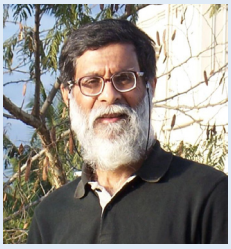
Clinical and Practical **Psychopharmacology**

Mean Difference, Standardized Mean Difference (SMD), and Their Use in Meta-Analysis:
As Simple as It Gets

Chittaranjan Andrade, MD



Each month in his online column, Dr Andrade considers theoretical and practical ideas in clinical psychopharmacology with a view to update the knowledge and skills of medical practitioners who treat patients with psychiatric conditions.

Department of Clinical Psychopharmacology and Neurotoxicology, National Institute of Mental Health and Neurosciences, Bangalore, India (*candrade@psychiatrist.com*).

ABSTRACT

In randomized controlled trials (RCTs), endpoint scores, or change scores representing the difference between endpoint and baseline, are values of interest. These values are compared between experimental and control groups, yielding a mean difference between the experimental and control groups for each outcome that is compared. When the mean difference values for a specified outcome, obtained from different RCTs, are all in the same unit (such as when they were all obtained using the same rating instrument), they can be pooled in meta-analysis to yield a summary estimate that is also known as a mean difference (MD). Because pooling of the mean difference from individual RCTs is done after weighting the values for precision, this pooled MD is also known as the weighted mean difference (WMD). Sometimes, different studies use different rating instruments to measure the same outcome; that is, the units of measurement for the outcome of interest are different across studies. In such cases, the mean differences from the different RCTs cannot be pooled. However, these mean differences can be divided by their respective standard deviations (SDs) to yield a statistic known as the standardized mean difference (SMD). The SD that is used as the divisor is usually either the pooled SD or the SD of the control group; in the former instance, the SMD is known as Cohen's *d*, and in the latter instance, as Glass' delta. SMDs of 0.2, 0.5, and 0.8 are considered small, medium, and large, respectively. SMDs can be pooled in meta-analysis because the unit is uniform across studies. This article presents and explains the different terms and concepts with the help of simple examples.

J Clin Psychiatry 2020;81(5):20f13681

To cite: Andrade C. Mean difference, standardized mean difference (SMD) and their use in meta-analysis: as simple as it gets. *J Clin Psychiatry.* 2020;81(5):20f13681.

To share: https://doi.org/10.4088/JCP.20f13681

When different authors address the same research question but obtain different results and we have to choose from among them, which result should we believe? One approach could be to accept the study that has the most credible methodology. What if many or all of the studies are methodologically sound? Averaging or pooling results across studies, using meta-analysis, has become an important option to consider, here.

Pooling Means

Meta-analysis, explained in the simplest possible way, is a mathematical procedure that averages results across several similar studies. The average value that is obtained is known as a summary estimate or a pooled estimate. As an example, imagine that 5 different studies presented weight gain data after 12 weeks of treatment with clozapine. In these 5 studies, the mean weight gain at 12 weeks was 3 kg, 3 kg, 4 kg, 5 kg, and 5 kg. If we average 3, 3, 4, 5, and 5, we get 4 as our pooled estimate. We conclude that, on average, patients gain 4 kg after 12 weeks of treatment with clozapine.

Meta-analysis averages data in much the same way, except that studies with more precise values for the mean are given greater weightage in the averaging process.

> Here, precision is determined by the standard error of the mean (SEM).[1] The SEM is smaller when the standard deviation (SD) is smaller and when the sample size is larger; a smaller SEM indicates greater precision of the mean.
>
> This is easily understood. When the SD is large, it is because individual patient values are scattered widely, implying a wider margin of error, and so the sample mean may not accurately represent the population mean. When the sample size is small, there is a greater likelihood that the sample is not a good representation of the population, again making the value of the mean dubious.
>
> Many reference sources state that precision and hence weighting is determined by how precise (narrow) the 95% confidence interval (CI) around the mean is. This is also a correct explanation because the CI is derived from the SEM.

In brief, when the sample size is large and when the SD associated with the mean is small, the mean is expected to be more precise and is assigned a higher weight when averaging results across studies in meta-analysis. Because the SEM (or the 95% CI) is a measure of variance, this method of determining weights is known as the inverse variance method. How weighting is done is beyond the scope of this article; interested readers may refer to Deeks et al.[2]

Pooling Means That Are Expressed in Different Units

How do we pool means when some studies present the 12-week clozapine outcomes as mean increase in the body mass index (BMI) and other studies present the data as mean increase in body weight?

As an example, imagine that the mean (SD) [M (SD)] gain values in 5 studies were 1.2 (0.6) kg/m^2, 1.5 (0.9) kg/m^2, 3.5 (2.5) kg, 4.0 (2.0) kg, and 5.5 (3.3) kg. These means cannot be averaged because some are presented in the unit of BMI (kg/m^2) and some are presented in the unit of weight (kg). There is, however, a way out.

> In the first study, the mean increase was 1.2 and the SD was 0.6. This means that clozapine increased the mean BMI by 2 SD. In the second study, clozapine increased the mean BMI by 1.5/0.9, or 1.7 SD. In the third study, clozapine increased mean body weight by 3.5/2.5, or 1.4 SD. The values for the last 2 studies are 4.0/2.0 and 5.5/3.3; that is, 2 SD and 1.7 SD.

If we divide each mean by its SD, we discover by how many SD BMI or weight have increased in the average patient who has received clozapine for 12 weeks. Because increase in BMI or weight is now expressed in the uniform unit of SDs, averaging of outcomes, with weights assigned (as explained in the previous section), can validly be performed.

> We found that, in the 5 studies, clozapine increased measures of body weight by 2.0, 1.7, 1.4, 2.0, and 1.7 SD. The average of these 5 numbers is 1.76. We can conclude that clozapine increases measures of weight by an average of 1.76 SD. Note that this calculation is for explanation, only; weights have not been assigned to each value, as would have been done in meta-analysis.

Mean Difference

Imagine now that, instead of examining the increase in weight in a single group of clozapine-treated patients, we examine the increase in weight in patients who are randomized to receive clozapine or haloperidol.

> The words *increase* and *gain* are used for convenience. Patients may gain weight, lose weight, or show no change in weight. Change in weight can therefore be positive, negative, or zero. Averaging can be performed regardless of the direction of weight change as long as the sign (positive or negative) is retained while averaging.

After 12 weeks of treatment, we calculate the increase in weight in each patient in each of the 2 groups. We next calculate the M (SD) weight gain in the clozapine group and the M (SD) weight gain in the haloperidol group. When we subtract the haloperidol mean from the clozapine mean, we learn by how much weight gain in the average clozapine patient exceeds weight gain in the average haloperidol patient. This value is the mean difference for this study.

Imagine, now, that there are several randomized clinical trials (RCTs) that examine weight gain with clozapine as compared with other antipsychotic drugs.

> It would be nice if the comparator antipsychotic is the same drug in all the RCTs; however, this is not essential. If the comparator is the same, we draw conclusions about clozapine versus this comparator. If the comparator is a different antipsychotic drug in different RCTs, we draw conclusions about clozapine versus "other antipsychotics" rather than versus a specific antipsychotic.

We get a mean difference value for clozapine versus comparator antipsychotic in each RCT. In meta-analysis, we can pool the values for mean difference across RCTs in exactly the same way that was described for the pooling of mean values in an earlier section of this article. This pooled estimate is also known as mean difference and is abbreviated as MD.

Because the mean difference in different RCTs would have been associated with different SDs and with different sample sizes (and hence different SEMs), different weights would need to be assigned to each RCT when the mean differences are pooled in meta-analysis. So the pooled MD is more accurately described as a weighted mean difference or WMD. That is, MD and WMD mean the same, but only when they refer to a pooled estimate in meta-analysis.[2]

> Weights are applied to relative risks (RRs), odds ratios (ORs), and other statistics, and not to means and mean differences, alone, when these are pooled in meta-analysis. However, we do not say "weighted RR" or "weighted OR" when we speak of the RR or OR as pooled estimates. Therefore, there does not seem to be much logic in preferring the term WMD over MD when describing a weighted pooled value for mean difference.[2] Most meta-analyses that present pooled mean differences therefore use MD as a descriptor, though the use of WMD is not uncommon.

Interpreting the Mean Difference

Rather obviously, a mean difference value of 0 means that there is no difference between the experimental and control groups. A positive value means that the experimental group is associated with an increase in the value of outcome, relative to the control group, and a negative value means that the experimental group is associated with a decrease in the value of the outcome.

The mean difference is usually expressed along with a 95% CI. If the entire 95% CI lies above 0, it means that there is a statistically significant increase in the value of the outcome. If the 95% CI includes 0, it means that there is no significant difference in outcomes between the groups being compared. If the entire 95% CI lies below 0, it means that there is a significant decrease in the value of the outcome.[3]

As imaginary examples to illustrate the point, an RCT found that clozapine was associated with a significant increase in weight relative to haloperidol (mean difference, 4.0 kg; 95% CI, 2.5–5.5 kg); perphenazine did not differ significantly from haloperidol with regard to weight gain (mean difference, 0.5 kg; 95% CI, −1.0 to 2.0 kg); and lurasidone was associated with significantly less weight gain than haloperidol (mean difference, −1.0 kg; 95% CI, −1.9 to −0.1 kg).

Standardized Mean Difference

What happens when some clozapine versus haloperidol RCTs present change as weight gain (kg) and others present the data as increase in BMI (kg/m^2)? We are again faced with having to combine data that are presented in different units. The solution, as described earlier, is to divide the

mean difference by the SD in each RCT. Thus, for each RCT we obtain a value that is known as the standardized mean difference (SMD); that is, the mean difference expressed in units of SD. The SMD for each RCT can now be pooled, with weights assigned to each SMD (as described earlier). The pooled estimate is also known as an SMD.

There is a little problem here. Let's look at imaginary data from a single RCT. Between study baseline and endpoint, the M (SD) weight gain was 5.0 (2.5) kg with clozapine and 2.0 (1.5) kg with haloperidol. Finding the mean difference is easy; 5 – 2 = 3, so the average patient gained 3 kg more in the clozapine arm than in the haloperidol arm of the RCT. Now, there are 2 SDs. One is the SD in the clozapine arm and the other is the SD in the haloperidol arm. Which SD do we choose when converting the mean difference (3 kg) into an SMD?

One solution is to use the pooled SD. In the example above, we need to pool 2.5 and 1.5. This is easily done using a simple formula.

> The formula is not provided here because it is not necessary to know. Statistical programs do the work for us. For those who do want to know, an online search will provide several answers, such as the formula for pooled SD when the sample sizes are equal in the 2 groups, the formula when sample sizes are unequal, and the formula when sample sizes are small (eg, < 20).

The mean difference divided by the pooled SD gives us an SMD that is known as Cohen's *d*. Because Cohen's *d* tends to overestimate the true effect size, especially when the sample size is small (< 20), a correction factor is applied, and this value for the SMD is known as Hedges' *g*. However, when the mean difference is divided not by the pooled SD but by the SD of the control group, the SMD is known as Glass' delta.[4]

The reader will now understand how, in meta-analysis, mean differences between treatments (eg, drug vs placebo) can be combined even when different studies assess outcomes using different rating scales. As with means and mean differences, SMDs from individual studies are weighted before they are pooled in meta-analysis.

As a final note in this section, in most scales that rate illness severity, lower scores indicate less severe illness. What if such scales are being pooled (using SMDs) in meta-analysis with 1 or more scales in which higher scores indicate better functioning and hence less severe illness? The simplest approach is to multiply the latter values by −1 so that all SMDs uniformly indicate that lower values indicate less severe illness.[2]

Which SMD Should We Use and When?

When calculating the SMD, the numerator is always the difference between means. Depending on what we use as the denominator, and depending on our use of a correction factor, there are 3 SMDs: Cohen's *d*, Hedges' *g*, and Glass' delta. Cohen's *d* is the SMD that is most often reported.[5] Hedges' *g* is usually reported in Cochrane reviews.[2] Glass' delta may be preferred when the intervention changes the SD in addition to the mean[5] and/or when there is considerable difference between the SDs of the 2 groups.[4]

A disadvantage of Glass' delta is that, because it uses the SD of only the control group, it is based on a smaller sample size and may hence be less accurate.[5] An advantage is that it tells us by how much the experimental group changes in terms of the control group SD; so the control group SD is the yardstick for measuring change. Another advantage of Glass' delta is that it may be more appropriate for use when there are several experimental groups and 1 control group. Note, however, that it is possible to calculate a pooled SD for > 2 groups.

Readers who are interested in other, less common, versions of the SMD may refer to Lakens.[4] The SMD can also be used to compare the mean difference in scores between 2 time points (eg, before vs after an intervention) in a single group[4]; this, however, is an uncommon circumstance.

SMD and Effect Size

How large is a treatment effect? There are different ways of assessing this. For example, with continuous data, such as depression rating scores, we can express the difference between experimental and control groups as a mean difference or as an SMD. With categorical data, such as treatment response or illness remission status, we can express the difference between groups in terms of percentages, RRs, or ORs. Mean difference, percentage difference, SMD, RR, and OR are all measures of effect size. For that matter, statistics such as the number needed to treat, the number needed to harm, the likelihood of being helped or harmed, and others, are also measures of effect size.

However, because Cohen's *d* was proposed as the original measure of effect size, when authors write ES (effect size), they almost always mean SMD or, by default, Cohen's *d*. SMD is the preferred term.[2] Even better, authors can state Cohen's *d*, or whatever version of the SMD they have used, instead of ES.

Interpreting the SMD

Cohen[6] suggested that *d* values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes (readers may now understand how Cohen's *d* became equated with ES). If 2 populations are normally distributed and if they are equal in size and variability, then, when $d = 0.2$, there is about 85% overlap between the distributions; so it can be hard to differentiate between the groups. When $d = 0.5$, the overlap shrinks to about 67%, and the difference between groups is fairly obvious to the eye. When $d = 0.8$, the overlap is only about 53%, and the difference between groups is very obvious.[6]

This guidance is now almost set in stone. SMD values of 0.2–0.5 are considered small, values of 0.5–0.8 are considered medium, and values > 0.8 are considered large. In psychopharmacology studies that compare independent groups, SMDs that are statistically significant are almost always in the small to medium range. It is rare for large SMDs to be obtained.

Of note, an SMD of 0 means that there is no difference between groups, and an SMD that is negative means that the experimental group has a lower mean score than the control group (this is when the numerator for SMD is calculated as experimental minus control and the negative sign is retained). If the 95% CI for the SMD includes 0, the SMD is not "statistically significant." As an example, if the SMD for a trial of weight gain with experimental drug versus haloperidol is 0.30 (95% CI, −0.70 to 1.30), there is no significant difference in weight gain between drug and haloperidol groups. If the SMD is 0.30 (95% CI, 0.05 to 0.55), the drug is associated with significant increase in weight relative to haloperidol (by a mean of 0.3 SD). If the SMD is −0.40 (95% CI, −0.10 to −0.70), the drug is associated with significant weight loss relative to haloperidol (by a mean of 0.4 SD).

Which Is Better, Mean Difference or SMD?

Mean difference and SMD are both important. The mean difference provides information in clinical units, and the SMD provides information in statistical units. Thus, for example, if we learn that, after 12 weeks of treatment, mean body weight increases by 4 kg (relative to haloperidol) in clozapine-treated patients, we are concerned because we know that 4 kg is a lot of weight to gain. If we are told that the corresponding SMD is 1.0 (Cohen's $d = 1.0$), we are likewise concerned, because this means that the curve representing the distribution of body weight has been shifted to the right by 1 whole standard deviation.

Sometimes, when SDs are small, the pooled SD is also small, and the corresponding SMD can be medium to large even when the absolute difference between groups (mean difference) is small. This is when the mean difference scores over the SMD. It goes without saying that knowing that clozapine increases weight by 4 kg is more clinically meaningful than knowing that clozapine increases weight by 1 SD.

We are all familiar with the real-life implications of an increase in body weight by a certain value, expressed in kg. Many of us would likewise be familiar with the clinical implications of a change by a certain value in common anxiety, depression, or psychosis rating scales. What if the reader is a novice who is unfamiliar with rating scales? What if the rating scale is an unfamiliar or new instrument? In such situations, the SMD, classified as small, medium, or large, is necessary for understanding the magnitude and importance of the treatment effect in an RCT.

Finally, if the sample size in a study is sufficiently large, even a small and clinically unimportant difference between groups can be statistically significant. This is seen as a small value for the mean difference between groups or as an SMD that is < 0.2, or as both. Having said so, there could be studies that are important for life and limb where even a small difference is better than no difference, and so all statistically significant SMDs that are < 0.2 should not necessarily be rejected as inconsequent.

Summary

When different studies present outcomes in the same units, or using the same rating instruments, mean differences between experimental and control groups can be directly pooled in meta-analysis. When the outcomes are presented in different units, or when different rating instruments are used in different studies, mean differences need to be converted into SMDs before they can be pooled in meta-analysis.

Published online: September 22, 2020.

REFERENCES

1. Andrade C. Understanding the difference between standard deviation and standard error of the mean, and knowing when to use which. *Indian J Psychol Med*. 2020;42(4):409–410.
2. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, West Sussex, England: Wiley-Blackwell; 2008:243–296.
3. Andrade C. A primer on confidence intervals in psychopharmacology. *J Clin Psychiatry*. 2015;76(2):e228–e231.
4. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol*. 2013;4:863.
5. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. 4th ed. Shelton, CT: People's Medical Publishing House; 2014.
6. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.