Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An algorithmic process to develop clinical practice guideline recommendations based on meta-analyses of the relevant evidence

Pierce Boyne

boynepe@ucmail.uc.edu

University of Cincinnati

Emily Fox

University of Florida

Dorian Rose

University of Florida

Michael Lewek

University of North Carolina at Chapel Hill

Christina Garrity

University of Cincinnati

Daria Pressler

University of Cincinnati

Julie Braun

University of North Florida

Steven Walczak

Brooks Rehabilitation

Jolene Foster

University of Cincinnati

Mark Bowden

Brooks Rehabilitation

James Lynskey

Creighton University

Method Article

Keywords: guideline development group, meta-analysis, evidence synthesis, evidence to recommendation

Posted Date: July 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10371300/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The development of clinical practice guidelines (CPGs) is a major initiative in the physical therapy profession to improve quality of care and reduce unwarranted variation in practice. When developing any CPG, a fundamental step is determining how to synthesize the evidence that addresses each clinical question. While meta-analysis is the strongest approach to evidence synthesis, few prior physical therapy CPGs have included meta-analysis and none (to our knowledge) have used the results of the meta-analysis to guide recommendations. In addition, current evidence-to-recommendation frameworks require strong subjective judgments, which may vary based on the opinions of the specific guideline development group members, making the guidelines less reproducible and subject to bias. This report from a CPG development team describes an algorithmic decision-making process for developing CPG recommendations based on meta-analyses of the relevant evidence. We also provide resources to facilitate this process and discuss the strengths and limitations of the proposed methodologies. These detailed methods and resources are intended to enable other CPG development teams to capitalize on the benefits of meta-analysis when developing recommendations, to reduce the need for subjective judgments and to strengthen reproducibility.

Introduction

A major initiative of the physical therapy profession is directed towards development and implementation of Clinical Practice Guidelines (CPGs). CPGs are intended to assist therapists with selection of treatment strategies that maximize outcomes, reduce unwarranted variability in care, and increase the value of physical therapy services. In the context of rising health-care costs, an aging population and increased survival rates among individuals with health impairments, these CPGs should provide recommendations about interventions to maximize function for specific tasks, which will improve the efficiency of rehabilitation services.

A fundamental step in developing any CPG is determining how to synthesize the evidence that addresses each clinical question. The three evidence synthesis methods considered to be acceptable are: (1) meta-analysis; (2) summarizing the range and distribution of effect sizes (e.g. treatment vs. control differences) across studies; and (3) counting the proportion of studies favoring the treatment vs control (without regard to statistical significance).[1] According to the Cochrane Handbook for Systematic Reviews, conventional methods that count the proportion of studies with statistically significant effects (or other vote counting rules based on statistical significance) are considered “unacceptable”, because they “have serious limitations and can lead to the wrong conclusion”.[1] When feasible, meta-analysis is the preferred approach to evidence synthesis because it: (1) provides an overall effect estimate and confidence interval for the body of evidence as a whole; (2) improves precision and statistical power by pooling numerical results across studies; (3) provides estimates of effect consistency across studies; and (4) can overcome issues arising from apparently conflicting studies.[1]

This information from meta-analysis could also be used to reduce the subjective judgment currently needed for determining CPG recommendations,[2,3] thus increasing reproducibility and making the process less dependent on the opinions of Guideline Development Group (GDG) members. For example, the American Physical Therapy Association (APTA) CPG manual suggests that a strong recommendation for an intervention should be made when there is a "high level of certainty of moderate to substantial benefit … based on a preponderance of Level 1 or 2 evidence …".[2] Instead of relying on expert opinion for subjective decisions about what constitutes a "moderate to substantial" benefit, the overall effect estimate (and confidence interval) from meta-analysis could be compared with pre-specified minimally important difference (MID) thresholds for more objective determinations. These overall effect estimates average together all the available evidence in an equitable manner (e.g. with studies weighted proportional to their sample size). Thus, subjective decisions about whether the "preponderance" of evidence favors an intervention could be minimized. Finally, subjective decisions about what constitutes a "high" level of certainty could be facilitated by using meta-analysis results to derive an evidence certainty rating using the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) system.[4]

Despite these potential benefits, few prior physical therapy CPGs have included meta-analysis[5] and none (to our knowledge) have used the output from meta-analysis to determine the recommendations. One likely contributor to this limited use of meta-analysis in physical therapy CPGs is that there is little to no available guidance on how meta-analysis output can be incorporated into recommendation decisions. Thus, this report from a GDG aims to describe an approach to develop recommendations based on meta-analyses of the relevant evidence. We also provide resources to facilitate this process.

CPG development involves numerous steps, e.g.: GDG identification, obtaining input from key interested parties, a literature search, appraisal of study quality, grouping similar studies for synthesis, study data extraction, synthesizing results across studies, developing recommendations based on the results synthesis, obtaining feedback from key interested parties, and disseminating recommendations. This report focuses on the results synthesis and its use to guide recommendation decisions. However, we begin by providing some information about earlier CPG development steps to provide relevant context.

Summary of the CPG providing context for this report

The CPG that prompted the methodological developments described in this report is focused on interventions to improve walking ability in patients less than six months post stroke, which is the timeframe when most usual care rehabilitation is delivered. To be considered for evidence synthesis, studies had to be randomized controlled trials in this population testing the effects of a physical therapy intervention on one or more walking outcomes. With input from individuals with stroke, we prespecified the walking outcomes of interest and grouped them by the construct assessed by each outcome (Table 1). The literature search and screening identified 296 eligible studies for quality appraisal.

Table 1
Outcome constructs and measures of interest

Construct	Measures representing that construct
Walking capacity	Self-selected gait speed (e.g. 10-meter walk test)
	Fast gait speed (e.g. 10-meter walk test)
	Timed distance walked (e.g. 6-minute walk test)
Walking independence	Functional ambulation category
	Functional independence measure locomotion item
	Proportion of participants who newly achieved or increased walking independence
Functional mobility	Timed up and go
Daily walking activity	Steps per day

Study appraisal methodology

Each eligible study was appraised by two-person teams of physical therapists, among a pool of 10 total appraisers who were trained and standardized on the appraisal process. Discrepancies between individual therapist appraisal ratings were resolved by discussion within the two-person appraisal team, with involvement of other GDG members in cases of uncertainty. Appraisals were performed using version 2 of the Cochrane risk-of-bias tool for randomized trials (RoB-2, available at: https://www.riskofbias.info/).[6] The RoB-2 evaluates the following bias domains, which are known to affect the results of randomized trials:[1,6]

- Domain 1: Risk of bias arising from the randomization process
- Domain 2: Risk of bias due to deviations from the intended interventions
- Domain 3: Missing outcome data
- Domain 4: Risk of bias in measurement of the outcome
- Domain 5: Risk of bias in selection of the reported result

Each domain consists of a series of questions requiring a response of yes, probably yes, no information, probably no, or no. These responses are used with a decision algorithm to determine whether there is “low risk”, “some concerns”, or “high risk” of bias for that domain.[6] This is done for each study outcome within a given study.

Using the published RoB-2 document and additional resources available at https://www.riskofbias.info/, we developed an accompanying spreadsheet to further operationalize the RoB-2 criteria for appraisal of the acute/subacute neurologic rehabilitation literature (available at: https://osf.io/m6u4w/files/osfstorage), which may be useful for other GDGs. This spreadsheet also

facilitated standardization and quantification of the study appraisal process, simplified decision making by collapsing response options that led to the same algorithm decision for a given question, and increased efficiency by automating the decision algorithm. We also calculated an overall risk of bias for each study outcome as follows:

- Low risk: ≥ 3 domains at low risk and no domains at high risk
- Some concerns: ≥ 1 domain at low risk and ≤ 1 domain at high risk
- High risk: 0 domains at low risk or 2 domains at high risk
- Very high risk: ≥ 3 domains at high risk

Study outcomes judged to be at very high risk of overall bias were excluded from the evidence synthesis. Overall risk of bias for a given study was taken as the best overall risk of bias across all study outcomes.

Grouping Evidence for Synthesis

When synthesizing evidence, it is important to only group studies together when they have similar intervention and control conditions. For example, studies comparing two active interventions targeting the same outcome are likely to have smaller between-group differences in that outcome compared with studies comparing an active intervention with a no-intervention or sham control. In our case, the 296 eligible and appraised studies tested a wide variety of interventions and used a variety of control conditions.

Instead of pre-specifying specific intervention vs. control comparisons to assess (e.g. gait training with vs. without body weight support), we used a data-driven approach that aimed to identify all available randomized comparisons with sufficient evidence in the literature. With input from the full GDG, two members (ML and MB) identified clusters of studies with intervention and control arms that were similar enough for evidence synthesis. This process necessitated consideration of critical components that define the intervention (e.g. a heart rate or perceived exertion target in the aerobic training zone for cardiovascular endurance exercise) for each intervention reported in the literature. These critical components were then used to identify studies that had experimental arms containing the components and control arms that did not.

When there were potentially important methodological differences (e.g. differing intervention or control treatment parameters or participant characteristics) between studies making the same comparison, subgroup analyses were considered within that subset of studies with similar characteristics. For example, when comparing cardiovascular endurance training to non-cardiovascular interventions, we included a subgroup analysis of studies that used walking vs. cycling as the mode of endurance training. To mitigate risks of false positives from multiple comparisons,[1,7] 0–3 subgroupings of particular interest were prespecified for each intervention vs. control comparison.

Meta-analysis Methodology

Overview

For each intervention vs. control comparison meeting a minimum threshold of evidence (at least two RCTs, including at least one with low risk of overall bias or two with some concerns of overall bias), meta-analysis was performed to quantitatively pool the results across studies. Within each intervention vs. control comparison, separate meta-analyses were done for each outcome *measure* (e.g. self-selected gait speed) and *construct* (e.g. walking capacity), while applying the same minimum evidence threshold above, and excluding outcomes judged to be at 'very high' risk of overall bias. For the prespecified subgroup analyses, this process was repeated within each subgrouping of studies.

Standardized effect sizes

For outcome *constructs* (Table 1), meta-analyses were performed in standard deviation units (i.e. standardized effect sizes), so that they could be pooled across outcome measures that used different units to represent the same construct. Hedges' *g* was used for standardized effect size calculation, with small sample adjustment, as recommended in the Cochrane Handbook.[1] Pooling across measures was done within each study. For example, the Hedges' *g* for the walking capacity construct for a given study was the average of the Hedges' *g* values for self-selected gait speed, fastest gait speed and/or 6-minute walk distance from that study (whichever of those outcomes were reported).

Unstandardized effect sizes

Meta-analyses for each outcome *measure* were performed in the original units of the measure (e.g. m/s for gait speed), to allow inferences regarding clinical meaning relative to MID thresholds. Since *between-group* MID thresholds (i.e. meaningful superiority margins) have not yet been established for the outcomes of interest, operational MID thresholds were selected for each outcome (Table 2) using *within-group* minimal clinically important differences (MCIDs) from study samples that were the closest to the target population to this CPG. When no published MCID thresholds were available for an outcome measure, we defaulted to the published minimal detectable change. When that was not available, the operational MID was determined via consensus of the GDG.

Table 2
Operational between-group MID values for this CPG

Outcome	Threshold	Reference
Self-selected and fastest gait speed	0.16 m/s	MCID from Tilson 2010[8]
6-minute walk distance	34 meters	MCID from Fulk 2018[9]
Functional ambulation category	0.5 points	GDG consensus
Functional independence measure locomotion item	1.3 points	MCID from Beninato 2006[10]
Probability of increased independence	25%	'GRADE' handbook[11]
Timed up and go	2.9 seconds	MDC from Flansbjer 2005[12]
Daily walking activity	1,828 steps/day	MCID from Hiragami 2024[13]

Abbreviations: MID, minimally important difference; MCID, minimal clinically important difference; GDG, guideline development group; MDC, minimal detectable change

Results extraction

Results were extracted from individual studies by a physical therapist with PhD training in epidemiology and biostatistics (PB) and two physical therapists who had completed PhD coursework in biostatistics (CG and DP). We developed a spreadsheet and R[14] scripts to facilitate the extraction process, automate the calculations, and facilitate quality control checks by calculating variability among values calculated from different estimates. These resources are provided at https://osf.io/m6u4w/files/osfstorage. Extracted results included characteristics of trial participants (e.g. mean stroke chronicity) and statistical outcomes. Outcome data extraction used the methods from the Cochrane Handbook,[1] similar to previous reports by one of our team members.[15,16] In studies that reported multiple time points, we used the first post-intervention time point.

For continuous outcomes, we prioritized extraction of within-group mean changes when available and otherwise extracted estimates of the between-group difference in mean change. When standard errors were not reported, they were calculated from confidence intervals, T statistics, F statistics, p values, or pre and post-intervention standard deviations (while conservatively assuming a repeated measures correlation of 0.60).[1,15,16] When only medians and interquartile ranges were reported, the approximate means and standard deviations were calculated using the formulas from Wan et al,[17] as recommended in the Cochrane Handbook.[1] Log transformed results were approximately back transformed using method 3 from Higgins et al.[18] When walking speed was reported in seconds rather than meters/second, means were approximately converted by dividing the test distance by the mean time. Approximate standard errors were calculated by dividing the new mean change/difference by the T-statistic from the

report (which was calculated from p-values if not directly reported). Standard deviations (SD) were approximately converted by assuming constant relative error, i.e.:

$$SD\left(speed_{m/s}\right) = \frac{SD\left(speed_{sec}\right)}{mean\left(speed_{sec}\right)} \times \ mean\left(speed_{m/s}\right)$$

Since the 6-minute walk test was the most common timed distance measure, other similar measures using different walking durations were converted to an approximate 6-minute walk distance. To do this for the 2-minute walk test, we used a published formula that was found to predict 98% of the variance in 6-minute walk distance in subacute and chronic stroke[19]:

$$6\,min\,walk\,distance\ (meters)\ = 2.79\ \times\ \ (2\,min\,walk\,distance) + 1.56$$

For other test durations, we interpolated the above formula, by assuming that: (1) the multiplier scales linearly from 93% of the relative test duration for the 2-minute walk test (2.79 / [6-minutes / 2-minutes] = 93%) to 100% of the relative test duration for the 6-minute walk test (1 / [6-minutes / 6-minutes] = 100%); and (2) the additive term scales linearly from 1.56 for the 2-minute walk test to 0 for the 6-minute walk test. For example, this yields the following formulas for 3-minute and 5-minute walk tests:

$$6\,min\,walk\,distance\ (meters)\ = 1.895\ \times\ \ (3\,min\,walk\,distance) + 1.17$$

$$6\,min\,walk\,distance\ (meters)\ = 1.179\ \times\ \ (5\,min\,walk\,distance) + 0.39$$

When study participants were non-ambulatory (i.e. requiring continuous physical assist from another person to walk), they were not testable for gait speed/distance or the Timed Up and Go (TUG). When extracting gait speed or distance outcomes in this situation, we prioritized extracting results that counted non-ambulatory participants at a given time point as having zero gait speed/distance at that time point (i.e. the worst possible value) to retain all participants in the analysis. However, this was not possible for the TUG since the worst value for that outcome is infinity seconds. When this preferred method was not possible, we attempted to extract results that were calculated entirely among participants who were ambulatory at baseline. Otherwise, the result was determined to be unusable.

For the dichotomous outcome (increased walking independence [yes/no]), we extracted the number of participants who achieved the outcome and the total number of participants analyzed who had not already achieved the outcome at baseline. These data were used to calculate the relative probability (a.k.a. risk ratio or probability ratio) of increased independence in the intervention versus control group.

When multiple intervention or control groups in the same trial were eligible for analysis, we pooled data from those groups together with exact averaging weighted by group sample size (for continuous outcomes).[1] For crossover studies, we extracted data from the initial parallel phase before any crossover. When intent-to-treat results were not reported, missing outcome data were imputed using control group means and standard deviations,[20] as previously described.[15]

Statistical methods

Random-effects meta-analysis was performed with R[14] and the 'meta' package,[21] using restricted maximum likelihood estimation. Hartung-Knapp correction was used for continuous outcomes.[22,23] For the dichotomous outcome, our analysis used the exact Mantel-Haenszel method and treatment arm continuity correction in studies with zero cell frequencies.[21] The R scripts used to automate the meta-analyses are provided at https://osf.io/m6u4w/files/osfstorage.

To assess between-study heterogeneity, we generated forest plots and calculated I^2, random effects model estimates of the between-study standard deviation and p-values from the Cochran's Q test.[21] To assess risk of bias, we added the RoB-2 results for each study to the forest plot and used the 'robvis' package[24] to generate summary plots that pooled RoB-2 results across studies, with each study weighted by its contribution to the pooled effect estimate. To assess risk of publication bias we generated funnel plots and used Egger's test (for the continuous outcomes) or the score test (for the dichotomous outcome) when there were at least 10 studies being pooled.[1] An example of the visualizations generated for one evidence synthesis is shown in Fig. 1.

<<< Insert Fig. 1 *here >>>*

Rating the certainty of each evidence synthesis

The GRADE approach[1,11,25,26] was used to provide a certainty rating (Table 3) for each evidence synthesis, i.e. each outcome measure and outcome construct for each treatment vs. control comparison. With this approach, randomized trials begin at high certainty and can be downgraded if there are concerns about risk of bias, inconsistency of results, indirectness of evidence, imprecision, or publication bias. Using the published GRADE materials,[1,11,25,26] and our meta-analysis output we further operationalized the GRADE criteria for this CPG as shown in Table 4.

Table 3
GRADE evidence certainty ratings11

Grade	Definition
High (++++)	We are very confident that the true effect lies close to that of the estimate of the effect.
Moderate (+++)	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
Low (++)	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.
Very Low (+)	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect.

Table 4
Operational criteria for GRADE ratings

Factor	Downgrade one level if …	Downgrade two levels if …
Risk of bias	> 50% of the information in the pooled effect estimate comes from study outcomes at high RoB for one RoB domain; OR Effect sizes are systematically different between high and low RoB study outcomes in a way that disproportionately influences the pooled effect estimate	> 50% of information in the pooled effect estimate comes from study outcomes at high overall RoB; OR > 50% of the information in the pooled effect estimate coming from study outcomes at high RoB for > 1 RoB domain; OR No information in the pooled effect estimate coming from study outcomes at low RoB across all RoB domains
Inconsistency of results	> 33% of the information in the pooled effect estimate coming from study results with 95% CIs outside the pooled mean point estimate In borderline cases, I^2 >60% used as secondary criterion	> 50% of the information in the pooled effect estimate coming from study results with 95% CIs outside the pooled mean point estimate
Indirectness of evidence	Not possible for this CPG since the synthesis only included directly applicable evidence in terms of patients, interventions, comparators and outcomes	
Imprecision	Pooled effect estimate confidence interval wider than 2 x MID*	Pooled effect estimate confidence interval wider than 4 x MID*
Publication bias	Statistically significant Egger or score test with evidence of funnel plot asymmetry (only considered when there were ≥ 10 studies)	N/A

*Minimally important difference (MID) thresholds were taken from Table 2. For the standardized effect size (Hedges' g), we used a threshold of 0.4.[27]

Abbreviations: RoB, risk of bias; CI, confidence interval; CPG, clinical practice guideline; N/A, not applicable

Adverse event data extraction and analysis

Given the limited reporting of adverse events in the rehabilitation literature,[15] we only attempted to extract data on serious adverse events (SAEs), which we expected to be reported if present. An SAE is any untoward medical occurrence that results in death, is life-threatening, requires inpatient hospitalization, or results in significant disability.[28] When SAEs were reported, the proportion of participants experiencing an SAE was compared between the treatment and control groups, pooled across all studies in the synthesis. If there was an apparent between-group difference, we planned to perform a formal meta-analysis using the methods described above for dichotomous outcomes.

Evidence to Recommendation Algorithm

We developed an algorithm to come to recommendation decisions that explicitly incorporated the meta-analysis results, GRADE certainty ratings and other inputs (Table 5). This algorithm is intended to be applied for each outcome *construct*. The possible recommendation categories are similar (but not identical) to those suggested in the GRADE handbook.[11,29]

<<< Insert Table 5 *here >>>*

New recommendation category: “No recommendation for or against”

We added one new category for situations in which the evidence ruled out a meaningful difference in outcomes between the experimental and control interventions. While this situation typically results in a recommendation against the experimental intervention,[11] our algorithm instead uses the category ‘No recommendation for or against’. This was deemed necessary to better fit the context of this CPG, because: (1) most control groups were receiving usual care rehabilitation (which would be unethical to withhold in acute/subacute stroke); (2) many of the experimental interventions being assessed were also part of usual care in many studies (e.g. overground gait training, therapeutic exercise). Thus, it could be incorrect to recommend against doing any of an intervention that is currently part of usual care when all we know is that doing more of it does not meaningfully improve outcomes. We note that the GRADE handbook was recently updated with a similar recommendation category called ‘Conditional recommendation for either’, which applies when two active treatments are found to be equivalent[29] (a slightly different situation).

Table 5
Recommendation categories and criteria

Recommendation category	Criteria	Recommendation language	Implications
Strong	Sufficient certainty (GRADE ≥ **+++**) that the treatment has: • Significant benefit over control interventions for the overall outcome construct; **AND** • Significant and potentially meaningful benefit over control interventions for individual outcome(s) within the construct (95% CI excludes zero and includes or exceeds an MID)	"Clinicians *should* use [the intervention] for [the outcome construct]"	Most individuals seeking to improve the outcome construct should receive the intervention. Adherence to this recommendation could be used to assess quality of care.
Conditional	Sufficient certainty (GRADE ≥ **+++**) that the treatment has: • Significant benefit over control interventions for the overall outcome construct; **OR** • Significant and potentially meaningful benefit over control interventions for individual outcome(s) within the construct	"Clinicians *should consider* using [the intervention] for [the outcome construct]"	The intervention may be indicated, but there is less certainty. Or it may be indicated for some individuals, but not others.
No recommendation for or against	Sufficient certainty (GRADE ≥ **++**) that the treatment: • Does not have meaningful benefit over control interventions for individual outcome(s) within the construct (95% CI includes 0 and excludes an MID)	"Clinicians should prioritize available evidence-supported interventions *other than* [the intervention] for [the outcome construct]"	If other evidence-supported interventions are available, the intervention should be de-prioritized. Otherwise, it may be appropriate.

Abbreviations: CI, confidence interval; MID, minimally important difference; N/A, not applicable

Recommendation category	Criteria	Recommendation language	Implications
Against	Sufficient certainty (GRADE ≥ **++**) that the treatment has: • Significantly greater risk of serious adverse events than control interventions; **OR** • Significantly worse than control interventions for the overall outcome construct or individual outcome(s)	"Clinicians *should* not use [the intervention] for [the outcome construct]"	Most individuals should not receive the intervention.
Insufficient evidence to make recommendation	None of the above criteria met, e.g.: • All 95% CIs include zero and an MID	N/A	More research is needed to make any recommendation.

Abbreviations: CI, confidence interval; MID, minimally important difference; N/A, not applicable

Determination of "potentially meaningful benefit"

To determine whether the treatment of interest has "potentially meaningful benefit" over control interventions for a given outcome construct, we examine the pooled treatment vs. control comparisons for the individual outcomes within the construct. If at least one of the outcomes has a 95% confidence interval (CI) that exclude zero and includes or exceeds the MID for that outcome, this criterion of "potentially meaningful benefit" is met. We also considered setting a higher bar by only declaring potentially meaningful benefit if the point estimate or full 95% CI exceeds the MID. However, we decided against these options because:

1. Our operational *between-group* MID values were mostly taken from *within-group* MCID studies designed to assess *total* change within individual patients, whereas our pooled effect estimates are based on a between-group comparison, i.e. the *net additional* benefit of the treatment over the control.[30–32] Thus, insisting on a between-group difference larger than an MCID to declare meaningful benefit would create a high risk of under-recommending effective treatments.[32]
2. The pooled treatment vs. control comparisons from our random effects meta-analyses estimate the average between-group difference across studies. If there is any between-study variability in treatment effects or participant response variability (as there typically is[15]), a sizable proportion of patients can be expected to have a meaningful net benefit from the treatment as long as the average effect estimate is approaching the MID.[15]

3. The specific operational MIDs selected for the outcomes of interest (based on published MCID values) are not hard biologic thresholds, as MCIDs tend to show large variability between derivation studies and in different subsamples, with fairly modest associations with meaningful change anchors.[e.g.33–35]

Handling conflicting results between measures

For outcome constructs represented by more than one measure, it is possible for individual measures to have results that conflict with each other or with the overall results for the construct. It is also possible for criteria to be met but with insufficient GRADE certainty. Since these issues decrease confidence in the evidence, we downgrade to a lower recommendation category when this occurs. Examples of this are provided below:

- If the criteria for a strong recommendation are met, but another outcome within the construct shows sufficient certainty (GRADE ≥ ++) that the treatment does not have meaningful benefit over control interventions, we downgrade to a conditional recommendation.
- If the criteria for a conditional recommendation are met, but another outcome within the construct shows sufficient certainty (GRADE ≥ ++) that the treatment does not have meaningful benefit over control interventions, we downgrade to no recommendation for or against.
- If the criteria for a strong recommendation are met, but with low certainty (GRADE ++), we downgrade to a conditional recommendation. If the certainty is very low (GRADE +), we downgrade to insufficient evidence.
- If the criteria for a conditional recommendation are met, but with insufficient certainty (GRADE ≤ ++), we downgrade to insufficient evidence.

Discussion of strengths and limitations of this approach

This report described an algorithmic decision-making process for developing CPG recommendations based on meta-analyses of the relevant evidence, while providing resources to facilitate this process for future GDGs. A major benefit of using meta-analysis results in the recommendation decision making process is that it avoids the flawed[1] approach of counting the proportion of studies with statistically significant effects, which has been prevalent in physical therapy CPGs. Another major benefit of this algorithmic decision-making process to translate evidence into recommendations is that it reduces the need for subjective judgments, which may vary based on the opinions of the specific guideline development group members. This differs from other recent developments in evidence to decision frameworks that focus more on making these judgments explicit.[3] By taking a more algorithmic approach, we expect that our findings will be more reproducible by our team or others in the future. We also expect that any recommendation changes in future updates will be based on newly available evidence rather than changes in GDG composition or opinions.

Of course, no algorithm could be optimized for every possible scenario. The greater objectivity and reproducibility in decision making that comes from using an algorithm comes at the cost of reduced flexibility. One particular challenge to optimizing an algorithm such as this for rehabilitation CPGs is the lack of clear cut MID thresholds for meaningful between-group differences, given the limitations of MCIDs for this application.[30–32] Having more definitive thresholds for the differences between intervention and control groups for our outcomes of interest would be particularly helpful for improving the approach described in this paper. Such developments could also better enable rehabilitation GDGs to capitalize on the newer "fully contextualized" GRADE methods that focus on whether confidence intervals cross decision thresholds when defining evidence certainty and assessing imprecision.[29]

Another limitation of our proposed methods is that meta-analysis can be time intensive, and it requires biostatistical expertise for extracting and converting quantitative results and for performing the analyses. In our case, we extracted 598 total outcome results from 186 studies for 20 intervention vs. control contrasts, and performed 309 total meta-analyses across all intervention comparisons, outcomes, and subgroup analyses. This high volume of work required us to develop tools to automate some of the procedures, which greatly increased time efficiency and may be of use to others. We are disseminating our detailed methods and tools in the hopes of reducing the time needed by future GDGs who want to implement similar approaches. Other GDGs could also opt to reduce their workload by making different decisions about how many intervention vs. control comparisons need to be synthesized with meta-analysis. However, we recognize that additional funding and support may still be necessary to enable other teams to use these methods in the future.

Conclusions

This report from an APTA CPG Guideline Development Group described an algorithmic decision-making process for developing CPG recommendations based on meta-analyses of the relevant evidence. We have also provided resources to facilitate this process. We hope that these detailed methods and resources will enable other CPG development teams to capitalize on the benefits of using meta-analysis results to drive recommendations.

Declarations

ACKNOWLEDGEMENTS

Funding: This work was supported by a clinical practice guideline development grant from the American Physical Therapy Association.

Conflicts of Interests: The authors declare no conflicts of interest.

References

1. Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (2022) Cochrane Handbook for Systematic Reviews of Interventions Version 6.3. Cochrane
2. APTA Clinical Practice Guideline Process Manual, Revised. Published online 2023. Accessed April 16 (2026) https://www.apta.org/patient-care/evidence-based-practice-resources/cpgs/cpg-development/cpg-development-manual
3. Beltran J, Neumann I, Brennan S et al (2026) Introduction to the Evidence-to-Decision Frameworks. In: *The GRADE Book*. The GRADE Working Group; Accessed May 18, 2026. https://book.gradepro.org
4. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ (2008) What is quality of evidence and why is it important to clinicians? BMJ 336(7651):995–998. 10.1136/bmj.39490.551019.BE
5. Osborne JA, Botkin R, Colon-Semenza C et al (2021) Physical Therapist Management of Parkinson Disease: A Clinical Practice Guideline From the American Physical Therapy Association. Phys Ther 102(4):pzab302. 10.1093/ptj/pzab302
6. Higgins JP, Altman DG, Gotzsche PC et al (2011) The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 343:d5928. 10.1136/bmj.d5928[doi]
7. Burke JF, Sussman JB, Kent DM, Hayward RA (2015) Three simple rules to ensure reasonably credible subgroup analyses. BMJ 351:h5651. 10.1136/bmj.h5651
8. Tilson JK, Sullivan KJ, Cen SY et al (2010) Meaningful gait speed improvement during the first 60 days poststroke: minimal clinically important difference. Phys Ther 90(2):196–208. 10.2522/ptj.20090079
9. Fulk GD, He Y (2018) Minimal Clinically Important Difference of the 6-Minute Walk Test in People With Stroke. J Neurol Phys Ther 42(4):235–240. 10.1097/NPT.0000000000000236
10. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J (2006) Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. Arch Phys Med Rehabil 87(1):32–39. 10.1016/j.apmr.2005.08.130
11. Schunemann H, Brozek J, Guyatt G, Oxman A (2013) GRADE handbook. Accessed May 18, 2026. https://gdt.gradepro.org/app/handbook/handbook.html#h.rqx5gicnzcpu
12. Flansbjer UB, Holmbäck AM, Downham D, Patten C, Lexell J (2005) Reliability of Gait Performance Tests in Men and Women with Hemiparesis After Stroke. J Rehabil Med 37(2):75–82
13. Hiragami S, Yoshida K, Otsuka T, Inoue Y (2024) Minimal clinically important difference in physical activity in patients with stroke. *F1000Research*. Preprint posted online April 12. 10.12688/f1000research.149214.1
14. Team RDC R: A Language and Environment for Statistical Computing. *R Found Stat Comput Vienna Austria*. Published online 2004. http://www.R-project.org
15. Boyne P, Miller A, Kubalak O, Mink C, Reisman DS, Fulk G (2024) Moderate to Vigorous Intensity Locomotor Training After Stroke: A Systematic Review and Meta-analysis of Mean Effects and Response Variability. J Neurol Phys Ther 48(1):15–26. 10.1097/NPT.0000000000000456

16. Boyne P, Welge J, Kissela B, Dunning K (2017) Factors Influencing the Efficacy of Aerobic Exercise for Improving Fitness and Walking Capacity After Stroke: A Meta-Analysis With Meta-Regression. Arch Phys Med Rehabil 98(3):581–595 doi:S0003-9993(16)31143-1 [pii]
17. Wan X, Wang W, Liu J, Tong T (2014) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 14:135. 10.1186/1471-2288-14-135
18. Higgins JPT, White IR, Anzures-Cabrera J (2008) Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. Stat Med 27(29):6072–6092. 10.1002/sim.3427
19. Valet M, Pierchon L, Lejeune T (2023) The 2-min walk test could replace the 6-min walk test in ambulant persons with subacute or chronic stroke: a two-stage retrospective study. Int J Rehabil Res Int Z Rehabil Rev Int Rech Readaptation 46(1):41–45. 10.1097/MRR.0000000000000556
20. Ebrahim S, Akl EA, Mustafa RA et al (2013) Addressing continuous data for participants excluded from trial analysis: a guide for systematic reviewers. J Clin Epidemiol 66(9):1014–1021e1. 10.1016/j.jclinepi.2013.03.014[doi]
21. Schwarzer G meta: General Package for Meta-Analysis. R package version 5.2-0. Published online 2022. https://github.com/guido-s/meta/
22. Langan D, Higgins JPT, Jackson D et al (2019) A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods 10(1):83–98. 10.1002/jrsm.1316[doi]
23. Knapp G, Hartung J (2003) Improved tests for a random effects meta-regression with a single covariate. Stat Med 22(17):2693–2710. 10.1002/sim.1482[doi]
24. McGuinness L, Kothe E robvis: Visualize the Results of Risk-of-Bias (ROB) Assessments. Published online 2019. https://github.com/mcguinlu/robvis
25. Guyatt GH, Oxman AD, Vist GE et al (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336(7650):924–926. 10.1136/bmj.39489.470347.AD
26. Neumann I, Schunemann H (2024) *The GRADE Book*. The GRADE Working Group; Accessed May 18, 2026. https://book.gradepro.org
27. Dobkin BH (2009) Progressive Staging of Pilot Studies to Improve Phase III Trials for Motor Interventions. Neurorehabil Neural Repair 23(3):197–206. 10.1177/1545968309331863
28. Institute NC Common Terminology Criteria for Adverse Events (CTCAE) Version 4.0. Published online 2009. http://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_8.5x11.pdf
29. Neumann I, Akl E, Meerpohl J, Working Group (2026) GRADE Recommendations. In: *The GRADE Book*. The GRADE ; Accessed May 18, 2026. https://book.gradepro.org
30. Peacock JL, Lo J, Rees JR, Sauzet O (2021) Minimal clinically important difference in means in vulnerable populations: challenges and solutions. BMJ Open 11(11):e052338. 10.1136/bmjopen-2021-052338

31. Ganju J (2025) Why Minimal Clinically Important Difference for Interpreting the Magnitude of the Treatment Effect Is Not Useful. Pharm Stat 24(3):e70015. 10.1002/pst.70015
32. Cook C (2026) Using MCID for Between–Group Comparisons Is a Methodological Mistake – Duke Center of Excellence in Manual and Manipulative Therapy. February 25, Accessed May 24, 2026. https://sites.duke.edu/cemmt/2026/02/25/using-mcid-for-between-group-comparisons-is-a-methodological-mistake/
33. Cook CE (2008) Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. J Man Manip Ther 16(4):E82–E83. 10.1179/jmt.2008.16.4.82E
34. Bohannon RW, Crouch R (2017) Minimal clinically important difference for change in 6-minute walk test distance of adults with pathology: a systematic review. J Eval Clin Pract 23(2):377–381. 10.1111/jep.12629
35. Bohannon RW, Glenney SS (2014) Minimal clinically important difference for change in comfortable gait speed of adults with pathology: a systematic review. J Eval Clin Pract 20(4):295–300. 10.1111/jep.12158[doi]

Figures

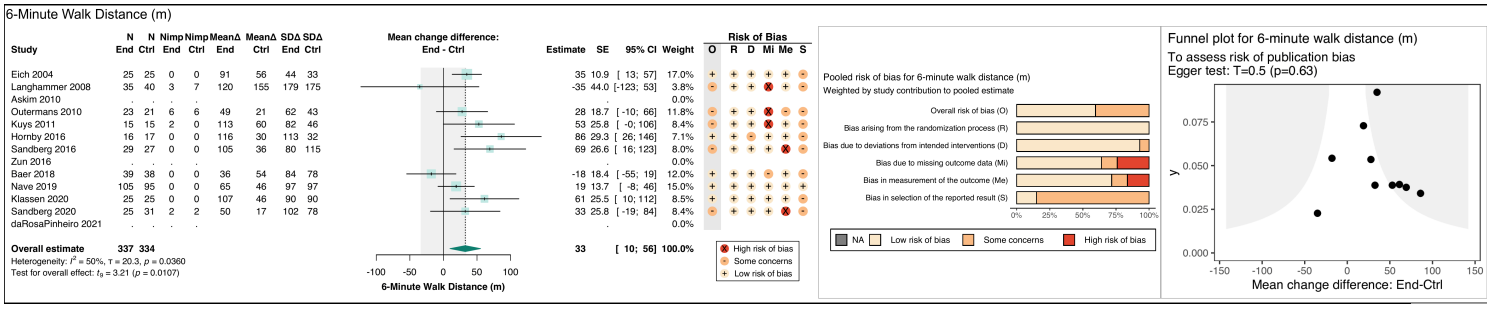


Figure 1

Example of meta-analysis outputs for one treatment vs. control comparison for one outcome. Abbreviations: N, number of participants; Nimp, number of participants with imputed missing data; **Δ**, change; SD, standard deviation; End, cardiovascular endurance training; Ctrl, control treatment; SE, standard error; CI, confidence interval.