BMJ Mental Health

STATISTICS

Understanding effect size: an international online survey among psychiatrists, psychologists, physicians from other medical specialities, dentists and other health professionals

Ferdinand Heimke ,[1] Yuki Furukawa,[2] Spyridon Siafis ,[1,3] Bradley C. Johnston,[4,5] Rolf R. Engel,[6] Toshi A Furukawa ,[7] Stefan Leucht [1,3]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjment-2023-300978).

For numbered affiliations see end of article.

Correspondence to
Professor Stefan Leucht, Department of Psychiatry and Psychotherapy, Technical University of Munich School of Medicine, München 81675, Germany; Stefan.Leucht@tum.de

Received 27 December 2023
Accepted 29 January 2024

ABSTRACT

Background and objective Various ways exist to display the effectiveness of medical treatment options. This study examined various psychiatric, medical and allied professionals' understanding and perceived usefulness of eight effect size indices for presenting both dichotomous and continuous outcome data.

Methods We surveyed 1316 participants from 13 countries using an online questionnaire. We presented hypothetical treatment effects of interventions versus placebo concerning chronic pain using eight different effect size measures. For each index, the participants had to judge the magnitude of the shown effect, to indicate how certain they felt about their own answer and how useful they found the given effect size index.

Findings Overall, 762 (57.9%) participants fully completed the questionnaire. In terms of understanding, the best results emerged when both the control event rate (CER) and the experimental event rate (EER) were presented. The difference in minimal importance difference units (MID unit) was understood worst. Respondents also found CER and EER to be the most useful presentation approach while they rated MID unit as the least useful. Confidence in the risk ratio ranked high, even though it was rather poorly understood.

Conclusions and clinical implications For dichotomous outcomes, presenting the effects in terms of the CER and EER could lead to the most correct interpretation. Relative measures including the risk ratio must be supplemented with absolute measures such as the CER and EER. Effects on continuous outcomes were better understood through standardised mean differences than mean differences. These can also be supplemented by dichotomised CER and EER.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Most effect size indices are poorly understood by clinicians.

⇒ The magnitude of effect size is most likely to be interpreted correctly when presented with dichotomous outcome measures.

⇒ How effect sizes are interpreted when they are presented using the control event rate (CER) and the experimental event rate (EER) instead of risk difference (ie, EER–CER) only has never been investigated.

WHAT THIS STUDY ADDS

⇒ Presenting results using the CER and the EER would lead to the best interpretation of the effect size.

⇒ Effect size presented with risk ratio is often misinterpreted while medical professionals indicate to have great confidence and perceived usefulness for risk ratio.

⇒ Relative outcome measures must be supplemented with absolute measures to avoid misinterpretation.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The findings of our study provide authors of scientific papers with a recommendation on how to present results in future papers in the most comprehensible way.

⇒ Further initiatives are needed to improve the education of health professionals in health research methodology.

BACKGROUND

Clinicians, patients and policymakers, when choosing among treatments, reach a decision based on numbers: numbers expressing the magnitude of their efficacy—effect size indices—derived typically from systematic reviews of randomised controlled trials (RCTs) or explicitly from large randomised trials.

The same effect can be expressed in various ways.[1] When the outcome of interest is a continuous variable, as is typically the case in mental health, the most common way of presentation is the standardised mean difference (SMD),[2 3] the difference in means in the experimental and control arms divided by their SD.[4] When the outcome is measured on the same scale, perhaps a more intuitive option is the simple mean difference (MD) of the two groups.[3] MD is also usually the primary index in the case of RCTs. There are also some other proposed ways of presenting continuous outcomes, including the ratio of means (RoM)[5] and the difference in means

Check for updates

To cite: Heimke F, Furukawa Y, Siafis S, *et al*. *BMJ Ment Health* 2024;**27**:1–7.

divided by each included instrument's minimal important differences (MID units).[6] For dichotomous variables, the commonly used effect sizes include the risk ratio (RR) and OR as relative measures as well as risk difference (RD) or number needed to treat (NNT) as absolute measures.[4 7] For dichotomous outcomes, the Cochrane Collaboration's Summary of Findings tables recommends showing both, the control event rate (CER) and the experimental event rate (EER), to facilitate evidence users' understanding of the results of systematic reviews.[8]

Previous research suggests that the same magnitude of effect, when expressed by different indices, may be interpreted differently by users of evidence. Akl *et al*[9] reviewed and compared the absolute risk reduction (which is the same as RD) and the relative risk reduction (which is converted from RR as (1.0–RR)×100, for example, 60% reduction instead of RR=0.4) and concluded that the two did not differ in terms of correct interpretations of the effect, but that the relative risk reduction greatly increased the willingness to adopt the intervention.[9] The RD and the NNT did not differ in terms of interpretability or persuasiveness.

Johnston *et al*[10] investigated clinicians' understanding and perception of the usefulness of six statistical formats for presenting outcomes from meta-analyses based on a hypothetical scenario about the treatment of chronic pain. Their results showed that all tested effect size indices were rather poorly understood (rate of correct responses lower than 50% for every tested index), especially those representing continuous outcomes (eg, SMD). However, Johnston *et al* did not examine some widely used or recommended indices, such as the NNT separately or the CER together with the EER. Moreover, the small versus large effect size values, presented in their questionnaire, were not calculated/determined consistently across indices, thereby making it difficult to interpret the proportion of correct responses for different indices. Finally, they surveyed clinicians in internal medicine and family medicine only and did not include doctors of other specialities or people from other healthcare professions.

OBJECTIVE

In this study, we examined various psychiatric, medical and allied professionals' understanding of eight different effect size indices for dichotomous and continuous outcome data, including SMD, MD, MID unit, RoM, RR, RD CER/EER and NNT. We aimed to find out which of those would be best suited to present the efficacy of medical treatments in the most comprehensible way. We also investigated respondents' confidence while dealing with these measures as well as their perceived usefulness. Finally, we evaluated the influence of various demographic characteristics on the understanding of the effect size indices. We chose chronic pain as an example to make the comparison with Johnston *et al*'s results easier and because it is a ubiquitous symptom for any professional or lay persons.

METHODS

We hereby report our study in accordance with the consensus-based checklist for reporting of survey studies.[11]

Participants

Since health research methodology and its understanding are integral to any health experts' activities, we decided to recruit broad range of medical professionals. Medical doctors of all specialities and training levels as well as dentists, medical and dental students and people from other healthcare professions (eg, psychologists, nurses, pharmacists) were eligible to participate. Participants had to be sufficiently proficient in English. We distributed the link to the online questionnaire (see online supplemental file 1) with an invitation and further explanations and descriptions about the project by email. We used mailing lists of hospitals, doctors' networks, and personal contacts.

Questionnaire

We created a digital questionnaire (see online supplemental file 1) to reach as many potential respondents as possible from as broad backgrounds as possible. To compare the results with the previous studies and because English is the lingua franca for science, the questionnaire was mainly in English, except for the introductory explanations which were presented in the local language when necessary to increase the accessibility of the questionnaire. To design and conduct the questionnaire, we used the online survey tool SoSci-Survey (V.3.3.13). The survey was completely anonymous. Tracing the participants' IP addresses was impossible and data protection was always guaranteed by secure internet communication and the secure website. The participants were informed about the processing of their data in the invitation. The questionnaire comprised two parts. In the first part, we asked about demographic and background information. The second part assessed participants' understanding and perceived usefulness of eight effect size indices: SMD, MID unit, MD, RoM for continuous outcomes; and RR, RD, CER and EER, and NNT for dichotomous effect measures.

Initially, we had presented a clinical scenario before we introduced the actual questions. The scenario described a hypothetical meta-analysis of randomised trials of interventions for patients with chronic non-cancer pain. Pain often shares a complex interplay with psychiatric diseases since the persistent nature of chronic pain can contribute to the development or exacerbation of psychiatric conditions such as depression and anxiety.[12]

Pain was measured on a visual analogue scale (VAS) between 0 (no pain) and 10 (worst pain ever). Before treatment, the average score on the VAS was approximately 6 points, as reported in a large-scale study of similar patients.[13] All subsequent questions were based on this scenario. For each of the eight effect size indices, we determined a small, medium and large treatment effect (table 1) in accordance with Cohen's rule of thumb, which defines a small effect as SMD=0.2, a medium effect as SMD=0.5 and a large effect as SMD=0.8.[14] Our exact approach for calculating and defining the required effect sizes is explained in online supplemental file 2.

In the digital questionnaire, each participant assessed the magnitude of the effect for all eight effect size indices. To reduce the respondents' burden and avoid response fatigue and errors, we chose only one of small, medium or large effects for each effect size index. Thus, for each index, the participants had to choose one of three possible answers (small effect, medium effect and large effect). The sequence of the eight indices as well as the presented effect size was randomised automatically to prevent order effects.

Additionally, the participants indicated how certain they felt about their own answers and how useful they found the given effect size index. For every effect size question, their confidence and perceived usefulness were assessed on a 7-point Likert scale, with response options ranging from 'not at all' (1 point) to 'extremely confident' (7 points) and 'not useful in understanding the size of the effect' (1 point) to 'extremely useful in understanding the size of the effect' (7 points), respectively. We carried out a pretest of our survey with the help of 20 individuals who were not involved in the project and who matched our survey target group.

Table 1 Small, medium and large treatment effects for all eight effect size indices

Effect size index	Small effect	Medium effect	Large effect
SMD	0.20	0.50	0.80
MD	0.50 points	1.25 points	2.00 points
MID unit	0.50	1.25	2.00
RoM	0.90 (10%)	0.75 (25%)	0.60 (40%)
RR	1.30 (30% more)	1.80 (80% more)	2.40 (140% more)
RD	0.06 (6%)	0.17 (17%)	0.28 (28%)
CER and EER	0.20/0.26 (20%/26%)	0.20/0.37 (20%/37%)	0.20/0.48 (20%/48%)
NNT	16.50	6.00	3.50

CER&EER, control event rate and experimental event rate; MD, mean difference; MID unit, difference in minimal importance difference units; NNT, number needed to treat; RD, risk difference; RoM, ratio of means; RR, risk ratio; SMD, standardised mean difference.

Survey period

The online questionnaire was accessible for participation for a period of 2 months, from 15 February 2022 to 15 April 2022.

Outcomes

Our primary outcome was the proportion of respondents who correctly understood each effect size index. Correct understanding was defined as the right estimation of the effect size (small, medium or large) that was presented with the respective index. Secondary outcomes were the respondents' confidence while dealing with these measures, their perceived usefulness as well as sociodemography and other factors that were associated with the understanding of the eight statistical formats.

Sample size

Assuming a proportion of correct answers at 50%[10] to achieve a 95% CI width of 10% (ie, margin of error of 5%), we calculated our required sample size to be at least 384.

Statistical analysis

We included only fully completed and returned questionnaires in the analysis and did descriptive statistics to summarise the respondents' characteristics. Then the proportion of correct answers for the questions about the magnitude of presented effects was calculated with corresponding 95% CI. We applied a multivariable logistic regression to examine the relative performance of the different indices. The index which produced the largest rate of correct answers was chosen as reference. We also compared the rate of correct answers for the small effect sizes with those affiliated with the medium and large effects. We displayed the influence of expertise in conducting systematic reviews and experience in health research methodology on the results and contrasted the performance of respondents of mental health professions with the rest of the participants. We summarised the respondents' confidence and perceived usefulness for each index as their mean scores on the 7-point Likert scale, reported with 95% CI. All statistical procedures were performed using Excel (V.2301) and R Software (V.4.2.2).

FINDINGS

Participants' characteristics

In total, 1316 people participated in our survey. Overall, 762 participants fully completed and returned the questionnaire, for a response rate of 57.9%. Respondents came from 13 different countries. Among those, Germany was the most frequent one (50.3%). 58.3% of the interviewed persons stated they had no experience in health research methodology. A small part of 124 participants (16.3%) had conducted at least one systematic review with meta-analysis by themselves (table 2).

Correct understanding of the effect size indices

The proportions of correctly evaluated magnitudes of effect-by-effect size indices varied between 43% and 56% (figure 1A). The best results were ascertained for CER and EER. Fifty-six per cent of the participants estimated a given effect size correctly if it was presented with CER and EER. The RD turned out to be the second best. SMD and NNT showed similar results. The RR ranked clearly lower, and the MID unit was the least understood.

Logistic regression

In the multivariable logistic regression taking CER and EER as reference, there was strong evidence that all the indices except for RoM and RD performed worse than the CER and EER by 5 percentage points or greater (table 3). Medium and large effect sizes tended to be more incorrectly estimated than small effect sizes. We also examined factors associated with correct understanding. The data suggested that education in health research methodology improved the assessment of given effect sizes to a small degree. There was no evidence that specialities (mental health vs others) or experience in conducting systematic reviews made any meaningful contributions.

Perceived confidence and usefulness

Respondents felt most confident about using CER and EER while they were least confident about using MID unit (figure 1B, online supplemental file 3). Likewise, they found CER and EER to be the most useful presentation approach while they rated MID unit as the least useful (figure 1C, online supplemental file 3). NNT, RD and RR were also highly appreciated. In both categories, RR ranked high even though it was rather poorly understood comparatively.

Understanding of the indices by the magnitude of the effect size

For all the effect size indices (except RR and NNT), the most correct answers were noted for small effect sizes (online supplemental file 4). When presented with large or medium effect sizes, the participants tended to underestimate the magnitude of the effect (ie, interpret them as representing smaller effects). Only in the case of the RR, larger effects were better interpreted, indicating that small or medium effects were misinterpreted as larger effects.

Table 2 Characteristics of all participants that fully completed the survey

Aspects	Number of respondents (%)
Sex	
Male	459 (60.2)
Female	299 (39.2)
Diverse	4 (0.5)
Country	
Germany	383 (50.3)
Japan	161 (21.1)
Spain	53 (7.0)
USA	41 (5.4)
United Kingdom	39 (5.1)
Canada	16 (2.1)
Italy	14 (1.8)
France	13 (1.7)
Austria	13 (1.7)
Australia	12 (1.6)
Others	17 (2.2)
Specialty	
Psychiatry	177 (23.2)
Internal medicine (including subspecialities)	82 (10.8)
General medicine/family medicine	56 (7.3)
Others	447 (58.7)
Professional status	
Student	75 (9.8)
Resident	116 (15.2)
Attending/staff physician	196 (25.7)
Consultant	132 (17.3)
Chief physician	40 (5.2)
Dentist	44 (5.8)
Psychologist	44 (5.8)
Other (eg, pharmacist, nurse, midwife, public health scientist, physical therapist etc.)	115 (15.1)
Graduation from university	
Not graduated yet	70 (9.2)
Graduated 2010 or later	287 (37.7)
Graduated between 2000 and 2009	225 (29.5)
Graduated between 1990 and 1999	106 (13.9)
Graduated 1989 or earlier	74 (9.7)
Experience/knowledge in health research methodology	
No official prior knowledge	444 (58.3)
Knowledge based on one or more formal courses in health research methodology	197 (25.9)
Knowledge based on a masters or PhD degree in health practice	121 (15.9)
Experience in conducting systematic reviews	
Has conducted a systematic review with meta-analysis himself	124 (16.3)
Has not conducted a systematic review with meta-analysis himself	638 (83.7)

'Others' in 'specialty' includes accident and emergency medicine; anaesthesiology; dentistry; dermatology; gynaecology; laboratory medicine; microbiology, virology; neurology; ophthalmology; orthopaedics; otorhinolaryngology; pathology; paediatrics; pharmacology; public health; physical and rehabilitative medicine; psychology; psychosomatic medicine; radiology, nuclear medicine, radiotherapy; surgery (including subspecialities); urology; student; other.

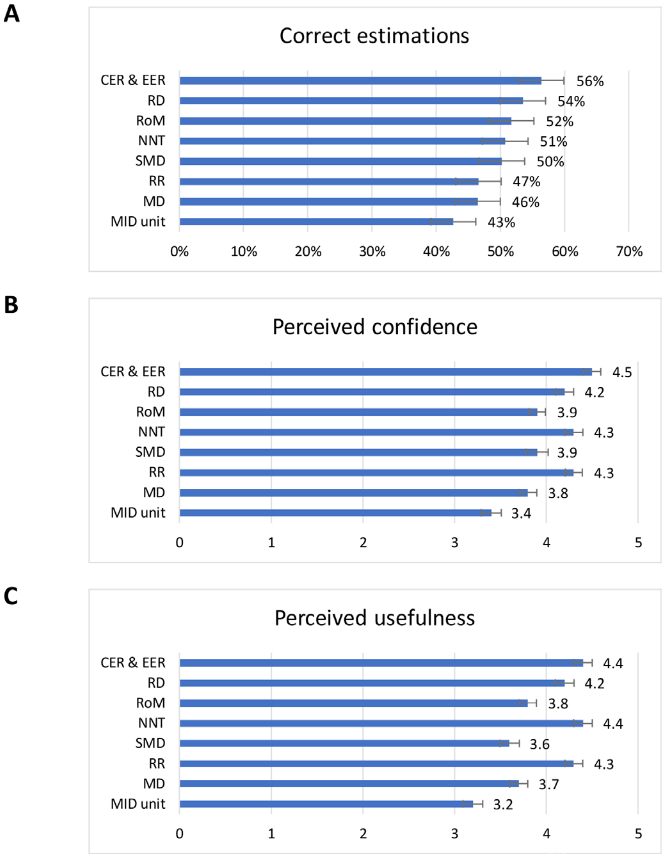

Figure 1 Proportion of correct answers, perceived confidence and perceived usefulness description. (A) Proportion of correct answers regarding the estimation of the size of treatment effects presented by the eight effect size measures. (B) Participants' perceived confidence (on a scale between 1 and 7) while dealing with the effect size measures. A higher value stands for higher confidence. (C) Participants' perceived usefulness (on a scale between 1 and 7) for the effect size measures. A higher value stands for higher perceived usefulness. Error bars=95% CI. CER & EER, control event rate and experimental event rate; MD, mean difference; MID unit, difference in minimal importance difference units; NNT, number needed to treat; RD, risk difference; RoM, ratio of means; RR, risk ratio; SMD, standardised mean difference.

DISCUSSION

The CER and EER as method of presentation was understood best followed by the RD. The lowest rate of correct answers was seen for the MID unit, followed by the MD. These results were generally in line with the respondents' indications regarding their perceived confidence and usefulness. The NNT, SMD and RoM ranked in-between. Experience in health research methodology showed a positive impact on the rate of correct answers.

The performance of the RR, arguably the preferred summary index in many meta-analyses for the dichotomous outcomes,[15] was peculiar. Although participants clearly understood CER and EER or RD more than RR, respondents indicated that they had greater confidence and perceived usefulness for RR. As perhaps expected, the misinterpretation of the RR lay in the direction of over-interpreting small effects, a common mistake by consumers of the evidence especially when presented with the relative risk reduction. The RR was also less correctly understood than the RD in Johnston's study.[10] These findings clearly suggest that the RR should not be the sole summary index to communicate the effect of an intervention. A recent survey found that the RR was the only reported data presentation method in abstracts of many RCTs in leading journals,[16] a practice that will likely mislead evidence users and that needs improvement. The NNT

Table 3 Results of the logistic regression analysis

	OR	95% CI	P value
Effect size indices			
CER and EER	(Ref)		
RD	0.91	0.73 to 1.11	0.332
RoM	0.83	0.67 to 1.02	0.070
NNT	0.80	0.65 to 0.98	0.033
SMD	0.78	0.63 to 0.96	0.018
RR	0.66	0.54 to 0.82	<0.001
MD	0.66	0.54 to 0.82	<0.001
MID unit	0.56	0.46 to 0.69	<0.001
Magnitude of effect size			
Small effect size	(Ref)		
Medium effect size	0.56	0.49 to 0.64	<0.001
Large effect size	0.51	0.45 to 0.58	<0.001
Experience/knowledge in health research practice (vs none)	1.13	1.00 to 1.28	0.048
Mental health professions (vs others)	0.96	0.85 to 1.08	0.487
Experience in conducting systematic reviews (vs none)	1.00	0.85 to 1.18	0.983

CER and EER, control event rate and experimental event rate; MD, mean difference; MID unit, difference in minimal importance difference units; NNT, number needed to treat; RD, risk difference; RoM, ratio of means; RR, risk ratio; SMD, standardised mean difference.

is sometimes advocated as the preferred way to make the RR more clinically interpretable,[7] however, given the methods for calculating the 95% CI can be confusing (eg, when the 95% CI of the RD is (−0.2 to 0.2), the correct NNT should be (−∞ to −5, 5 to ∞) but is often misunderstood as (−5 to 5)), it has been suggested that EER, CER and RD are better options,[17] as our findings bear out.

With regard to continuous outcomes, the MD is often defended as the more easily interpretable than the SMD, particularly if the instrument, in our case pain intensity on a 10 point scale, is familiar to the audience.[18] This was not the case in our survey. In Johnston's study, the MD was also one of the two least correctly interpreted indices (along with the MID unit).[10] The MD has been shown to be slightly less generalisable than the SMD.[19] It is possible that, if the MD represents *natural* units such as weight or a laboratory value instead of a 0–10 pain score as in our survey, it may be interpreted more correctly than SMD. Moreover, we must remember that in the current survey experience in health research methodology influenced the interpretability. The MD probably will continue to be perceived as a readily understood index of effect size when the results are presented to the lay public or to the less methodologically trained health professionals. However, we must keep in mind that, behind this apparent ease of understanding, perhaps their interpretation may remain misleading, especially when the unit is not familiar, for example, scores of a certain psychopathology scale. Once the evidence users became more experienced, interpretations based on the SMD were more correct than those based on the MD, and perceived to be equally helpful with equal confidence.

The index based on MID units was the least correctly interpreted and perceived to be the least helpful. In Johnston's study, it was one of the least correctly interpreted indices and the second least useful index. First of all, this was probably driven by the unfamiliarity of the current medical professionals with the concept of the MID, even though it has been around for three decades.[20] Second, it remains possible that using the MID, which represents the smallest important pre–post change, in the context of the between-group comparison may have been conceptually misguided.[21 22] The place and value of the MID unit approach to express the effect size need further research and educational outreach. By contrast, another newly proposed method to summarise a continuous outcome, the RoM, was as well understood as the SMD. Unfortunately, the use of the RoM is limited by the fact that it can be calculated only when the scores in the intervention and the control group are both, positive or negative.[4] Meeting this condition, the RoM remains a viable option as an effect size index.

It must be noted that even the best-performing indices led to correct interpretations in slightly more than half of the questions only, and the perceived confidence and usefulness hovered around the middle value on a scale of 1 to 7. This performance must be interpreted in the context of our questionnaire design in which, for the estimation of the presented effect sizes, participants only had to choose among three possible answers (small effect, medium effect, large effect). This design would have naturally increased the rate of right guesses, because, by chance alone, every third answer should be correct. Furthermore, a small effect could not be underestimated while a large effect could not be overestimated. However, the characterisation of various effects is bound to be subjective and achievement of perfect or near-perfect correct answers may not be pragmatically possible or to be expected. The fact that participants tended to underestimate large or medium effect sizes could also indicate that effects in pain interventions are mostly modest. Furthermore, it implies that what we define as a large effect in our survey may, in reality, only yield a limited impact and might not be considered as a large effect.

Limitations

The study faces limitations typical of survey studies, with voluntary participation potentially under-representing those uninterested in evidence-based medicine.[23] Also, people who do not feel confident with the English language might not have taken part in our study. Another reason for non-participation in our survey could be that individuals found the questionnaire too time-consuming during their daily work routine. For others, the number of questions may have caused response fatigue, prompting them to cease completing the questionnaire. Our participants, while diverse across 13 countries, were predominantly from high-income nations. The absence of respondents from middle-income to low-income country, where training in health research methodology is likely to be less common, limits generalisability. We did not include the OR in our survey because it is already known to be difficult to understand and can be easily misinterpreted as compared with the RR.[24] To reduce the respondents' time burden and response fatigue to avoid non-response, we decided to focus on the RR as a relative index of efficacy. Nevertheless, it would have been interesting to examine OR in our questionnaire. The study's definitions of small, medium and large effects could be seen as arbitrary. We concede this caveat but argue that no other alternative could have been any more plausible. Following Johnston *et al,* we examined both perceived confidence and usefulness but these concepts probably overlap. In online supplemental file 3, we present a scatterplot, which indeed demonstrates that these measures very likely are related.

CONCLUSION AND CLINICAL IMPLICATIONS

Our findings suggest that studies presenting the results using the CER and EER would lead to the most correct interpretation of the effect size. It was also associated with the highest perceived confidence and usefulness among various healthcare and related workers. However, all the tested effect size indices were only moderately correctly interpreted, with only a 13% difference in correct answers between the best (CER and EER: 56%) and worst performing index (MID unit: 43%). While relative measures including the RR (and the OR) remain the most externally valid summary index,[15] they should be supplemented with absolute measures such as the CER and EER. The current study provides strong empirical support for the way the Summary of Findings tables are structured in the Cochrane Library.[25] Especially in the field of psychiatry, the presentation of both CER and EER is particularly crucial, given the often high placebo response rates. For instance, in the acute treatment of schizophrenia with antipsychotics, a meta-analysis revealed that approximately 30% of patients exhibited at least minimal improvement after about 6 weeks under placebo, while 50% demonstrated such improvement under drug treatment.[1] In the case of antidepressants for major depression, approximately 37% respond to placebo,[26] compared with 52% responding to antidepressants.[27] When the outcome is continuous such as pain intensity, our results indicate the use of the SMD over the MD, which is less correctly interpreted, and which is less externally valid. The interpretability of the SMD must apparently be cultivated with education in health research methodology among the professionals, while we must be aware of the faux interpretability of the MD for the lay public when conveying the results of continuous outcomes. The SMD can also be converted into the CER and EER using the validated conversion method[28 29]: supplementing the summary SMD with the converted CER and EER may be as helpful as supplementing the RR with the same. Since, in our survey, even the best-performing indices led to correct interpretations in only slightly more than half of the questions, further initiatives are needed to improve the education of health professionals in health research methodology, including skills in interpreting effect size.

Author affiliations
[1]Department of Psychiatry and Psychotherapy, School of Medicine and Health, Technical University of Munich, Munich, Bavaria, Germany
[2]Department of Neuropsychiatry, University of Tokyo Hospital, Tokyo, Japan
[3]German Center for Mental Health (DZPG), partner site München/Augsburg, Munich, Germany
[4]Department of Nutrition, College of Agriculture and Life Sciences, Texas A&M University, College Station, Texas, USA
[5]Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, College Station, Texas, USA
[6]Department of Psychiatry and Psychotherapy, Klinikum der Ludwig-Maximilians-Universität München, Munich, Bavaria, Germany
[7]Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

Twitter Toshi A Furukawa @Toshi_FRKW

Acknowledgements This project will be part of Ferdinand Heimke's doctoral thesis.

Contributors All authors have contributed significantly and agree with the content of the manuscript. FH, YF, SS, BJ, RE, TAF and SL conceived this study. FH, YF, TAF and SL organised the distribution of the questionnaire and coordinated data collection. FH and RE conducted the statistical analysis. FH, TAF, BJ and SL drafted and prepared the manuscript. All authors took part in the interpretation of results, revised the manuscript for important intellectual content and approved the final version. SL acts as guarantor for the overall content.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests In the last 3 years, SL has received honoraria as a consultant and/or advisor and/or for lectures and/or for educational material from Alkermes, Angelini, Eisai, Gedeon Richter, Janssen, Lundbeck, Medichem, Medscape, Merck Sharpp and Dome, Mitshubishi, Neurotorium, NovoNordisk, Otsuka, Recordati, Roche, Rovi, Sanofi Aventis, TEVA. TAF reports personal fees from Boehringer-Ingelheim, DT Axis, Kyoto University Original, Shionogi and SONY and UpToDate, and a grant from Shionogi, outside the submitted work; In addition, TAF has patents 2020-548587 and 2022-082495 pending, and intellectual properties for Kokoro-app licensed to Mitsubishi-Tanabe.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Ethics Committee of the Technical University of Munich; 682/21 S-SR. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer-reviewed.

Data availability statement Data are available upon reasonable request. The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Ferdinand Heimke http://orcid.org/0009-0001-7042-4007
Spyridon Siafis http://orcid.org/0000-0001-8264-2039
Toshi A Furukawa http://orcid.org/0000-0003-2159-3776
Stefan Leucht http://orcid.org/0000-0002-4934-4352

REFERENCES

1 Leucht S, Siafis S, Engel RR, *et al*. How efficacious are antipsychotic drugs for schizophrenia? An interpretation based on 13 effect size indices. *Schizophr Bull* 2022;48:27–36.
2 Guyatt GH, Thorlund K, Oxman AD, *et al*. GRADE guidelines: 13. preparing summary of findings tables and evidence profiles—continuous outcomes. *J Clin Epidemiol* 2013;66:173–83.
3 Valentine JC, Aloe AM. How to communicate effect sizes for continuous outcomes: a review of existing options and introducing a new metric. *J Clin Epidemiol* 2016;72:84–9.
4 Higgins JPT, Li T, Deeks JJ. Chapter 6: choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane handbook for systematic reviews of interventions version 6.3 (updated February 2022)*. Cochrane, 2022. Available: www.training.cochrane.org/handbook
5 Friedrich JO, Adhikari NKJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol* 2011;64:556–64.
6 Johnston BC, Thorlund K, Schünemann HJ, *et al*. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. *Health Qual Life Outcomes* 2010;8:116.
7 Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;310:452–4.
8 Guyatt GH, Oxman AD, Santesso N, *et al*. GRADE guidelines: 12. preparing summary of findings tables—binary outcomes. *J Clin Epidemiol* 2013;66:158–72.
9 Akl EA, Oxman AD, Herrin J, *et al*. Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database Syst Rev* 2011;2011:CD006776.
10 Johnston BC, Alonso-Coello P, Friedrich JO, *et al*. Do clinicians understand the size of treatment effects? A randomized survey across 8 countries. *CMAJ* 2016;188:25–32.
11 Sharma A, Minh Duc NT, Luu Lam Thang T, *et al*. A consensus-based checklist for reporting of survey studies (CROSS). *J Gen Intern Med* 2021;36:3179–87.
12 Lerman SF, Rudich Z, Brill S, *et al*. Longitudinal associations between depression, anxiety, pain, and pain-related disability in chronic pain patients. *Psychosom Med* 2015;77:333–41.
13 Busse JW, Wang L, Kamaleldin M, *et al*. Opioids for chronic noncancer pain: a systematic review and meta-analysis. *JAMA* 2018;320:2448–60.
14 Cohen J. *Statistical power analysis for the behavioral sciences*. 2. ed. Hillsdale, NJ: Erlbaum, 1988.
15 Furukawa TA, Guyatt GH, Griffith LE. "Can we Individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses". *Int J Epidemiol* 2002;31:72–6.

16 Seta T, Takahashi Y, Yamashita Y, *et al*. Outcome measures reported in abstracts of randomized controlled trials in leading clinical journals: a Bibliometric study. *J Gen Fam Med* 2020;21:119–26.

17 Hutton JL. Number needed to treat and number needed to harm are not the best way to report and assess the results of randomised clinical trials. *Br J Haematol* 2009;146:27–30.

18 Schünemann HJ, Vist GE, Higgins JPT, *et al*. Chapter 15: interpreting results and drawing conclusions. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane handbook for systematic reviews of interventions version 6.3*. Cochrane, 2022. Available: www.training.cochrane.org/handbook

19 Takeshima N, Sozu T, Tajika A, *et al*. Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference *BMC Med Res Methodol* 2014;14:30.

20 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407–15.

21 Furukawa TA. Measuring clinical importance in a trial of interventions for mixed urinary Incontinence. *JAMA* 2020;323:479.

22 Furukawa TA, Scott IA, Guyatt G. Measuring patients' experience. In: Guyatt G, Rennie D, Meade MO, et al., eds. *Users' Guides to the Medical literature: a manual for evidence-based clinical practice*. 3rd ed. New York, NY: McGraw-Hill Education, 2014: 219–34.

23 Singh S, Sagar R. A critical look at online survey or questionnaire-based research studies during COVID-19. *Asian J Psychiatr* 2021;65:102850.

24 Persoskie A, Ferrer RA. A most odd ratio:: interpreting and describing odds ratios. *Am J Prev Med* 2017;52:224–8.

25 Schünemann HJ, Higgins JPT, Vist GE, *et al*. Chapter 14: completing 'summary of findings' tables and grading the certainty of the evidence. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane handbook for systematic reviews of interventions version 6.3*. Cochrane, 2022. Available: www.training.cochrane.org/handbook

26 Furukawa TA, Cipriani A, Atkinson LZ, *et al*. Placebo response rates in antidepressant trials: a systematic review of published and unpublished double-blind randomised controlled studies. *Lancet Psychiatry* 2016;3:1059–66.

27 Stone MB, Yaseen ZS, Miller BJ, *et al*. Response to acute monotherapy for major depressive disorder in randomized, placebo controlled trials submitted to the US Food and Drug Administration: individual participant data analysis. *BMJ* 2022;378:e067606.

28 Furukawa TA. From effect size into number needed to treat. *Lancet* 1999;353:.:9165.

29 da Costa BR, Rutjes AWS, Johnston BC, *et al*. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. *Int J Epidemiol* 2012;41:1445–59.