REVIEW AND DISCUSSION PAPERS

Determining clinically significant patient change with effect sizes: considerations for clinicians and researchers

AUTHORS

DUNCAN MCKECHNIE BNurs(Hons), DipPublicSafety, GradCertRehabNurs, PhD, RN[1,2]

1 Clinical Nurse Consultant, Royal Rehab, Sydney, NSW, Australia

2 Casual Lecturer, Susan Wakil School of Nursing and Midwifery, University of Sydney, NSW

CORRESPONDING AUTHOR

DUNCAN MCKECHNIE Royal Rehab, PO Box 6, Ryde, Sydney, NSW 1680 Australia.
E: duncan.mckechnie@royalrehab.com.au

ABSTRACT

Objective: To aid nurses' understanding of effect size utilisation in clinical and research contexts.

Design and data sources: Methodological discussion paper that is based on the author's experiences as a clinician and researcher and is supported by literature.

Primary arguments: Patient change is a key consideration for clinical nurses and nurse researchers. Nurses routinely use measurement instruments to identify and quantify such change informing intervention outcomes, clinical decision-making, and health research conclusions. Whether improvement or deterioration, patient change should be operationalised through the magnitude of change (i.e., effect size). Effect sizes relative to the context of change (clinical vs empirical) and the reliability of the instruments used are important considerations here. However, despite discourse on the utilisation of effect sizes in health, aspects of effect sizes can be poorly understood, misapplied or overlooked. Furthermore, nurse researchers may default to Cohen's *d* for use in power analysis and results reporting where they should be considering an effect size derived from other methods in the first instance. In part, this is due to the literature surrounding aspects of effect size being inherently complex, impacting on nurse and nurse researchers' capacity to acquire a thorough understanding of the topic.

Conclusions: Effect size in health can be particularly complex. Nevertheless, nurses and nurse researchers should have some understanding about effect sizes and their role in measuring patient change in clinical and empirical contexts. They need to be aware of how measurement instruments detect, track and quantify patient change and the resultant magnitude of effect relative to the clinical significance of the change for the patient. This paper aids nurses to effect robust change based on informed decision making thus strengthening their evidence-based practice.

What is already known about the topic?

- Patient change informs nurses clinical decision-making strategies; however, nurses may not consider the magnitude of change relative to the context of change and the reliability of the instruments used to identify and quantify the change
- Effect size is one of the four criteria needed for power analysis and is perhaps the most difficult to identify
- Underpowered studies result in imprecise estimation of the true effect, which could be an over- or an under-estimation

What this paper adds:

- This paper dispenses with the inherently complicated and technical terminology on effect size often found in the literature that can impact understanding. Consequently, this paper equips nurses to critique research literature and apply this knowledge to their clinical practice
- Provides other methods for identifying a suitable effect size for use in power analysis and results reporting as opposed to defaulting to Cohen's *d*
- Draws attention to the importance of reporting effect size and associated confidence interval with research results data

Keywords: Effect size, measurement, instrument, clinically significant

INTRODUCTION

The measurement of patient change is important for clinical nurses and nurse researchers to understand. Patient change is characterised as a deviation from a patient's baseline in medical, physical, behavioural, cognitive, functional, capacity and/or mental health domains. A change in a patient's condition is complex as it could be benign or clinically important, subtle or overt, sudden or gradual and positive or negative in nature. Regardless of the nature, however, nurses consider patient change with each encounter to inform clinical decision making. Indeed, a key goal for clinicians is to identify, treat, and modify care interventions based on patient change. Clinical researchers may use patient change, such as comparing two antihypertensives in a randomised controlled trial (RCT), as evidence regarding the efficacy of an intervention. Identifying reliable measurement instruments to predict, identify, and quantify patient change is important for both clinicians and researchers. Effect and effect size (ES) have a central role in these considerations.

In health patient change is often described as an effect. An effect in this context relates to patient change due to an action, intervention, or other cause. Nurses generally observe effects from two perspectives: functional (e.g., changes to capacity, mobility, or continence) and medical (e.g., changes to temperature or blood pressure) contexts. Clinical researchers use statistics to identify an effect when comparing outcomes in two populations (e.g., between two treatments in an RCT), treatment effects within the same group, or between low and high-risk groups. This information is dichotomous in nature informing whether an effect exists or not. On the other hand, ESs provides information about the magnitude, direction, and strength of an effect in relation to results as they occur and as such, are termed magnitude of effect.[1–3] For this reason, effect and ES are important concepts for nurses and nurse researchers to understand. The purpose of this paper is to aid the understanding of ES utilisation in determining clinically significant patient change by nurses and beginner nurse researchers whose knowledge on these concepts may be limited.

BACKGROUND

This paper is part of a series of articles about methodological aspects of health research. The overarching aim of this series is to assist nurses and beginning nurse researchers to critique research literature and conduct research that informs evidence-based practice. In this paper aspects of ES in measuring patient change are discussed. There are numerous methodological papers on ES, so from this perspective this paper is not new. What is new, however, is the aim to do without the inherently complicated, dense, and technical discourse that is often found in the literature regarding the subject that can impact the ability of beginning researchers and nurses to understand and apply the information. This paper provides a straightforward perspective on some long-standing ES concepts.

Effect sizes have a role in many aspects of clinical health and research. Quantifying patient change, power analysis, establishing the responsiveness and minimal detectable change (MDC) of health-related measurement instruments and minimal clinically important difference (MCID) are all relevant here. Researchers need to translate their results to some quantifiable meaning, such as an ES, and then provide a qualitative explanation of the effect regarding clinical significance from which clinicians can then apply to their practice. These aspects of ES have important implications for researchers, patients, and nurses. This paper begins by providing some background context to these concepts in relation to ES.

THE IMPORTANCE OF ES IN POWER ANALYSIS

A study's power (aka statistical power) is the probability of detecting a true effect when it exists. Establishing a study's power often involves a priori (before the study) power analysis. This type of power analysis is a process for determining sample size or number of observations needed to avoid a Type II error (false-negative), given a desired significance level, statistical power and population ES.[4,5] Historically a notable proportion of published research has been underpowered.[6–8] This is a concern as statistically significant results in underpowered studies can reflect an

 https://doi.org/10.37464/2025.422.1949

imprecise estimate of the true effect of an intervention as 'underpowered studies have to detect much larger effects to achieve statistical significance.'[9,1(p.125)] Overall, in underpowered studies true and smaller effects can be missed, there is an increased risk of false positive statistically significant results and ESs of statistically significant results can be exaggerated to appear larger than they actually are resulting in little useful information about the effect size.[6,8,10] This is why identifying a suitable population ES for inclusion in power analysis is a primary consideration.

STATISTICAL SIGNIFICANCE VS CLINICAL SIGNIFICANCE

Statistical significance and clinical significance are not the same thing, and the relationship between them is inherently complex. The former should be thought of as a necessary condition but not sufficient for judging a treatment to be effective.[11] However, a pervasive myth in clinical research is that the smaller the *p*-value (i.e., statistical significance) the stronger the hypothesis that an effect, relationship or association exists.[12] This is due to the *p*-value only informing the likelihood of the results occurring by random chance and consequently, it doesn't tell you if the null hypothesis is true or false. Another consideration here is that 'a sufficiently powerful test will almost always generate a statistically significant result irrespective of the effect size'.[1(p.16)] That is, with large samples extremely small effects can result in statistically significant results even when there is little to no clinical significance.[13]

Statistically significant results in health research are commonly interpreted as important and meaningful patient change. This is not entirely accurate. When researchers and clinicians consider a statistically significant result from a statistical test, the utility of the effect in terms of clinical significance is perhaps more important to consider. This is because a statistically significant result only informs whether an effect exists which may not be synonymous with any clinical or practical significance for the patient as it does not convey the magnitude of the effect. Consequently, the reliance on '*p*-values as a basis for evidence-based clinical decision-making is a major source of error' and should not be used as the sole inference for clinical significance.[14,12(p.302)]

Clinical significance goes beyond statistical significance as it identifies whether the statistically significant difference, or score on a measurement instrument, is large enough to have clinical implications for the patient.[15,16] This is where the utility of a statistically significant finding in terms of the associated ES and confidence interval (CI) needs to be considered. However, this is not so straightforward as some 'commonly used effect sizes are limited in conveying clinical significance' as they have limited interpretability as an ES misleading clinical decision-making, for example, odds ratio.[12,14(p.990),17] It is recommended that ESs number needed to treat (NNT), success rate difference (SRD) and if relevant area under the receiver operating characteristic curve (ROC) be reported to convey clinical significance when comparing two populations.[12,14] Reporting such ES with their CIs allows consumers of research to better judge the clinical significance of research results as they apply to their own contexts and standards.

While ESs are used to report research results when sampling a population of patients, a more relevant issue for clinical nurses is how to measure, quantify, and track individual patient change. Further, even if a statistically significant result is clinically significant and can be generalised to the population of interest, it may have little importance to an individual patient. There are simply too many research confounders, individual patient factors, and contextual factors to account for. This is where the use of measurement instruments and identifying MCID to quantity patient change is beneficial.[15,18]

MCID AND MDC

Minimal clinically important difference (MCID) and minimal detectable change (MDC) have a role in determining whether patient change is clinically significant. MCID (aka minimally important difference) is fundamentally an outcome ES derived from health-related measurement instruments. MCID can be used as a reference point for identifying the magnitude of treatment effects based on patient change.[19] Jaeschke et al provides a self-explanatory definition of MCID as being 'the smallest difference in score in the domain of interest [e.g., outcome measure or scale] which patients perceive as beneficial which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient's management.'[20(p.408)] Guyatt et al recommended adding 'or harmful' to the definition to address patient deterioration.[21]

Several factors make the concept of MCID useful in health. First, MCID can be used 'for judging the magnitude of treatment effects [i.e., clinical significance] not only in routine clinical practice but also in clinical trials and systematic reviews, facilitating the establishment of treatment recommendations for patients.'[22(p.2)] Second, MCID is an ES that can be used for sample size estimates regarding the desired MCIDs one wishes to detect. Third, MCID 'emphasizes the primacy of the patient's perspective and implicitly links that perspective to that of the physician.'[21(p.377)] Finally, the MCID construct is easily understood by clinicians as they routinely use the instruments used to determine MCIDs (e.g., Visual Analog Scale and Functional Independence Measure) and are knowledgeable of the patients' presenting condition and associated deviation from baseline.

The MDC criterion is tied to clinical significance and MCID. This is mainly due to the difficulty in operationalising MCID without a minimum reference point which MDC provides.

MDC reflects a threshold for minimum point change of an outcome measurement instrument or scale; this relates to its ability to detect actual patient change beyond measurement error within a defined level of statistical confidence (e.g., 95% CI).[15,23] Consequently, an instrument's standard error of measurement (SEM) needs to be determined to identify MDC. There is a relationship between SEM, MDC, ES and MCID characterised by: (i) the higher the instrument's reliability the lower the ES needed to achieve an MCID;[15] and (ii) MDC needs to be smaller than an MCID to ensure that the change score is beyond measurement error (i.e., SEM).[23]

RESPONSIVENESS OF HEALTH-RELATED MEASUREMENT INSTRUMENTS

A fundamental role of health-related measurement instruments is to identify patient change, whether improvement or deterioration. Responsiveness of instruments is one of their psychometric properties warranting consideration here. The responsiveness of a measurement instrument relates to its ability to accurately *detect* and *track* clinically meaningful patient change. This primarily relates to an instrument's change score which is obtained by the arithmetic differences between serially gathered data, such as before and after treatment or comparing a control group and an intervention group.[24] Responsiveness can be further divided into internal and external responsiveness. The former relates to an instrument's precision in tracking patient change over time or change before and after an intervention which can be defined as MDC. With external responsiveness, a reference instrument is compared to an external criterion, index, or measure from which MCID can be determined.[16,25,26]

AIM

To aid nurses' understanding of effect size utilisation in clinical and research contexts.

DESIGN AND DATA SOURCES

A methodological discussion paper that is based on the author's experiences as a clinician and researcher and is supported by literature.

DISCUSSION

ES TYPES AND CATEGORIES

There are many different types of ES that are generally based on how they are derived and from which data source. For instance, researchers may need to identify a *population* ES for priori power analysis while research results may be used to compute a *sample* ES. ES can further be divided into *absolute* (raw), such as the difference between cohort means, and *relative* (standardised) ESs. Any indices (e.g., squared correlations and kappa) that convey the magnitude of change are considered a relative ES.[27]

Relative ES can be categorised as the difference between groups or measures of association known as the *d* and *r* family, respectively.[1,28] In the *d* family ES includes comparisons between binary variables (e.g., yes/no data) that can be expressed as relative risk or SRD. These indices represent the difference between two proportions classified as the probability of being in one of the two categories, such as in the Chi-square test.[3] In this family, comparisons between a continuous variable (e.g., height and weight) means and their associated standard deviations (SDs) are used to calculate standardised differences expressing ES in SD units.[29] Cohen's *d* is an example here. SDs on their own can also be considered as a discrete ES statistic as they represent the variation of each group around the mean.[1,30]

The *r* family of ES covers the direction and strength of a relationship between two or more binary or continuous variables.[1] Some examples include ANOVA (*f*), Pearson product moment correlation coefficient (*r*) and Spearman's rank correlation coefficient (p or r_s). Proportion of variance indexes are also part of this family including coefficient of determination (r^2) and multiple regression (R^2).

In addition to statistical test and variable type impacting on which ES should be considered, some ES are only valid when statistical assumptions are met.[31] For instance, in comparing two treatments or interventions in an RCT the ES may be expressed as a hazard ratio (HR), which is only valid if two survival curves are being compared that satisfy the proportional hazards assumption in that population.[12,32] The validity of Cohen's *d*, Hedges' *g* or Glass's delta, depend on the outcome measures in the two populations having a normal distribution with equal variances.[12,31,33] Another consideration if assumptions are met, is that some ES can be converted to others. For example, conversions between Cohen's *d*, HR, NNT and SRD are possible and assist with clinical interpretability and determining the clinical significance of results.[12]

DETERMINING ES

There are many types of ES and methods for determining an ES. Based on their context of use, ESs are generally derived from two methods: distribution-based and anchor-based.

Distribution-based approaches for determining ES

The distribution-based method for determining relative ES involves the underlying distribution and magnitude of change measured in SD units around the mean.[22] ES can be expressed in three ways using this method:

(i) between-person SD units (person 1 mean minus person 2 mean);

(ii) within-person SD units (post-test mean minus pre-test mean); and

(iii) the standard error of measurement (SEM).[21]

There are numerous methods for calculating relative ES within the distribution-based category. Perhaps the most widely used and reported method involves comparing the means of continuous variables (see Table 1). While there are online calculators for determining some of the ESs in Table 1, the formulas are provided as clinicians generally only need means, SDs and cohort numbers to calculate their own ES if not reported aiding the interpretation of results. It is important to note that when using SDs as an ES and for research in general, the population SD and sample SD are calculated differently and are represented by different indices (= population SD, s = sample SD). There are also two different mean indices (= population mean, = sample mean).

Population ES is one of the four criteria needed for power analysis in quantitative studies and is perhaps the most difficult to identify. Being derived from distribution-based methods, an ES is required for all types of power analysis except sensitivity power analysis where the goal is to identify an ES based on a known sample size. Regarding the former types of power analysis, researchers in the first instances should always attempt to identify a population ES worth investigating as they would be applied to the relative clinical or empirical context.[10] This could be based on previous similar studies, a systematic review, expert clinical judgment or informed clinical opinion. Study design and methods and types of variables, statistical tests and measurement instruments used are also considerations here.[4]

TABLE 1: FORMULAS FOR DETERMINING DISTRIBUTION-BASED ES

Relative effect size	Characteristics	Formula	Considerations
Cohen's *d*[34]	Either group SD if they are homogeneous	$d = \frac{m_1 - m_2}{s}$	
Cohen's *d*[34]	Pooled SD_{pooled} if SDs are about the same	$d = \frac{m_1 - m_2}{\sqrt{\left[\frac{(s_1{}^2 + s_2{}^2)}{2}\right]}}$	Can overestimate the true population ES
Glass's delta (Δ)[35]	Control group SD if the SD of each group are sufficiently different	$\Delta = \frac{m_1 - m_2}{s_{control}}$	Also referred to as relative change[16]
Hedge's *g*[35]	Weighted & pooled SD if group sizes are different	$g = \frac{m_1 - m_2}{\sqrt{\frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{(n_1 + n_2 - 2)}}}$	Suitable for sample sizes <20 & unbiased estimates of the population ES opposed to Cohen's *d*
Kazis formula[11]	Pre-intervention SD used as a proxy for control group SD	$ES = \frac{m_1 - m_2}{s_{preintervention\ group}}$	Can be used for within-person and between-person. Same as Glass's Δ when the assumed control group is the pre-intervention group.
SRM[25,36]	Mainly used to determine internal responsiveness	$SRM = \frac{m_1 - m_2}{s_{change\ score}}$	Also termed responsiveness to treatment coefficient or efficiency index
SEM	For calculating MDC	$SEM = SD \times \sqrt{1 - r}$	SD = from total sample at baseline[23] or pooled initial and re-test SDs. r = reliability coefficient of the reference tool being test-retest[22] including ICC[23] or internal consistency (Cronbach's α)[15,36]
MDC[37]	Identifying MDC of a measure	$MDC_{CI} = SEM \times z \times \sqrt{2}$	z-values depend on the desired CI. E.g., 1.64 for 90% CI and 1.96 for 95% CI.
GRI[26]	Mainly used to determine MDC and internal responsiveness	$GRI = \frac{\Delta}{\sqrt{2 \times MSE}}$	Δ = mean change of treatment group.[23] MSE = ANOVA for multiple baseline measures prior to intervention or SD of reference group for two observations (before and after intervention)[25]
Responsiveness statistic[26]	Mainly used to establish responsiveness	$ES = \frac{m_1 - m_2}{s_{stable\ group}}$	
Relative change[11]	Quantitative descriptor of patient change[16]	$RC = \frac{m_1 - m_2}{m_1}$	
Norman index[18]	Mainly serves as a starting baseline for estimating MCID	$ES = 0.5 \times S_{\text{preintervention}}$	Control group SD can be used. Manly used for patient-reported outcome measures.
RCI[38]	Mainly used with MCID estimates to ascertain if the score change (e.g. before and after an intervention) is statistically significant[39]	$RCI_{CI} = \frac{m_2 - m_1}{\sqrt{2 \times (s_1\sqrt{1 - r})^2}} \times z$	Formula in brackets is SEM using SD of pre-intervention group. E.g., if z = 1.96 and RCI >1.96 then there is a statistically significant change based on 95% CI.

Abbreviations: *m*, mean; *s*, standard deviation; *n*, number in group; MDC, minimal detectable change; MCID, minimal clinically important difference; SEM, standard error of measurement; ICC, interclass correlation coefficients; SD, standard deviation; CI, confidence interval; SRM, standardised response mean; GRI, Guyatt Responsiveness Index; RCI, Reliable Change Index

TABLE 2: PROPOSED ES THRESHOLDS FOR COMMON STATISTICAL TESTS

Description	Example of statistical test	Indices	Proposed effect sizes		
			Small	Medium	Large
***d* ES family for mean differences**					
Independent means of continuous variables	Student's *t* test	*d*, Δ, *g*	.20	.50	.80
***r* ES family for correlation indexes**					
Binary variables	Chi-square test	ω, φ, *V*, *C*	.10	.30	.50
Two interval or ratio scale variables	Pearson coefficient	*r*			
Comparison of two correlations	Fisher's *r* to *z*	*q*			
Average Spearman Rho	Friedman test	*p* (r_s)			
***r* ES family for proportion of variance indexes**					
Difference between proportions	Sign Test	Cohen's *g*	.05	.15	.25
For independent proportions	*z*-test	*h*	.20	.50	.80
Mean dispersion in multiple groups	ANOVA	*f*	.10	.25	.40
	Eta/Omega[2]	η^2, Ω^2	.01	.06	.14
Multiple regression • Multiple & hierarchical regression • Bivariate regression		R^2 f^2 r^2	.02 .02 .01	.13 .15 .09	.26 .35 .25
Other					
Group mean differences	Student's *t*-test	*d*, Δ, *g*	.41	1.15	2.70
Relative risk (risk ratio)	Chi-square test	*RR*	2	3	4
Correlation indexes (range –1 to 1)	Pearson and Spearman's coefficient	*r*, *R*, ρ, β, *tau*, φ	± .2	± .5	± .8
Proportion of variance indexes (range 0 to 1)	Regression modelling	r^2, R^2, η^2, ε^2, ω^2	.04	.25	.64

Note: Adapted from Cohen[34], Ferguson[40], Ellse[1]

Motivated by the prevalence of underpowered studies,[15] Cohen developed three operational definitions to describe distribution-based ES that could be used when no better basis for identifying an ES is available.[34] These include:

1. *small* ES being noticeably smaller than medium but not so small that it is trivial, however, cannot be detected by the naked eye but detected by a statistical test;[2]
2. *medium* ES being an effect *likely* to be detectable by a careful or trained observer; and
3. *large* ES being an effect detectable by an untrained observer[2] represented by an effect that is as far above a medium effect as small is below it.[34]

Cohen further identified ES thresholds for several statistical tests based on these three definitions (see Table 2) which assists in operationalising desired (i.e., sample size estimates for power analysis) and clinically significant effects (i.e., interpreting results).[34] It is important to note that Cohen describes his definitions as arbitrary conventions and the associated ES thresholds as subjective judgements.[34] Consequently, they should serve as a guide only and not detract researchers from identifying relevant context-specific ES from the research literature based on empirical data and reasoned arguments.[28] Cohen's *d* can also be converted to other ESs (see Table 3). Formulas for these conversions are readily available in the literature.[3,34]

TABLE 3: CONVERSIONS BETWEEN COHEN'S *d* AND OTHER EFFECT SIZES

d effect size descriptors	*d*	*r*	r^2	SRD	NNT	HR[a]	HR[a]
	0	.000	.000	.00	∞	1.00	1.00
	.1	.050	.003	.06	17.7	.89	1.12
Small	.2	.100	.010	.11	8.9	.80	1.25
	.3	.148	.022	.17	6.0	.71	1.40
	.4	.196	.038	.22	4.5	.64	1.57
Medium	.5	.243	.059	.28	3.6	.57	1.76
	.6	.287	.082	.33	3.0	.51	1.98
	.7	.330	.109	.38	2.6	.45	2.22
Large	.8	.371	.138	.43	2.3	.40	2.50
	.9	.410	.168	.48	2.1	.36	2.81
	1.0	.447	.200	.52	1.9	.32	3.17
	2.0	.707	.500	.84	1.2	.09	11.71

Abbreviations: *d*, Cohens *d*; *r*, Pearson correlation coefficient; r^2, coefficient of determination based on Pearson correlation; SRD, success rate difference; NNT, number needed to treat; HR, hazed ratio.

a – which HR used depends on whether the event is undesirable (HR <1 if population 1 is better than population 2) or desirable (HR >1).

Note: Adapted from Cohen[34] and Kraemer et al[12(p.303)]

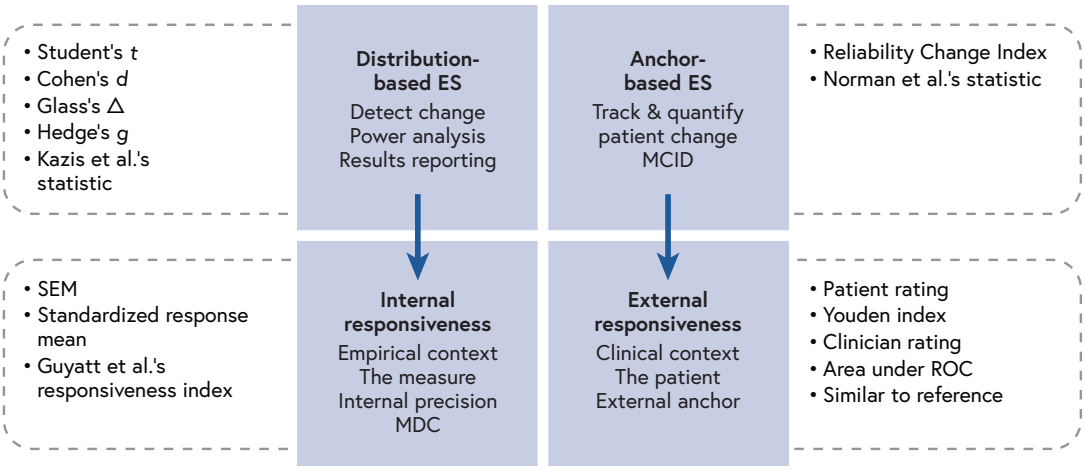

FIGURE 1: ES APPLICATION AND APPROACHES FOR DETERMINING ES AND RESPONSIVENESS OF A MEASURE

Distribution-based methods are also used to determine the internal responsiveness (aka internally referenced or precision) and MDC of measurement instruments. Standard error of measurement (SEM) of an instrument in identifying its MDC is an important aspect here. This is because MDC serves as an anchor of sorts. That is, only when an individual's change score exceeds the SEM can clinicians be confident that it is an actual change rather than a product of instrument measurement error.[24] While many of the distribution-based approaches for determining ES in Table 1 could be used to examine the precision of a measure,[24] the standardised response mean (SRM) and Guyatt et al., responsiveness index (GRI) are mainly used. Proposed ES benchmarks for SRM and GRI include 0.20-<0.50, 0.50-0.80, and >0.80 representing small, moderate and large responsiveness, respectively.[25,26,41] A summary of distribution-based approaches for determining ESs and the internal responsiveness of a measure are illustrated in Figure 1.

Anchor-based approaches for determining ES

While ESs derived from distribution-based methods have a key role in identifying population effects from an intervention in research and MDC, they have limited operational utility in guiding clinical decision-making based on individual patient change.[15] This is where MCID is important to consider. MCID represents a small ES as it is the minimum point gain on a measurement instrument indicating clinical improvement.[42] Anchor-based (aka externally referenced) methods are used to identify MCID. Anchor-based methods primarily involve using an independent and external instrument or criterion (i.e., anchor) that measures change in the patient's condition, function, or activity to examine the MCID of the reference instrument. The advantage of this method is that a robust clinically important difference (RCID) can be established as one or more independent measures can be compared to a single reference instrument.[39]

Identifying MCID associated with this method is a complex process involving multiple steps and statistical methods. This process largely depends on the anchors selected and a statistical platform will be needed for analysis. For instance, clinicians and researchers should be familiar with all the measures used, a patient-reported outcome measure should be the primary anchor (e.g., Global Impression of Change Scale) and empirical correlation (usually Spearman's correlations coefficient) of at least 0.5 (>0.7 is preferable) between the anchor/s and the reference measure is needed.[39] Several different types of anchors can be used, some of which are considered ESs (see Figure 1). If more than one anchor is used triangulation of the resultant MCID values will be needed. Complicating matters, patient-reported outcome measures as a primary anchor may not be possible in some patient cohorts due to cognitive capacity. In this instance, similar measures to the reference measure, a checklist using the clinician's perspective regarding discrete patient change (e.g., independence in transfers) and/or a functional outcome measurement instrument (e.g., the Functional Ambulation Categories) can be used.[22]

Anchor-based methods should be the primary method used for estimating MCID over distribution-based methods This is because this method quantifies patient change relative to a measurement instrument.[15] However, distribution-based methods do also have a role. For instance, the Norman et al., formula (see Table 1) can be used to reveal small but important patient change from an intervention as indicated on the reference instrument and the Reliable Change Index statistic can show whether a score change is statistically significant.[18,22] Due to the complex nature of incorporating both methods in determining MCID a full breakdown is out of the scope of this paper. Of the many methodological papers in the literature that can assist here, the paper by Malec and Ketchum is a standout as it provides step-by-step instructions.[39]

 https://doi.org/10.37464/2025.422.1949

REPORTING ES

It is essential that results of clinical research be conveyed to consumers in ways that accurately informs clinical significance and ultimately decision-making; '*p*-values do not serve that function. Nor do statistics like Odds Ratio'.[12(p.307)] Furthermore, researchers should not treat *p*-values as a surrogate for ESs as they are not synonymous with clinical significance. This is due to a *p*-value primarily reflecting the quality of research design decisions including statistical tests, analytical procedures and reliability of measures used and above all, sample size.[12] Consequently, in addition to the *p*-value researchers should report both absolute and relative ES along with SDs and CIs with all general results. This, however, is not common practice. Consumers of research need ESs as they show the size of the substantive significance (magnitude) of an effect which aids in determining the clinical significance of research results. Reporting more than one ES is a consideration here. For example, Kraemer et al., recommends reporting SRD, NNT and ROC curves for studies that compare two samples, such as in RCTs.[12]

Solely reporting an ES by itself is meaningless for a consumer as it can mean almost anything. A small ES can have clinical significance in one context but not another, whereas a large ES might have relatively less importance or persuasive. Consequently, reported a ES needs some narrative contextualising it against some frame of reference, such as a well-known scale, outcome, patient experience, previous study, or functional based change.[1] Narrative around the index (e.g., Cohen's *d*) used for obtaining the ES, quantifying the magnitude of the effect and a qualitative explanation of the effect regarding everyday practice is also needed to fully appreciate the clinical significance and utility of the effect. Finally, reporting ES as part of general results can aid future research as they can be used for priori power analysis.

The above recommendations are not new. In 1999 The Task Force on Statistical Inference of the American Psychological Association (APA) outlined similar expectations of researchers as part of their common reporting standards across research designs.[43] These recommendations continue today in the current APA Manual (v7) that notes the importance of reporting ES so the consumer can fully understand the importance of a study's results. Even before these APA recommendations, Kazis et al., advocated for 'the use of effect sizes as a method for estimating and communicating the extent of health status change that occurs in a group.'[11(p.188)] They further go on to advise that ES should supplement statistical significance testing in interpreting results and when reported, assist in comparing results across studies.[11]

Reporting CIs around ES should be considered but are not commonly reported.[9,29] This is because the connection between an ES and statistical significance (i.e., *p*-value) is via CI width. CIs provide a range of internal result estimates being a measure of imprecision or uncertainty of the true effect. That is, 'a confidence interval can also be defined as a point estimate of a parameter (or an effect size) plus or minus a margin of error.'[1(p.17)] The wider the CI the less certain or precise the true effect is and if the CI crosses zero, a result is not statistically significant. A CI of 95% means the true effect lies within the lower and upper CI limits 95% of the time. It is important to consider that a CI width is inversely proportional to sample size so the larger the sample size the narrower the probability and more precise the effect distribution (i.e., CI) is likely to be. The relationship between these aspects is illustrated in Figure 2 through four hypothetical scenarios.

This figure illustrates that while a study's results may be statistically significant, they may not be clinically significant due to the desired ES (> 0.8) being based on a therapeutic decision limit (4% drop in blood pressure). It also illustrates how CI width is inversely proportional to sample size (scenario 4).

INTERPRETING ES

Interpreting the clinical significance of an ES as it relates to results and the clinical context can be difficult. This is mainly due to the numerous factors needing consideration. Some include the type of ES reported, how the ES was derived, what the ES is to be used for, the context (empirical vs clinical), the relevant diagnosis related group (DRG), and how the ES was identified (i.e., using an outcome measure or statistical test). Findings from four inpatient rehabilitation studies are used to illustrate how ESs can be interpreted (see Table 4).

Drop in BP (%)	0 1 2 3 4 5 6 7 8	Significance
Scenario 1 2 mg Perindopril		Not statistically significant & no clinical significance. ES < 0.2, *N* = 50
Scenario 2 4 mg Perindopril		Statistically significant but no clinical significance. ES < 0.2, *N* = 50
Scenario 3 8 mg Perindopril		Statistically significant & clinical significance. ES > 0.8, *N* = 50
Scenario 4 8 mg Perindopril		Not statistically significant with possibly clinical significance. ES > 0.8, *N* = 10

4% drop in BP = Clinical significance decision limit

FIGURE 2. RELATIONSHIP BETWEEN STATISTICAL AND CLINICAL SIGNIFICANCE, SAMPLE SIZE (N) AND EFFECT SIZE (ES)

 https://doi.org/10.37464/2025.422.1949

TABLE 4: EXEMPLAR OF STUDIES ILLUSTRATING THE DIFFERENCE BETWEEN ES AND DRGS IN INPATIENT REHABILITATION

Studies	DRG	Measures	AFG (SD)	*d*	Δ	Cohen's U_3 index	MDC_{95}
Van der Putten et al[30]	MS	FIM total BI (0-20)	6.9 (8.3) 2.1 (2.4)		0.30 0.37	61.8% 64.4%	
	Stroke	FIM total BI (0-20)	21.9 (19.0) 5.2 (4.4)		0.82 0.95	79.4% 82.9%	
Houlden et al[45]	BI vascular TBI	FIM total	17.3 (15.1) 17.4 (15)		0.59 0.52	72.2% 69.8%	
	BI vascular TBI	BI (0-20)	3.9 (3.4) 3.95 (3.4)		0.65 0.55	74.2% 70.9%	
McKechnie et al[44]	TBI without RTAC TBI with RTAC	FIM total FIM total	28.2 (25.8) 33.3 (32.3)		0.85 1.21	80.2% 88.7%	11.9
Arcolin et al[22]	Hip fracture	FIM total BI (0-100)	24.4 (11.8) 23.4 (15.1)	1.39 1.35		91.8% 91.1%	10.3

Abbreviations: DRG, diagnosis related group; AFG, absolute functional grain (discharge mean – admission mean); SD, standard deviation; *d*, Cohen's SD_{pooled}; Δ, Glass's delta; MS, Multiple Sclerosis; BI, brain injury; TBI, traumatic brain injury; RTAC, readmission to acute care; FIM, Functional Independence Measure (18-126 score range); BI, Barthel Index

Note: Cohen's U_3 index[34] based on Cohen's SD_{pooled} or Glass's delta and used for comparison between pre vs post treatment and not between two independent groups

The Functional Independence Measure (FIM) and/or Barthel Index (BI) were used in the four exemplar studies to examine the effectiveness of inpatient rehabilitation for five DRGs. As per the results in Table 4, hip fracture,[22] brain injury,[44] and stroke patients,[30] appeared to have a moderate to large benefit from inpatient rehabilitation as they all had ESs above 0.5. The FIM and BI derived ESs for the multiple sclerosis (MS) cohort equated to small effects being 0.30 and 0.37, respectively. This is an important finding as it shows that the effectiveness of inpatient rehabilitation differs between DRGs, with MS patients possibly only obtaining modest benefits from this intervention compared to other DRGs. This is further evidenced by the MS cohort having low absolute function gain scores. MS is a degenerative condition that possibly mitigates some capacity for MS patients to improve from rehabilitation compared to other DRGs with newly acquired conditions who are likely to have more capacity to improve.

Several other inferences can be drawn from the ES results data in Table 4. First, rehabilitation had a large effect on traumatic brain injury (TBI) patients irrespective of their rehabilitation program being interrupted resulting from readmission to acute care (ES = 1.21).[44] Second, TBI patients in a specialist TBI inpatient rehabilitation unit had larger functional gains (ES = 0.84 and 1.21) compared to those admitted to a general neurological rehabilitation unit (ES = 0.52 and 0.55).[44,45] Third, the FIM and BI had similar responsiveness for these DRGs as their ESs were comparable. Finally, the FIM's MDC is approximately 11 points of total FIM score which stands to reason as FIM is an 18-item ordinal scale with item scores ranging from 1 for completely dependent to 7 for independent.

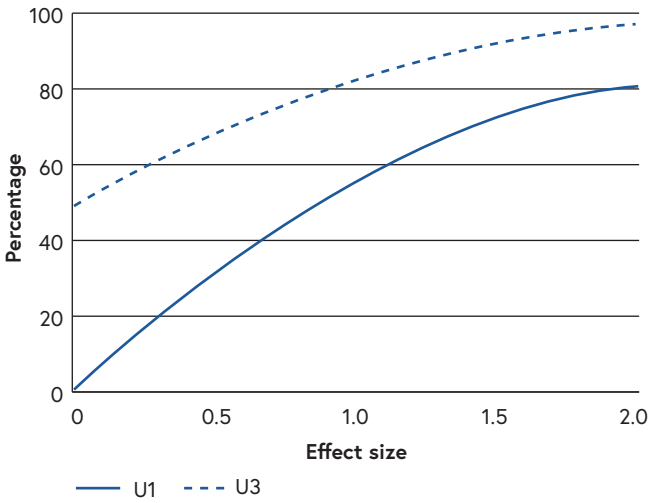

FIGURE 3: EQUIVALENTS OF COHEN'S *d* REPRESENTED AS U INDICES

Another relatively unfamiliar way to interpret the clinical significance of ES is by using an improvement index to convert the ES value into percentile gain manifested by the target group.[3] Being derived from equivalents of Cohen's *d*, the U indexes can be used here (see Table 4 and Figure 3).[34(pp.21–22)] For example, an ES of *d* = .30 indicates that 61.8% of the treatment group will be above the mean of the control group (Cohen's U_3) representing a 62% improvement in the treatment group. Cohen's U indexes can also be interpreted in terms of percentage of non-overlap (U_1) between treatment group and untreated group scores on a bell curve. In this instance, an ES of .30 indicates that 21% of the two groups' scores will not overlap (79% overlap).

 https://doi.org/10.37464/2025.422.1949

An automated online tool for these calculations that also provides an interpretation of the results is available at https://rpsychologist.com/cohend/. The website's author notes differences in their results compared to Cohen's regarding the percentage of non-overlap (Cohen's U_1) and provides a detailed rationale for the inconsistencies stating that his calculations are more robust.[46] Using the above example, $d = .30$ equates to 12% non-overlap based on his calculations. The improvement per cent index (Cohen's U_3) remains the same.

IMPLICATIONS FOR RESEARCH, POLICY AND PRACTICE

Clinical nurses use patient change and measurement instruments to rationalise clinical significance and their resultant interventions. They should also consider the magnitude of effect and the responsiveness of instruments they use for more robust evidenced-based clinical decision making. Nurse researchers in the first instances should always attempt to identify population ESs worth investigating as they would be applied to their context. Both clinical nurses and nurse researchers need to understand aspects of ES to realise these goals. Considerations for measuring and quantifying patient change with ESs has been discussed. In doing so, this paper aids clinical nurses and nurse researchers in using ES to inform clinical decision making and report clinically meaningful research results.

CONCLUSION

This paper provides clinical nurses and nurse researchers with a broad overview on determining clinically significant patient change using ESs. In doing so, it provides guidance on how to critique research literature and apply ES in clinical and research contexts. This paper aids nurses to effect change based on informed decision making thus strengthening their evidence-based practice.

Disclosure of funding: The author reports that there was no funding for all or any part of this study.

Declaration of conflict of interest: The author reports no conflicts of interest.

REFERENCES

1. Ellis PD. Essential Guide to Effect Sizes. Cambridge, UK: Cambridge University Press; 2010.
2. Schneider Z, Whitehead D, LoBiondo-Wood G, Haber J. Nursing and Midwifery Research: methods and appraisal for evidence based practice. 5th ed. Sydney, Australia: Elsevier Australia; 2016.
3. Durlak JA. How to select, calculate, and interpret effect sizes. J Pediatr Psychol. 2009;34(9):917-28.
4. Portney LG. Foundations of Clinical Research: Applications to Evidence-Based Practice. 4th ed. Philadelphia, PA: F. A. Davis Company; 2020.
5. Fisher MJ, Fethney J. Analysing data in quantitative research. In: Schneider Z, Whitehead D, LoBiondo-Wood G, Haber J, editors. Nursing and Midwifery Research: methods and appraisal for evidence based practice. 5th ed. Sydney, Australia: Elsevier Australia; 2016. p. 213-36.
6. Krzywinski M, Altman N. Power and sample size. Nat Methods. 2013;10(12):1139-40.
7. Lenth RV. Post hoc power: Tables and commentary. 2007:1-13.
8. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-76.
9. Aberson CL. Applied Power Analysis for the Behavioral Sciences. 2nd ed. New York, USA: Taylor & Francis; 2019.
10. Nuzzo RL. Statistical Power. AAPM&R. 2016;8(9):907-12.
11. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27(3).
12. Kraemer HC, Neri E, Spiegel D. Wrangling with p-values versus effect sizes to improve medical decision-making: A tutorial. Int J Eat Disord. 2020;53(2):302-8.
13. Lin M, Lucas HC, Shmueli G. Too Big to Fail: Large Samples and the p-Value Problem. Inform Syst Res. 2013;24(4):906-17.
14. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry. 2006;59(11):990-6.
15. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoeconomics Outcomes Res. 2011;11(2):171-84.
16. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. Estimating clinically significant differences in quality of life outcomes. Qual Life Res. 2005;14:285-95.
17. Newcombe RG. A deficiency of the odds ratio as a measure of effect size. Stat Med. 2006;25(24):4235-440.
18. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life. Med Care. 2003;41(5):582-92.
19. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ. 2020;369:1-11.
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. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002;77(4):371-83.
22. Arcolin I, Godi M, Giardini M, Guglielmetti S, Bellotti L, Corna S. Minimal clinically important difference of the functional independence measure in older adults with hip fracture. Disabil Rehabil. 2023;ePub:1-8.
23. Williams VJ, Piva SR, Irrgang JJ, Crossley C, Fitzgerald GK. Comparison of reliability and responsiveness of patient-reported clinical outcome measures in knee osteoarthritis rehabilitation. J Orthop Sports Phys Ther. 2012;42(8):716-23.
24. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine. 2000;25(24):3192–9.

25. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53(5):459-68.
26. Guyatt GH, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171-8.
27. Cohen J. Things I have learned (so far). Am Psychol. 1990;45(12):1304-12.
28. Volker MA. Reporting effect size estimates in school psychology research. Psychol Sch. 2006;43(6):653-72.
29. Vacha-Haase T, Thompson B. How to Estimate and Interpret Various Effect Sizes. J Couns Psychol. 2004;51(4):473-81.
30. Van der Putten JJ, Hobart JC, Freeman J, Thompson AJ. Measuring change in disability after inpatient rehabilitation: Comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. J Neurol Neurosurg Psychiatry. 1999;66(4):480-4.
31. Marfo P, Okyere GA. The accuracy of effect-size estimates under normals and contaminated normals in meta-analysis. Heliyon. 2019;5(6):e01838.
32. Deo SV, Deo V, Sundaram V. Survival analysis-part 2: Cox proportional hazards model. Indian J Thorac Cardiovasc Surg. 2021;37(2):229-33.
33. Grissom JR, Kim JJ. Review of assumptions and problems in the appropriate conceptualization of effect size. Psychol Methods. 2001;6(2):135–46.
34. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2 nd ed. New York, USA: Lawrence Erlbaum Associates; 1988.
35. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. J Educ Stat. 1981;6(2):106-28.
36. Mouelhi Y, Jouve E, Castelli C, Gentile S. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. Health Qual Life Outcomes. 2020;18(1):1-17.
37. Beaton DE, Bombardier C, Katz JN, Wright JG, G. W, Boers M, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. J Rheumatol. 2001;28(2):400-5.
38. Jacobson NS, Truax P. Clinical significance: A statistical approach to denning meaningful change in psychotherapy research. JCCP. 1991;59(1):12-9.
39. Malec JF, Ketchum JM. A standard method for determining the Minimal Clinically Important Difference for rehabilitation measures. Arch Phys Med Rehabil. 2020;101(6):1090-4.
40. Ferguson CJ. An effect size primer: A guide for clinicians and researchers. Prof Psychol Res Pr. 2009;40(5):532-8.
41. McDowell I. Measuring Health: A Guide to Rating Scales and Questionnaires. 3rd ed. Oxford, New York: Oxford University Press, Inc.; 2006.
42. Copay AG, Subach BR, Glassman SD, Polly DW, Jr., Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J. 2007;7(5):541-6.
43. Wilkinson L, Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. Am Psychol. 1999;54:594–604.
44. McKechnie D, Pryor J, McKechnie R, Fisher MJ. Predictors of readmission to acute care from inpatient rehabilitation: An integrative review. PM&R. 2019;11(12):1335-45.
45. Houlden H, Edwards M, McNeil J, Greenwood R. Use of the Barthel Index and the Functional Independence Measure during early inpatient rehabilitation after single incident brain injury. Clin Rehabil. 2006;20(2):153-9.
46. Magnusson K. Interpreting Cohen's d Effect Size: An interactive visualization 2024 [Available from: https://rpsychologist.com/cohend/.