

Effect size measures and their benchmark values for quantifying benefit or risk of medicinal products

Volker Rahlfs  | Helmuth Zimmermann

idv – Data Analysis and Study Planning,
Gauting, Germany

Correspondence

Volker Rahlfs, idv – Data Analysis and Study
Planning, Tassilo Str. 6, 82131 Gauting,
Germany.

Email: v.rahlfs@t-online.de

Abstract

The standardized mean difference is a well-known effect size measure for continuous, normally distributed data. In this paper we present a general basis for important other distribution families. As a general concept, usable for every distribution family, we introduce the relative effect, also called Mann–Whitney effect size measure of stochastic superiority. This measure is a truly robust measure, needing no assumptions about a distribution family. It is thus the preferred tool for assumption-free, confirmatory studies. For normal distribution shift, proportional odds, and proportional hazards, we show how to derive many global values such as risk difference average, risk difference extremum, and odds ratio extremum. We demonstrate that the well-known benchmark values of Cohen with respect to group differences—small, medium, large—can be translated easily into corresponding Mann–Whitney values. From these, we get benchmarks for parameters of other distribution families. Furthermore, it is shown that local measures based on binary data (2×2 tables) can be associated with the Mann–Whitney measure: The concept of stochastic superiority can always be used. It is a general statistical value in every distribution family. It therefore yields a procedure for standardizing the assessment of effect size measures. We look at the aspect of relevance of an effect size and—introducing confidence intervals—present some examples for use in statistical practice.

KEYWORDS

binary, continuous data, clinical relevance, effect size measures, Mann–Whitney measure, transformation of measures, ordinal

1 | INTRODUCTION

Effect size measures to calculate size of benefit or risk are the basis of data analysis in clinical research. They are important tools for quantifying benefit or risk of a medicinal product. Benchmark values have been developed for defining relevance of effect size quantities in various medical fields. Examples in drug research are those given for the relative risk or risk ratio of a standard 2×2 table of frequencies (Skipka et al., 2016). Effect size measures in the current literature are commonly based on continuous data (raw scale or standardized mean difference) or binary data (risk difference, risk ratio, odds ratio, hazard ratio). Appropriate procedures for the analysis of ordinal data are rarely seen, although the odds ratio has been discussed in the handbook about

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2019 The Authors. *Biometrical Journal* Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

General Methods of the Institute for Quality and Efficiency in Healthcare, in short IQWiG (version 4.2, dated 22.4.2015), citing a note in the guidelines of the *Cochrane Collaboration Handbook* (Chapter 5). The analysis of risk values seems to be by far the most popular procedure, the binary data being genuine dichotomies or derived by dichotomization of a continuous or ordinal scale. (See also *Cochrane Handbook for Systematic Review of Intervention, Version 5.1.0* (2011), chapter 12.6.3.)

We give a detailed overview of the most useful parameters and their often used functions. This overview is based on the measure of superiority $P(Y < X)$, known also as the Mann–Whitney effect size measure MW .

This measure always exists. It can be used to compare effect size measures given by other procedures. Although some of these measures have been published in an isolated fashion in scientific journals, we present a structured overview of the most important parameters and some of their helpful functions. Only a small part of our scheme has appeared previously in the literature and comparisons between distribution families are new.

For the normal distribution MW can be expressed by the parameter standardized mean difference δ . For the Lehmann family MW is a function of the hazard rate ratio γ . Finally for the proportional odds family the formula for calculating MW from the proportional odds parameter β is given.

We demonstrate some very useful characteristics of the Mann–Whitney measure. Care should be taken when searching the literature for references to the Mann–Whitney measure because this measure has been invented, reinvented, and advocated in the statistical literature with a plethora of different names: “Measure of stochastic superiority of Y2 over Y1” (Klotz, 1966), “Common Language Effect Size Indicator” (McGraw & Wong, 1992), “Relative Effect” (Brunner, Domhof, & Langer, 2002; Brunner & Langer, 1999), “Area Under the Curve” (of a receiver operating characteristic), AUC (Kraemer & Kupfer, 2006) “probability index” (Acion, Peterson, Temple, & Arndt, 2006), among many others.

Advantages of the Mann–Whitney effect size measure are:

- It is defined for all distribution families.
- It is a universally applicable measure for superiority with a very general probabilistic interpretation and can be expressed by functions of the defining parameters for special distribution families.
- It is associated with a well-known graphical procedure, the percentile–percentile (P-P) plot, identical to the “Receiver Operating Characteristic” (ROC), well-known in the field of diagnostic research.
- It can be used for the analysis of continuous, quasi-continuous, ordinal, and even binary data.

We show that there are three simple functions of the Mann–Whitney effect size measure, all three of which provide useful interpretations of the data: Mann–Whitney odds (MW_{odds}), risk difference average (RD_{average}), and Mann–Whitney difference (MWD).

To fix ideas about the parameters and their relationships, we begin in Section 2 with a detailed description of formulas within several distribution families and their embedded parameters. Section 3.1 presents the classification scheme as a structured overview, with parameters and formulas, as well as benchmark values equivalent. Section 3.2 describes more practical aspects for the Mann–Whitney measure. Section 3.3 finally provides examples showing many different ways of comparing effect size measures and their interpretation using confidence intervals. The conclusion is a discussion of various related topics (Section 4).

2 | THE MANN–WHITNEY EFFECT SIZE MEASURE, GLOBAL, AND LOCAL DEFINITIONS

2.1 | The Mann–Whitney measure

In the following we present some technical details for the robust Mann–Whitney measure (MW) and two other useful robust measures derived from the MW measure. We then show some interesting features of the MW measure assuming a certain distribution family. These features make interpretations of the MW measure possible, even if the assumptions hold only approximately in practice.

If we have two random variables X, Y with distributions functions $F(t) = P(X \leq t)$, and $G(t) = P(Y \leq t)$ then the Mann–Whitney measure is defined as the probability $MW = P(Y \leq X) = \int_0^1 G dF$.

There are derived measures that are often cited in the literature. One is $MW_{\text{odds}} = MW/(1 - MW)$ that will be shown later to be the estimator of the hazard rate ratio γ under the Lehmann alternative. Another, less well known, is the average risk difference, defined as $RD_{\text{average}} = \int_0^1 (G - F) dF$ that reduces to $RD_{\text{average}} = MW - 0.5$. This is half the well-known Mann–Whitney difference (MWD), sometimes incorrectly called risk difference.

There even exist averages of other measures such as RR or OR . These can be formally defined as $RR_{\text{average}} = \int_0^1 RR dF$ and $OR_{\text{average}} = \int_0^1 OR dF$. These integrals are usually approximated numerically.

Also of interest are measures and relations based on the assumption of some distribution family. There are three important families: shift, proportional odds, and proportional hazard rate. In the following the Mann–Whitney measure is used to give benchmarks for parameters of these families that result in the same MWs .

2.2 | Global definitions

In the following we will distinguish between local measures at some threshold t_0 , $F(t_0) = CER$ (control event rate) and $G(t_0) = TER$ (test event rate), so called event rates and global parameters of distribution families. The most locally important derived measures are $RD = TER - CER$, $RR = TER/CER$, $OR = TER/(1 - TER)/CER(1 - CER)$.

With the relation $OR \cdot (RR - 1 - RR \cdot RD) = RR \cdot (RR - 1 - RD)$, we can find any of the three measures RD , RR , and OR given the other two. From any two of these, it is easy to calculate CER and TER .

The interesting fact for the three distribution families is that the MW measure can be used to calculate the defining parameters of all three families: δ for normal distribution, β for proportional odds, γ for proportional hazard.

2.2.1 | Location distribution (shift)

Given are two distributions characterizing shift of parameters of some base distribution assuming equal $\sigma = 1$ for formula simplicity. $F(t) = Y(t - \mu_x)$ and $G(t) = Y(t - \mu_y)$ or $Y^{-1}(F(t)) = t - \mu_x$ and $Y^{-1}(G(t)) = t - \mu_y$.

The standardized mean difference is given by $\delta = (\mu_y - \mu_x) = Y^{-1}(F(t)) - Y^{-1}(G(t))$. If δ is the effect size parameter of the normal distribution family, we find $MW = P(Y \leq X) = \Phi(-\delta/\sqrt{2})$ and thus $RD_{\text{average}} = \Phi(-\delta/\sqrt{2}) - 0.5$.

Two other useful measures can also be defined for the normal distribution family: One is the extreme risk difference given by $RD_{\text{extreme}} = \Phi(-\delta/2) - \Phi(\delta/2)$. Here, it is easy to see that this is smaller (in absolute value) than the Mann–Whitney difference given by $MWD = \Phi(-\delta/\sqrt{2}) - \Phi(\delta/\sqrt{2})$. (This is true for convex relations in general as shown in Zimmermann and Rahlfs (2014).) The other is the extreme odds ratio given by $OR_{\text{extreme}} = \{\Phi(-\delta/2)/\Phi(\delta/2)\}^2$. (This relation was derived by Trichtler (1995) using a different approach.)

2.2.2 | Proportional odds family

Within this family, $F(t)$ and $G(t)$ differ according to relation (we omit the variable t if not necessary) by $G/(1 - G) = \beta \cdot F/(1 - F)$, where β is the constant odds ratio within the family used as effect size parameter. Here, we find $MW = P(Y \leq X) = \beta/(\beta - 1) \cdot \{1 - \log(\beta)/(\beta - 1)\}$. $MW_{\text{odds}} = MW/(1 - MW)$, the odds of MW , is not equal to β above.

Here $RR_{\text{average}} = \beta \cdot \log(\beta)/(\beta - 1)$ can be calculated in closed form. The extreme risk difference is given by $RD_{\text{extreme}} = (\sqrt{\beta} - 1)/(\sqrt{\beta} + 1)$ at $CER = 1/(\sqrt{\beta} + 1)$. For the logistic distribution, there is the special relation $\beta_{\text{logistic}} = \exp[\pi(\mu_y - \mu_x)/\sqrt{3}]$. This must not lead, however, to simply translating Cohen's normal shift to the shift in logistic distributions because the resulting MW will not be the same.

For the very simple case of F uniform and G as above, it can be seen that these distributions have very different shapes and therefore cannot be shifted although they belong to the proportional odds family. (The logistic distribution is the only member of the proportional odds family that also belongs to the shift family.)

2.2.3 | Proportional hazard rate

The hazard rate of a distribution function F with density $F' = f$ is defined as $hr = f/(1 - F)$. The hazard rate ratio of the two functions F, G , therefore, is $HR = dG/dF(1 - F)/(1 - G)$.

Assuming the so-called Lehmann alternative (HR constant = γ), we find the solution of the differential equation above as $1 - G = (1 - F)^\gamma$. Again we calculate $MW = \int_0^1 1 - (1 - F)^\gamma dF = \gamma/(1 + \gamma)$ and thus $\gamma = MW_{\text{odds}}$ can be used as the effect size parameter of proportional hazard. With $help = -\log(\gamma)/(\gamma - 1)$, we find $RD_{\text{extreme}} = help - help^\gamma$. The extreme odds ratio here is just the parameter γ or 1.

Again there is no need to assume exponential distribution. We can take F as uniform and define $F = t$ and $G = 1 - (1 - t)^\gamma t \in (0, 1)$. Both, F and G , are quite different from an exponential distribution but fulfill the Lehmann alternative.

2.3 | From local to global measures

If in the case of a simple 2×2 table only event rates CER and TER are given, we can proceed to global measures assuming a specific distribution model. Assuming that the rates used are exact (or nearly so), we find good estimates from them.

2.3.1 | Normal shift model assumed

Cohen's δ can be estimated with $\delta = \Phi^{-1}(CER) - \Phi^{-1}(TER)$ for every pair (CER, TER) . Here, $MW = \Phi(-\delta/\sqrt{2})$. Within this model, all the locally defined measures as RD , RR , OR , and HR can be derived. The formula for HR is: $HR = f_y/f_x \cdot (1 - CER)/(1 - TER)$ where $f_x = \exp[-\{\Phi^{-1}(CER)\}^2/2]$ and $f_y = \exp[-\{\Phi^{-1}(TER)\}^2/2]$.

2.3.2 | Proportional odds model assumed

The odds ratio here is calculated as $\beta = TER/(1 - TER) \cdot (1 - CER)/CER$ for any pair (CER, TER) and MW is given as $MW = \beta/(\beta - 1) \cdot \{1 - \log(\beta)/(\beta - 1)\}$. Within the proportional odds model the formula for HR can be derived using $dG/dF = G/F(1 - G)/(1 - F)$ and one gets HR equal to RR . This relation seems not to be well known but could be very interesting for interpretations.

2.3.3 | Lehmann alternative assumed

The hazard rate ratio is estimated as $\gamma = \log(1 - TER)/\log(1 - CER)$ for every pair (CER, TER) . This yields $MW = \gamma/(1 + \gamma)$

3 | OVERVIEW OF EFFECT SIZE MEASURES, THE PROMINENT ROLE OF THE MW MEASURE, AND EXAMPLES FOR THE PRACTICAL WORKER

3.1 | Classification scheme for measures and their relationships with given distribution assumptions

Table 1 gives an overview of various effect size measures expressed as functions of $MW = P(Y \leq X)$. Four of them—Mann–Whitney effect size measure (MW), the odds of MW , the average (expected) risk difference (RD), the Mann–Whitney difference (MWD)—enjoy general validity and thus are fundamental for working with real-world data. They are placed at the top of the table.

The larger part of the table presents known and some less-known effect size measures expressed as functions of MW for a given specific distribution family.

We demonstrate the usefulness of the classification scheme by filling each box below the formula with equivalent quantities for well-known benchmark values of “small,” “medium-sized,” and “large.” Although any effect size measure or quantity could have been chosen as a starting point, we based all calculations on the popular benchmark values of Cohen's standardized mean value δ : 0.2, 0.5, and 0.8 (Cohen, 1969, 1977).

The classification scheme is a base for handling continuous data. For binary data, locally measured values such as risk difference (RD) or odds ratio (OR) are commonly used. For continuous data, it is possible to use globally defined measures such as risk difference average, risk difference extremum, or odds ratio extremum.

Using the MW values defined by Cohen's benchmarks, we find all corresponding effect sizes for all families: normal, proportional odds or proportional hazard. This is possible because their parameters are comparable through their relationship to the MW measure.

3.2 | The MW effect size measure: Practical aspects and interpretation

The MW effect size measure is a general measure for describing any useful data situation of superiority (or inferiority) of one group compared to another in a typical two-group comparison of an experimental/clinical study. The interpretation is often described as follows: the MW measure is the probability that a randomly selected patient from a test group fares better than a randomly selected patient from the reference group. A more concrete definition is based on numbers: Assume that there are two teams, X and Y . They play a match where every player of one team plays against every player of the other. How often the X -players win is determined and this number is divided by the total number of matches. The result is the MW value, often written as $P(Y < X)$ without tied values (no win or lose), or as $P(Y < X) + 0.5 P(Y = X)$ including tied values. Thus, we have a probability measure. Simply, the result of a group comparison can be formulated as “the probability of a better outcome” (Colditz, Miller,

TABLE 1 Definitions of effect size measures and pathways between them as well as transformation formulas are given and effect sizes derived from Cohen’s benchmark values: SMD = 0.2 (small), 0.5 (medium-sized), and 0.8 (large) for relevance of a difference

Effect size measures with relationships				
Robust/assumption free				
Magnitude	MW	MWD	MW _{odds}	RD _{average}
Small	0.444	-0.113	0.798	-0.056
Medium	0.362	-0.276	0.567	-0.138
Large	0.286	-0.428	0.400	-0.214
Measures with distribution assumption				
	Normal distribution Cohen's δ	Proportional odds Odds ratio β	Lehmann alternative Hazard rate ratio γ	
Parameters	$\delta = -\Phi^{-1}(MW) \cdot \sqrt{2}$	$\beta =$ Numerical solution of $MW = \frac{\beta * (\beta - 1 - \ln(\beta))}{(\beta - 1)^2}$	$\gamma = \frac{MW}{1 - MW}$	
	0.200	0.713	0.798	
	0.500	0.428	0.567	
	0.800	0.256	0.400	
Odds ratio extreme	$\left(\frac{\Phi(-\delta/2)}{\Phi(+\delta/2)}\right)^2$	Constant = β	From $\gamma \rightarrow 0$	
	0.727	0.713	0.798	
	0.450	0.428	0.567	
	0.277	0.256	0.400	
Risk difference extreme	$\Phi\left(\frac{-\delta}{2}\right) - \Phi\left(\frac{+\delta}{2}\right)$	$\frac{\sqrt{\beta} - 1}{\sqrt{\beta} + 1}$	$^{(1-\gamma)}\sqrt{\gamma}\left(1 - \frac{1}{\gamma}\right)$	
	-0.080	-0.085	-0.083	
	-0.197	-0.209	-0.206	
	-0.311	-0.329	-0.326	

Remark: the direction of the effect size measures could also be reversed: SMD could be minus instead of plus; MW, reflected around 0.5; OR, larger than 1 or reciprocal. The direction of superiority, however, must be defined.

& Mosteller, 1988). This measure by its very nature is assumption-free and thus conforms to the requirement of “minimizing the required assumptions of analysis procedures” (LaVange, Durham, & Koch, 2005, Saville, LaVange, & Koch, 2011).

Statistical properties of the procedure are well known. The *MW* measure can be estimated for arbitrary distributions even in the binary case. Therefore, it can be used as a universal tool for comparing benefit or risk across different scales in a clinical study or across several studies in a meta-analysis.

Until now effect size measures were mostly based on binary data (2 × 2 table), either genuine binary data (e.g. mortality) or those derived by dichotomizing a continuous or ordinal scale at some cutoff point. Against this there are two objections. First, the choice of the cutoff point is more or less arbitrary unless it has been accepted by the scientific community. Second, although the so-called responder analysis is sometimes useful for clinical interpretation, the process of dichotomization wastes patient information as has been demonstrated by many researchers. (see Uryniak et al., 2011). Thus, whenever continuous, quasi-continuous or ordinal data are available, they should be analyzed using the Mann–Whitney efficacy measure and related measures.

The use of the *MW* measure for obtaining a good measure of relevance in clinical research has been recommended for many years. We cite Brunner and Munzel, 2002, Colditz et al., 1988, D’Agostino, Campbell, and Greenhouse, 2006, Munzel and Hauschke, 2003, Newcombe 2006, Wei and Lachin, 1984, and Wolfe and Hogg, 1971, among others. Recently, the *MW* measure was proposed as the best measure of relevance in clinical research, inasmuch as it is a global measure not bound to thresholds such as local RD, RR, and OR, and also generally more efficient than the hitherto fashionable responder rate analysis based on binary data (Kieser, 2014; Kieser, Friede, & Gondan, 2013). We also cite Bordley (2009) who described *MW* as the probability that a treatment fulfills the Hippocratic oath “to help or, at least, do no harm.”). Demidenko (2016) recently recommended the Mann–Whitney measure (called the “D-measure”) as a useful measure for personalized medicine when treatment is sought, not on a group, but on an individual level.

The *MW* effect measure is related to a special graph type, the percentile–percentile plot (P-P plot), a scatterplot of the empirical distribution functions (EDF) of two groups. This graph is also well known in the field of diagnostic research as the receiver operating curve (ROC) (see Brumback, Pepe, & Alonzo, 2006; Newcombe, 2006, part I; Schistermann, Reiser, & Faraggi, 2006, and others). The area under the empirical function (AUC) is just $P(Y \leq X)$ and the area between the curve and the diagonal is nothing else than RD_{average} just showing the deviation from pure chance. If the curve is completely above the diagonal, then the difference is called stochastic ordering or stochastic superiority. In the other case, it is called stochastic tendency or relative effect.

3.3 | Confidence intervals and their use in statistical practice

For test-based confidence intervals (CIs) there is a duality of CI and test with the advantage that the CI gives the acceptance region for the null hypothesis and herewith a hint about the precision of the estimator. In addition to this use of CIs, the toolbox of the statistical researcher is substantially enlarged when using the formulas given in Table 1 in Section 3.1. Effect size measures, bounds for confidence intervals, and also benchmark values for relevance or irrelevance, they all can be compared with values of an alternative model. Using the Mann–Whitney measure of superiority, which exists for any model, it is possible to compare parameters of different models (normal distribution, proportional odds ratio, proportional hazards ratio, etc.). In the following, we give examples for interpretation and interpretation after reexpression of values of a clinical study.

3.3.1 | Examples: Different outcomes and effect size measures

The handling of the problem of different outcomes and different effect size measures will be exemplified by inspecting the so-called “summary of findings table” that is the recommended overview table for the presentation of study results (Guyatt et al., 2013). We discuss Table 5 of the Guideline (p. 179). This table gives five entities of outcome scales (called A to E), taken from a summary of 16 studies with a total of $N = 816$ patients. We discuss the adequacy of tests and the merit of our conversion approach for parts A to D (part E is in principle a repeat of part A). All confidence intervals in the table are 95% CIs. Both sides of the CI are given, but in general, when testing for relevance (or irrelevance), only one side is of interest.

(A) Part A of Table 5 presents a summary measure for different continuous scales, obviously referring to the same entity. The effect size resulting is given as $SMD = 0.72$ (0.48–0.96). This value tells us that the null hypothesis of no difference ($SMD = 0.0$) can be rejected inasmuch as it is not contained in the CI. Using the CI, we can also provide statements about the clinical relevance of the observed difference. For interpreting SMD values, we use the benchmark values of Cohen 0.2, 0.5, and 0.8 for small, medium, and large, respectively, which are also cited in the comment column of Table 5 as “a rule of thumb.” Thus the estimator 0.72 being larger than 0.5 demonstrates at least medium-sized relevance. The lower bound of the CI is the important boundary for proving relevance. With a lower bound $CI = 0.48$, superiority is proven for at least the value 0.2, based on the data of this study (a study with more patients would have led to a smaller CI). The same is true if we convert SMDs to MW values. The resulting MW is 0.31 (0.37–0.25), leading to the same conclusions using the MW values of Table 1 corresponding to Cohen’s benchmark values. These can then be used to obtain benchmark values for other models.

Remark: The direction of the effect size measures could also be reversed: SMD could be minus instead of plus; MW, reflected around 0.5; OR, larger than 1 or reciprocal. The direction of superiority, however, must be defined.

(B) Part B of Table 5 gives again a summary of scales. All are compressed to a 7-point scale of ordered categories. The result is given as a mean difference in favor of the treatment group, $MD = 0.71$ (0.48–0.94). Apparently, it was assumed that the clinical researcher has experience with the scale of natural units going from 1 to 7.

Here, calculation of the WMW test and its associated MW values would have been helpful because this procedure is recommended for the analysis of scales with ordered categories (no interval scale). Then the MW value and its CI could have been interpreted using the benchmark values given for the MW measure. These benchmark values can be used for any other model, as for example for proportional odds as used in Table 5C.

(C) Part C of Table 5 presents effect size measures for the proportion of patients with improvement. The table shows risk difference as $RD = 0.31$ (0.22–0.40) based on 0.3 for control group and $OR = 3.36$ (2.31–4.86). (Although we could not reproduce this OR value using the data shown, we will use it in the following for demonstration purposes; its difference from our value is not large.)

The question is now whether the OR value indicates a relevant superiority for the test drug. Until now benchmark values for an OR value were not really known. But values equivalent to Cohen’s proposal can be obtained as 1.40, 2.34, and 3.91 for the proportional odds model (see also Table 1 in Section 3.1). The value 2.31, the lower bound of the CI, is larger than 1.40 so that at least a small treatment effect is proved (and nearly a medium-sized benefit). Thus the statement of relevance is nearly the same as that obtained with the continuous data SMD.

We calculated MW values derived from an OR value given in Table 5 and obtained $MW = 0.69$ (0.64–0.74) that equivalently can be interpreted with reference to the benchmark values 0.56 (small), 0.64 (medium), 0.71 (large).

- (D) Part D in Table 5 is supposed to give a ratio of mean values, but the studies cited in the table do not present results expressed as a ratio of means. This would have been an interesting information. Assuming a proportional hazard family with exponential distribution, the ratio of two means could be interpreted as an estimator for the so-called hazard rate ratio γ . For demonstration purposes, we now use the data of part D in Table 4 that gives the ratio of means $RM = 0.87$ (0.78–0.98). Using the formula $MW = \gamma/(1+\gamma)$, we obtain $MW = 0.467$ (0.438–0.494), that is no indication of relevance. Note, however, that the example in Table 4 is based on completely different studies.

4 | DISCUSSION

Currently, measures of relevance in clinical research are primarily based on binary data (*RR*, *OR*, *RD*, and *NNT*) and on some types of continuous data (*MD*, *SMD*, and *HR*) and in rare circumstances also on ordinal data (generalized proportional *OR*). In this paper, we present well-known and less-well-known measures/parameters for continuous data, all organized in a classification scheme with defined relationships based on algebraic formulas. Assuming a special data situation, normal distribution, proportional odds, or proportional hazards, all measures can be transformed to others within a family of distributions. For practical purposes, the normal shift and proportional odds situations are very similar, so that interchanging situations is reasonable, at least for interpreting measures of relevance. Thus, each of the well-known benchmark values for relevance, for instance Cohen's effect sizes, can now be compared to an equivalent quantity of another measure in the sense of equal stochastic superiority (see Table 1 in Section 3.1 in this paper). Operationalization of clinical relevance can be obtained with reference to the *SMD* measure. Recently, a proposal was made for defining Cohen's effect size $SMD = 0.2$ (small difference) as clinically irrelevant (IQWiG, Vers. 5.0, 2017). Having accepted this value, it is easy to translate it to an equivalent value (with the same Mann–Whitney value) for the parameters of other distribution families.

Our classification scheme is based on the robust assumption-free Mann–Whitney measure of stochastic superiority (*MW*) and its robust derivations, odds (not odds ratio!), and average risk difference (RD_{average}). The *MW* measure is not only interesting in itself, it is also a key measure for many useful relations to other measures. The value of *MW* given $OR = \beta$ constant was determined in Section 2.2 when discussing the proportional odds family. Previously, this relation has never been cited in the literature (to the best of our knowledge), although it is extremely important because the odds ratio is often used when analyzing ordinal data (Bender & Grouven, 1998; Bolland, Sooriyarachchi, & Whitehead, 1998; Lu & Tilley, 2001; McHugh et al., 2010; Savitz, Lew, Bluhmki, Hacke, & Fisher, 2007; Tilley et al., 1996; Tilley, 2012; Whitehead, 1993; Whitehead et al. 2010). It is of interest that Agresti's α , defined as $P(Y > X)/P(X > Y)$ is sometimes alleged to be the odds ratio (Agresti, 1980; Fujii, 2004; Kieser et al., 2013; Newcombe, 2006, part 1), although it is—statistically speaking—the odds.

The number-needed-to-treat (*NNT*) has become a fashionable measure in the working group of Evidence Based Medicine (Altman et al., 2001; Cook & Sackett, 1995; Guyatt, Rennie, Meade, & Cook, 2008). This measure denotes the number of patients to be treated to find an average of one more success (or failure) patient compared to the reference group. Now the simple risk difference in a 2×2 table is equal to the Mann–Whitney difference, $MWD = P(Y < X) - P(Y > X) = 2MW - 1$. This feature has led some researchers to assume that it is correct to generalize the Mann–Whitney difference (*MWD*) to ordinal and continuous data (Kraemer 2006; Kraemer & Kupfer, 2006; Kraemer et al., 2003), but as shown by Zimmermann and Rahlfs (2014), this definition gives rather unreasonable results for continuous data (see also the hint in Section 2). If *NNT* based on risk difference is to be used at all (for objections to the *RD*, see Skipka et al., 2016), we propose using $RD_{\text{average}} = MW - 0.5$ or the RD_{extreme} derived from the parameter of the distribution. Concerning the often used procedure of dichotomization for obtaining a 2×2 table, Senn and Julious (2009, p. 3204) remark: "...it is totally unacceptable to create dichotomies purely in order to be able to calculate NNTs." It is interesting that Edwardes and Baltzan (2000) recommended some generalizations of 2×2 table measures. Their generalized OR_G , however, is Agresti's alpha and their RD_G is the method of Kraemer (2006). Both are now misleading.

The relative risk (*RR*) does not fit in this scheme: Contrary to other measures that are global the *RR* is strictly local, meaning that it is dependent on a specified rate in the reference group, called control event rate (*CER*).

Indeed, there are inherent disadvantages for *RR*. Because *RR* is not symmetric about the center of the distribution, whether *RR* results are based on the event or the counter-event (e.g. death or survival), makes a considerable difference. The relative risk can never increase beyond a certain upper limit (see also Skipka et al., 2016).

For death or survival, the category death is usually chosen because the *RR* in most cases is smaller than a predefined value that is less than 1.0. If, however, *RR* is used at all, we recommend calculating *RR* for both categories, for example for death and

survival. It could be that one side shows a relevant RR whereas the other side does not. In the following, we give a simple data example for the 2×2 table and assume that $RR < 0.5$ is a value of major relevance:

	Death	Survival
TER treatment	0.6	0.4
CER control	0.9	0.1

For death, we have $RR = 0.67$ so that the result is not relevant ($RR > 0.5$). For survival, we have $RR = 4.0$, or taking the equivalent reciprocal value, we have $RR = 0.25$ that is relevant ($RR < 0.5$). A decision with respect to the two sides need not be made for the odds ratio.

There also are difficulties with the risk difference measure. The IQWiG method manual (2015, p. 191) remarks that the risk difference is highly dependent on the risk of the control group and therefore cannot be a useful effect size measure. This statement is, indeed, true as can be seen by looking at Table 6 in Guyatt et al. (2013) that gives RD s derived from a single δ based on “control group response rate” (reexpression formula of Furukawa, 1999). Therefore, the *Cochrane Collaboration Handbook* recommends always presenting a variety of NNT values based on a different “assumed control risk,” which is, of course, awkward. Contrary to that there are some well-defined global RD values in our scheme of measures, all based on the complete set of continuous data, when derived from a global δ , β , or γ , etc. There is then a reasonable final result: either expected (\sim average) or extreme RD .

Effect size measures and relationships are now appearing everywhere in the literature. We have structured these measures in such a way that the data distribution used for a specific reexpression is clear. Our classification scheme is based on continuous data, so that ordinal data and even binary data can be tackled with transfer formulas. Thus, if continuous data are available, there is no need for dichotomization associated with the well-known loss of data information. Our formulas provide standards for interpreting the relevance of all measures in the scheme. The cornerstone of the scheme is the relative effect also known as Mann–Whitney measure. This is rarely used and cited in the mainstream literature, although it has definite advantages: it is assumption-free and also efficient because it exploits all data information. We use it as a general basis for comparisons of effect size measures.

The formulas are extremely important when confronted with the task of a meta-analysis based on several different effect size measures, but they are also of interest for the data analysis in a single study. Here, we recommend starting by calculating the robust MW measure of stochastic superiority and then—for the purpose of the interpretation—transferring the result. Example: MW to RD_{average} . From this, using specific distribution assumptions, we can get their defining parameters and resulting averages and extreme values. No dichotomization necessary.

CONFLICT OF INTEREST

Both authors are members of idv – Data Analysis and Planning of Studies and receive no honoraria for this scientific work.

ORCID

Volker Rahlfs  <https://orcid.org/0000-0002-4827-0734>

REFERENCES

- Acion, L., Peterson, J. P., Temple, S., & Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25, 591–602.
- Agresti, A. (1980). Generalized odds ratios for ordinal data. *Biometrics*, 36, 59–67.
- Altman, D., Schulz, K., Moher, D., Egger, M., Davidoff, F., Elbourne, D. CONSORT GROUP. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*, 134, 663–694.
- Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, 51, 809–816.
- Bolland, K., Sooriyachchi, M. R., & Whitehead, J. (1998). Sample size review in a head injury trial with ordered categorical responses, *Statistics in Medicine*, 17, 2835–2847.
- Bordley, R. F. (2009). The hippocratic oath, effect size, and utility theory, *Medical Decision Making*, 377–379.
- Brumback, L. C., Pepe, M. S., & Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials, *Statistics in Medicine*, 25, 575–590.

- Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial designs* New York: Wiley.
- Brunner, E., & Langer, F. (1999). *Nichtparametrische analyse longitudinaler daten*. München: Oldenbourg.
- Brunner, E., & Munzel, U. (2002). *Nichtparametrische datenanalyse* Berlin: Springer.
- Cochrane Handbook for Systematic Reviews of Interventions*, version 5.1.0 (2011). Oxford: The Cochrane Collaboration. In Higgins, J. P. T., & Green, S., (Eds.). Retrieved from www.cochrane-handbook.org.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. (1977). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.: 1969, rev. New York.
- Colditz, G. A., Miller, J. N., & Mosteller, F. (1988). Measuring gain in the evaluation of medical technology: The probability of a better outcome. *International Journal of Technology Assessment in Health Care*, 4, 637–642.
- Cook, R. J., & Sackett, D. L. (1995). The number needed to treat: A clinical useful measure of treatment effect, *British Medical Journal*, 310, 452–454.
- D'Agostino, R. B., Campbell, M., & Greenhouse, J. (2006). The Mann-Whitney statistic: Continuous use and discovery (special papers for the 25th anniversary of Statistics in Medicine). *Statistics in Medicine*, 25, 541–542.
- Demidenko, E. (2016). The *p*-value you can't buy. *The American Statistician*, 70, 33–38
- Edwardes, M. D., & Baltzan, M. (2000). The generalization of the odds ratio, risk ratio and risk difference to $r \times k$ tables. *Statistics in Medicine*, 19, 1901–1914.
- Fujii, Y. (2004). Inference based on $P(X < Y/PX > Y)$ in two sample problems. *Bulletin of Informatics and Cybernetics*, 36, 137–145.
- Furukawa, T. A. (1999). From effect size into number needed to treat. *The Lancet*, 353, 1680.
- Guyatt, G. H., Rennie, D., Meade, M. O., & Cook, D. J. (2008). *User's Guides to the medical literature. A manual for evidence-based clinical practice* (2nd ed.). New York: McGraw-Hill.
- Guyatt, G. H., Thorland, K., Oxman, A. D., Walter, S. D., Patrick, D., & Furukawa, T. A. (2013). GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles—continuous outcomes. *Journal of Clinical Epidemiology*, 66, 173–183.
- IQWiG (2015). Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficacy in Health Care), *Allgemeine Methoden*, Vers. 4.2, 22.04.
- IQWiG (2017). Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficacy in Health Care), *Allgemeine Methoden*, Vers. 5.0.
- Kieser, M. (2014). Statistical methods for the assessment of clinical relevance. In K. van Montfort, J. Oud, & W. Ghidry (Eds.), *Developments in statistical evaluation of clinical trials* (pp. 195–207). Berlin, Heidelberg: Springer-Verlag.
- Kieser, M., Friede, T., & Gondan, M. (2013). Assessment of statistical significance and clinical relevance, *Statistics in Medicine*, 32, 1707–1719.
- Klotz, J. H. (1966). The Wilcoxon, ties and the computer. *Journal of the American Statistical Association*, 61, 772–787.
- Kraemer, H. C. (2006). Correlation coefficients in medical research: From product moment correlation to the odds ratio. *Statistical Methods in Medical Research*, 15, 525–545.
- Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, 42, 1524–1529.
- Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, 59, 990–996.
- LaVange, L. M., Durham, T. A., & Koch, G. G. (2005). Randomization-based nonparametric methods for the analysis of multicenter trials. *Statistical Methods in Medical Research*, 14, 281–301.
- Lu, M., & Tilley, B. C. (2001). Use of odds ratio or relative risk to measure a treatment effect in clinical trials with multiple correlated binary outcomes: Data from the NINDS t-PA stroke trial. *Statistics in Medicine*, 20, 1891–1901.
- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361–365
- McHugh, G. S., Butcher, I., Steyerberg, E. W., Marmarou, A., Lu, J., Lingsma, H. F. ... Murray G. D. (2010). A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: Results from the IMPACT* project. *Clinical Trials*, 7, 44–57.
- Munzel, U., & Hauschke, D. (2003). A nonparametric test for proving noninferiority in clinical trials with ordered categorical data. *Pharmaceutical Statistics*, 2, 31–37.
- Newcombe, R. G. (2006). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, 25, 543–557.
- Saville, B. R., LaVange, L. M., & Koch, G. G. (2011). Estimating covariate-adjusted incidence density ratios for multiple time intervals in clinical trials using nonparametric randomization-based ANCOVA. *Statistics in Biopharmaceutical Research*, 3, 242–246.
- Savitz, S. I., Lew, R., Bluhmki, E., Hacke, W., & Fisher, M. (2007). Shift analysis versus dichotomization of the modified ranking scale outcome scores in the NINDS and ECASS-II trials. *Stroke*, 38, 3205–3212.

- Schistermann, E. F., Reiser, B., & Faraggi, D. (2006). ROC analysis for markers with mass at zero. *Statistics in Medicine*, 25, 623–638.
- Senn, S., & Julious, S. (2009). Measurement in clinical trials: A neglected issue for statisticians?. *Statistics in Medicine*, 28, 3189–3209.
- Skipka, G., Wieseler, B., Kaiser, T., Thomas, S., Bender, R., Windeler, J., & Lange, S. (2016). Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biometrical Journal*, 58, 43–58.
- Tilley, B. C. (2012). Contemporary outcome measures in acute stroke research. Choice of primary outcome measure and statistical analysis of the primary outcome in acute stroke trials. *Stroke*, 43, 935–937.
- Tilley, B. C., Marler, J., Geller, N. L., Lu, M., Legler, J., Brott, T. ... Grotta, J. (1996). Use of a global test for multiple outcomes in stroke trials with applications to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. *Stroke*, 27, 2136–2141.
- Tritchler, D. (1995). Interpreting the standardized difference. *Biometrics*, 51, 351–353.
- Uryniak, T., Chan, I. S. F., Fedorov, V. V., Jiang, Q., Oppenheimer, L., Snapinn, S. M., ... Zhang, J. (2011). Responder analyses—A PhaRMA position paper. *Statistical Biopharmaceutical Research*, 3, 476–487.
- Wei, L. J., & Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, 79, 653–661.
- Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine*, 12, 2257–2271.
- Whitehead, J., Branson, M., & Todd, S. (2010). A combined score test for binary and ordinal endpoints from clinical trials. *Statistics in Medicine*, 29, 521–532.
- Wolfe, D., & Hogg, R. (1971). On constructing statistics and reporting data. *The American Statistician*, 25, 27–30.
- Zimmermann, H., & Rahlfs, V. W. (2014). Comments on number-needed-to-treat derived from ordinal scales. *Statistical Methods in Medical Research*, 23, 107–110.

How to cite this article: Rahlfs V, Zimmermann H. Effect size measures and their benchmark values for quantifying benefit or risk of medicinal products. *Biometrical Journal*. 2019;61:973–982. <https://doi.org/10.1002/bimj.201800107>