is considered, without risking losing otherwise recommended treatment modalities.

We are aware of only one other empirical study that evaluated the 0.05 versus 0.005 *P*-value threshold for published clinical evidence, focusing on the results of RCTs (not of meta-analyses, as we did) in major medical journals.[33] The authors evaluated 272 primary outcomes from 203 RCTs and recorded 174 outcomes with a *P*-value < 0.05. They found that shifting the threshold of significance to 0.005 would affect 51 of 174 outcomes (29.3%). Across the entire biomedical literature indexed in PubMed, about one-third of the papers that claim statistically significant results would fit to the 0.005–0.05 bin,[34] but the impact on decision making and recommendations for clinical practice has not been evaluated.

There is extensive theoretical and other work criticizing the use and misuse of *P*-values and statistical significance.[1,7,15,35–39] Examples of *P*-value misuse are outcome switching or selective outcome reporting practices based on the perceived significance of the outcome of interest. Should the adoption of lower *P*-value thresholds be upheld, attention must be drawn to the use of surrogate outcomes that may pass the threshold more easily, since these may be prioritized by researchers in an attempt to retrieve findings that satisfy the new threshold.[40,41] For this reason, some argue abandoning statistical significance thresholds or even *P*-values in most applications, and enhancing focus on effect sizes and their uncertainty and other inferential tools, e.g. Bayes factors[42,43] and false-discovery rates.[44,45] However, given that *P*-values are still so widely used and practically ubiquitous in the published literature, it will take continuous, extensive training of the scientific workforce to achieve drastic changes.

RCTs and their meta-analyses guide clinical practice and inform clinicians on treatment decisions. The fact that almost all recommended interventions have $P < 0.005$ for their evidence on the same or similar indications does not diminish the need to place evidence into the appropriate context for each patient. This includes the postulated magnitude of benefits and harms in the specific patient being considered and the specific setting, as well as cost and convenience, and alternative options and their merits and drawbacks. In addition, the fact that we also found a non-negligible amount of treatments having a $P < 0.005$ and not being recommended for clinical practice, at least within the electronic source we used, upholds the claims about the necessity for disentanglement between statistical and clinical significance.[5] This was further elucidated by our primary analysis findings which demonstrated that a very low *P*-value for an outcome of interest could also be followed by a recommendation against the use of a treatment modality in clinical practice.

Some limitations should be discussed. First, the *P*-value derived in a meta-analysis depends on the statistical model used. Our main analysis used the systematic reviewers' choice of statistical methods to summarize data (either fixed or random effects). An analysis using solely calculated *P*-values by random effects yielded similar patterns. Usually, in the presence of heterogeneity, random effects give estimates with higher *P*-values,[46] but exceptions may occur. Second, we focused on a set of SRs that would give a clean answer for evaluation of the efficacy/effectiveness of treatments versus no treatment/placebo. We excluded any head-to-head comparisons of active interventions that would hinder the determination of the net treatment effects of an intervention. Head-to-head comparisons of active treatments have more complex statistical inferences and decisions may often be based on non-inferiority rather than superiority. Moreover, as network meta-analyses and indirect comparisons become more popular, the indirect evidence may also affect the perception about the effectiveness of a treatment. Additionally, we did not appraise outcomes related to harms or toxicity. Harms are also appraised in different ways to primary efficacy outcomes in RCTs. The use of statistical inferences to substantiate evaluation of outcomes pertaining to adverse effects may be even more problematic or misleading, as assessment of harms/toxicity involves confirming a null effect rather than confirming the presence of an effect.

Finally, UpToDate[19] recommendations are not a perfect gold standard. The authors of this resource may express personal opinions. Moreover, some opinions were not easy to categorize using our preconceived categories. UpToDate uses in-house trained physician editors, along with the use of GRADE, to provide information about the strength of the recommendations for clinical practice.[47] UpToDate frames recommendations around a specific topic or clinical question based on the PICO (participant, intervention, comparator, outcome) format. The recommendations are based on the best available evidence (preferably from well-designed systematic reviews), while patient values and a cost–benefit trade-off are considered as well. Furthermore, different evidence tools, e.g. guidelines by professional societies or other agencies, may have reached different conclusions, and actual extent of clinical use in real life may not fully square with UpToDate recommendations. Nevertheless,

we preferred to use UpToDate because it has wide coverage of medicine[48] and it is generally considered to be less affected by bias from professional societies' guidance.[49] A selection bias may exist if UpToDate reviewers focused on discussing preferentially interventions that had evidence with relatively lower *P*-values. However, we do not believe this is likely, since the recommendations are provided by a panel of in-house trained medical editors based on the best available evidence overall. In addition, it is unlikely that interventions that do not even reach the lenient $P < 0.05$ for their evidence on primary outcomes would be recommended.

Allowing for these caveats, we note that clear recommendations favouring the use of a treatment are mostly associated with evidence from RCTs that reach $P < 0.005$ when summarized for efficacy in meta-analyses. Asking for a more conservative $P < 0.005$ for claiming 'statistical significance' would not translate into the overriding of many effective and clearly recommended treatments. Statistical significance (regardless of threshold used) should not be confused with clinical significance: many treatments with very low *P*-values for some outcomes are still not recommended for clinical use when the broader picture (i.e. patient values and cost–benefit trade-off) is considered. However, the vast majority of interventions that are clearly recommended seem to have substantial statistical support with $P < 0.005$.

Supplementary data

Supplementary data are available at *IJE* online.

Funding

METRICS is supported from a grant by the Laura and John Arnold Foundation. J.I. is supported by an unrestricted gift by Sue and Bob O'Donnell to Stanford Prevention Research Center. The research was conducted independent of any involvement of the funders. Funders were not involved in any aspect of the study design, data collection, data interpretation, writing, or the decision to submit the article for publication.

Authors' contributions

All authors contributed to the design of the study. D.K. and M.S. performed data extraction. D.K. with help from N.P. and J.P.A.I. performed statistical analyses. All authors interpreted the data and results. D.K. wrote the first draft with help from J.P.A.I. and all authors revised the paper and approved the final version. J.P.A.I. is the guarantor. All authors had full access to all of the data.

Conflict of interest: None declared.

References

1. Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med* 1999;**130**:995–1004.
2. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: an explanation for new researchers. *Clin Orthop Relat Res* 2010;**468**:885–92.
3. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;**311**:485.
4. Sedgwick P. Clinical significance versus statistical significance. *BMJ* 2014;**348**:g2130.
5. Mellis C. Lies, damned lies and statistics: clinical importance versus statistical significance in research. *Paediatr Respir Rev* 2018;**25**:88–93.
6. Page P. Beyond statistical significance: clinical interpretation of rehabilitation research literature. *Int J Sports Phys Ther* 2014;**9**:726–36.
7. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat* 2016;**70**:129–33.
8. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;**567**:305–307.
9. Ioannidis J. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA* 2019;**321**:2067.
10. Benjamin DJ, Berger JO, Johannesson M *et al*. Redefine statistical significance. *Nat Hum Behav* 2018;**2**:6–10.
11. Ioannidis J. The proposal to lower P value thresholds to.005. *JAMA* 2018;**319**:1429–30.
12. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 2015;**349**:aac4716.
13. Camerer CF, Dreber A, Forsell E *et al*. Evaluating replicability of laboratory experiments in economics. *Science* 2016;**351**:1433–36.
14. Lakens D, Adolfi FG, Albers CJ *et al*. Justify your alpha. *Nat Hum Behav* 2018;**2**:168–71.
15. Greenland S, Senn SJ, Rothman KJ *et al*. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;**31**:337–50.
16. Guyatt G, Oxman AD, Akl EA *et al*. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;**64**:383–94.
17. Schünemann HJ, Oxman AD, Higgins JP, Vist GE, Glasziou P, Guyatt GH. *Chapter 11: Presenting Results and Summary of Findings Tables [Internet]*. https://handbook-5-1.cochrane.org/chapter_11/11_presenting_results_and_summary_of_findings_tables.htm (13 April 2018, date last accessed).
18. Fleming PS, Koletsi D, Ioannidis JPA, Pandis N. High quality of the evidence for medical and other health-related interventions was uncommon in Cochrane systematic reviews. *J Clin Epidemiol* 2016;**78**:34–42.
19. Waltham M. *UpToDate Inc*. Post TW (ed). https://www.uptodate.com.
20. Milligan R, Paul M, Richardson M, Neuberger A. Vaccines for preventing typhoid fever. *Cochrane Database Syst Rev* 2018;**5**:CD001261.
21. Shea B, Swinden MV, Tanjong Ghogomu E *et al*. Folic acid and folinic acid for reducing side effects in patients receiving methotrexate for rheumatoid arthritis. *Cochrane Database Syst Rev* 2013;**5**:CD000951.
22. Herath SC, Poole P. Prophylactic antibiotic therapy for chronic obstructive pulmonary disease (COPD). *Cochrane Database Syst Rev* 2013;**11**:CD009764.

23. Law S, Derry S, Moore RA. Triptans for acute cluster headache. *Cochrane Database Syst Rev* 2013;**7**:CD008042.
24. Adamu B, Abdu A, Abba AA, Borodo MM, Tleyjeh IM. Antibiotic prophylaxis for preventing post solid organ transplant tuberculosis. *Cochrane Database Syst Rev* 2014;**3**:CD008597.
25. Alfirevic Z, Stampalija T, Dowswell T. Fetal and umbilical Doppler ultrasound in high-risk pregnancies. *Cochrane Database Syst Rev* 2017;**6**:CD007529.
26. Sigterman KE, Pinxteren B. V, Bonis PA, Lau J, Numans ME. Short-term treatment with proton pump inhibitors, H2-receptor antagonists and prokinetics for gastro-oesophageal reflux disease-like symptoms and endoscopy negative reflux disease. *Cochrane Database Syst Rev* 2013;**5**:CD002095.
27. Danso-Appiah A, Olliaro PL, Donegan S, Sinclair D, Utzinger J. Drugs for treating Schistosoma mansoni infection. *Cochrane Database Syst Rev* 2013;**2**:CD000528.
28. Owers DS, Webster AC, Strippoli GFM, Kable K, Hodson EM. Pre-emptive treatment for cytomegalovirus viraemia to prevent cytomegalovirus disease in solid organ transplant recipients. *Cochrane Database Syst Rev* 2013;**2**:CD005133.
29. Meer R. V D, Willemsen MC, Smit F, Cuijpers P. Smoking cessation interventions for smokers with current or past depression. *Cochrane Database Syst Rev* 2013;**8**:CD006102.
30. Thomsen T, Villebro N, Møller AM. Interventions for preoperative smoking cessation. *Cochrane Database Syst Rev* 2014;**3**:CD002294.
31. Cahill K, Lancaster T. Workplace interventions for smoking cessation. *Cochrane Database Syst Rev* 2014;**2**:CD003440.
32. Muchtar E, Vidal L, Ram R, Gafter-Gvili A, Shpilberg O, Raanani P. The role of maintenance therapy in acute promyelocytic leukemia in the first complete remission. *Cochrane Database Syst Rev* 2013;**3**:CD009594.
33. Wayant C, Scott J, Vassar M. Evaluation of lowering the P value threshold for statistical significance from.05 to.005 in previously published randomized clinical trials in major medical journals. *JAMA* 2018;**320**:1813–15.
34. Chavalarias D, Wallach JD, Li AHT, Ioannidis J. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA* 2016;**315**:1141–48.
35. Szucs D, Ioannidis J. When null hypothesis significance testing is unsuitable for research: a reassessment. *Front Hum Neurosci* 2017;**11**:390.
36. Gigerenzer G. Mindless statistics. *J Socio Econ* 2004;**33**:587–606.
37. Gigerenzer G, Krauss S, Vitouch O. The null ritual: what you always wanted to know about significance testing but were afraid to ask. In: *The SAGE Handbook of Quantitative Methodology for the Social Sciences [Internet]*. Thousand Oaks, CA, USA: SAGE Publications, Inc., 2004, pp. 392–409. http://methods.sagepub.com/book/the-sage-handbook-of-quantitative-methodology-for-the-social-sciences/n21.xml (18 February 2019, date last accessed).
38. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol* 2008;**45**:135–40.
39. McCloskey D, Ziliak S. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives [Internet]*. Ann Arbor, MI: University of Michigan Press, 2008. https://hdl.handle.net/2027/fulcrum.cr56n193q (18 February 2019, date last accessed).
40. Fleming PS, Koletsi D, Dwan K, Pandis N. Outcome discrepancies and selective reporting: impacting the leading journals? *PLoS One* 2015;**10**:e0127495.
41. Falk Delgado A, Falk Delgado A. Outcome switching in randomized controlled oncology trials reporting on surrogate endpoints: a cross-sectional analysis. *Sci Rep* 2017;**7**:9206.
42. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. *Clin Trials* 2005;**2**:282–90; discussion 301–**304**, 364–78.
43. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med* 1999;**130**:1005–13.
44. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995;**57**:289–300.
45. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Statist* 2001;**29**:1165–88.
46. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 2010;**1**:97–111.
47. Agoritsas T, Merglen A, Heen AF *et al*. UpToDate adherence to GRADE criteria for strong recommendations: an analytical survey. *BMJ Open* 2017;**7**:e018593.
48. Johnson E, Emani VK, Ren J. Breadth of coverage, ease of use, and quality of mobile point-of-care tool information summaries: an evaluation. *JMIR Mhealth Uhealth* 2016;**4**:e117.
49. Kwag KH, González-Lorenzo M, Banzi R, Bonovas S, Moja L. Providing doctors with high-quality information: an updated evaluation of web-based point-of-care information summaries. *J Med Internet Res* 2016;**18**:e15.