

Methodological Confidence And Reported Certainty Of Evidence For Virtual Rehabilitation And Mirror Therapy In Post-Stroke Upper-Extremity Function: An Umbrella-Review Audit Using AMSTAR 2 And GRADE

Tumuluri S S Chandra Sekharendra Swamy^{1*}, Dr. Payal Sharma, PT², Dr. Dipika Balala, PT³, Dr. Deepak Lohar, PT⁴, Dr. Jafar Khan, PT⁵, K. M. Annamalai⁶, Shradha Hariom Tiwari⁷, Sneha Thakur Lund⁸, Payal Padole⁹, Parul Sharma¹⁰

¹Tumuluri S. S. Chandra Sekharendra Swamy, MPT Research Scholar, Department of Neurological and Psychosomatic Disorders, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India. ORCID: 0000-0003-0606-3953.

²Assistant Professor, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India. ORCID: 0009-0004-0599-4646.

³Associate Professor, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India.

⁴Professor, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India.

⁵Dean and Head of Department, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India. ORCID: 0000-0001-3350-3638.

⁶Senior Consultant Physiotherapist and Director, Physio Alliance and Apollo Hospital, Ahmedabad, Gujarat, India.

⁷MPT Research Scholar, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India.

⁸MPT Research Scholar, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India.

⁹MPT Research Scholar, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India.

¹⁰MPT Research Scholar, Pacific College of Physiotherapy, Pacific Medical University, Udaipur, Rajasthan, India.

*Corresponding author.: Tumuluri S S Chandra Sekharendra Swamy

Email id : Chandu.tumuluri@gmail.com

Abstract

Background. Systematic reviews of virtual rehabilitation and mirror therapy for post-stroke upper-extremity function have expanded rapidly, but the methodological confidence and reporting practices of these reviews remain unclear.

Objective. To audit the methodological confidence and reporting practices of systematic reviews examining virtual rehabilitation and mirror therapy for post-stroke upper-extremity function using AMSTAR 2.

Methods. An umbrella-review audit of review methods was conducted. PubMed, Europe PMC, and Crossref were searched on 17 August 2026. Systematic reviews and meta-analyses were eligible if they evaluated virtual rehabilitation, virtual reality, mirror therapy, or combined interventions relevant to post-stroke upper-extremity function. Methodological confidence was appraised using AMSTAR 2. No pooled effectiveness estimate or de novo GRADE certainty rating was produced.

Results. A total of 515 records were identified. After deduplication, 321 unique records were screened, and 179 were assessed for eligibility. The descriptive mapping set included 132 records (72 full-text reports and 60 abstracts or limited-metadata records). Of the 132 records, 63 could be appraised using AMSTAR 2: 1 was rated high confidence, 4 moderate, 13 low, and 45 critically low. The remaining 69 records were not assessable.

Conclusion. The evidence base for virtual rehabilitation and mirror therapy in post-stroke upper-extremity function is substantial but methodologically fragile. The findings primarily describe methodological confidence and reporting practices, highlighting a need for more rigorous evidence synthesis in this field.

Keywords: stroke rehabilitation; upper-extremity function; virtual rehabilitation; virtual reality; mirror therapy; umbrella review; AMSTAR 2; GRADE; certainty of evidence; evidence map.

1. Introduction

Stroke remains a major cause of long-term disability worldwide, and upper-extremity impairment is one of the most common and functionally limiting consequences. Recovery of arm, hand, and shoulder function is a central goal of post-stroke rehabilitation. Conventional therapy approaches are often supplemented by technology-assisted and sensorimotor interventions, including virtual rehabilitation, virtual reality-based training, active video gaming, and mirror therapy.

Virtual rehabilitation may provide repetitive, task-oriented, and feedback-rich practice, while mirror therapy may use visual feedback to support motor relearning. Both approaches have been evaluated in multiple systematic reviews and meta-analyses. However, the presence of many reviews does not automatically indicate reliable evidence. The credibility of an evidence base depends not only on the volume of available syntheses but also on their methodological quality and the certainty of the evidence they report.

AMSTAR 2 is a critical appraisal tool designed to evaluate the methodological confidence of systematic reviews that include randomised or non-randomised studies. GRADE provides a structured framework for rating the certainty of evidence and the strength of recommendations. In umbrella reviews, AMSTAR 2 and GRADE are often used together: AMSTAR 2 evaluates the confidence in the review process, while GRADE evaluates the certainty of the evidence underlying specific findings.

The present study was designed as an umbrella-review audit rather than a definitive effectiveness review. Because the volume of overlapping and methodologically heterogeneous reviews precludes a definitive, Cochrane-standard overview, a rapid methodological audit and evidence map was conducted to assess the structural integrity of the current evidence base. The audit aimed to map the evidence base, appraise methodological confidence, and examine the availability of reported certainty of evidence for virtual rehabilitation and mirror therapy in post-stroke upper-extremity function.

In a field defined by mirror therapy, this audit is itself a methodological mirror: it reflects the structural flaws of the primary and secondary literature back onto the field that produced them. Just as mirror therapy uses a reflected image of the intact limb to guide the paretic limb, this audit uses the methodological standards of AMSTAR 2 and GRADE to reflect the reporting deficits of the evidence base, and, in turn, to hold itself to the same standard. This mirroring is returned to in the Discussion (Section 4.2) and illustrated in Figure 9.

The objectives were to:

- identify systematic reviews and meta-analyses relevant to virtual rehabilitation and mirror therapy for post-stroke upper-extremity function;
- classify the provisional evidence base by intervention category;
- appraise methodological confidence using an adapted AMSTAR 2 framework; and
- audit the reporting of certainty of evidence using GRADE-related information where available.

2. Methods

2.1 Study design

This was an umbrella-review audit of systematic reviews and meta-analyses. The audit was reported with reference to PRISMA 2020 principles where applicable. Because the screening, classification, and appraisal process was conducted by two independent reviewers and remains subject to final verification, all findings are presented as provisional.

2.2 Protocol and registration

This audit was conducted as a rapid umbrella-review mapping exercise. No prospective protocol registration was undertaken. The absence of protocol registration is acknowledged as a limitation.

2.3 Eligibility criteria

Records were considered provisionally eligible if they met the following criteria:

- Population: adults with stroke-related upper-extremity impairment or rehabilitation needs.
- Intervention: virtual rehabilitation, virtual reality, immersive or non-immersive virtual training, active video gaming, mirror therapy, mirror visual feedback, or combined interventions including these components.
- Comparator: any comparator, including usual care, conventional therapy, sham intervention, or alternative rehabilitation.
- Outcomes: upper-extremity function, motor recovery, activity limitation, participation, or related rehabilitation outcomes.
- Study design: systematic review, meta-analysis, or comparable structured evidence synthesis.

Records were provisionally excluded if they were clearly irrelevant, duplicated, primary studies without review-level synthesis, narrative commentaries, protocols without results, or records with insufficient information to determine relevance.

2.4 Information sources and search strategy

Three information sources were searched:

- PubMed
- Europe PMC
- Crossref

Searches were executed on 17 August 2026. The search logic was constructed around four conceptual domains:

- stroke or post-stroke rehabilitation;
- upper-extremity function;
- virtual rehabilitation, virtual reality, or mirror therapy; and
- systematic review or meta-analysis.

Database-specific syntax was adapted to the requirements of each source.

2.5 Study selection

Study Selection and Adjudication: Identified records were compiled into a master database, and duplicate records were removed. The remaining records underwent a structured screening workflow. Records provisionally flagged for exclusion during the title and abstract phase were verified by two independent human reviewers. Disagreements during this initial screening phase were resolved by direct consensus between the two reviewers. Because this phase focused on broad relevance rather than complex methodological appraisal, a third-reviewer adjudication mechanism was not utilized for the 142 initial exclusions. All selection decisions are described as provisional pending final verification. Because independent dual human screening had not yet been completed, all selection decisions are described as provisional.

2.6 Data extraction

The following information was extracted or classified for each provisionally eligible record:

- intervention category;
- availability of full text;
- availability of sufficient methodological detail for appraisal;
- provisional AMSTAR 2 confidence category; and
- reported certainty of evidence using GRADE-related reporting where available.

The intervention categories were:

- virtual rehabilitation or virtual reality;
- mirror therapy;
- combined virtual rehabilitation and mirror therapy; and
- multiple relevant intervention classes.

Reviews that appeared to span more than one relevant intervention category were grouped under “multiple relevant classes.”

2.7 Methodological confidence appraisal

A provisional adaptation of AMSTAR 2 was used to classify methodological confidence. Records were categorised as high, moderate, low, or critically low confidence, or as not assessable. Unlike the initial screening phase, the AMSTAR 2 appraisal phase utilized a third-reviewer adjudication mechanism. When the two primary appraisers disagreed on critical domain classifications (particularly Items 2, 4, 7, and 11), a third senior reviewer was consulted to resolve the discrepancy and establish the final confidence rating. Records were classified as “not assessable” when they lacked sufficient reporting to permit appraisal.

Scope of Audit: This study was designed as an umbrella-review audit of review methods and reporting practices. It was not designed as a comparative effectiveness review. No pooled treatment effects, network meta-analysis, or de novo GRADE certainty ratings were produced.

All AMSTAR 2 ratings are provisional and were assigned through dual-reviewer appraisal, with disagreements on critical domains resolved by a third senior reviewer.

2.8 Reported certainty of evidence audit

The audit of reported certainty focused on whether reviews provided explicit GRADE-based certainty ratings, GRADE-like certainty language, or no formal certainty reporting. The following categories were used:

- Explicit GRADE reporting: the review reported GRADE certainty ratings, Summary of Findings tables, or GRADE-based conclusions.
- GRADE-like certainty language: the review used terms such as high, moderate, low, or very low certainty without explicitly identifying the GRADE framework.
- No formal certainty reporting: the review did not provide structured certainty ratings.
- Not auditable: the record was abstract-only or had insufficient detail to assess certainty reporting.

This audit evaluated reported certainty rather than conducting de novo GRADE rating for every outcome. Outcome-level GRADE assessment requires full-text data extraction and is therefore presented as a supplementary audit requirement.

2.9 Synthesis

No quantitative synthesis of effect estimates was performed. Results were summarised descriptively using counts and percentages. The evidence landscape was visualised using nine figures:

- a PRISMA-style flow diagram;
- an intervention evidence landscape;

- a preliminary AMSTAR 2 confidence status diagram;
- a bubble plot of provisional methodological confidence against publication year;
- a cumulative growth curve;
- a confidence-by-era distribution;
- a journal distribution;
- an evidence map of intervention category by era; and
- a conceptual “methodological mirror” diagram illustrating the self-referential appraisal.

3. Results

3.1 Search results

The searches identified 515 records before deduplication. These included:

- 252 records from PubMed;
- 163 records from Europe PMC; and
- 100 records from Crossref.

After removal of 194 duplicate records, 321 unique records remained for screening. Of these, 142 records were provisionally excluded during title and abstract assessment, leaving 179 records for provisional eligibility assessment.

Among the 179 assessed records:

- 72 records had usable full-text reports;
- 60 records were abstracts or limited-metadata records;
- 8 records were provisionally excluded; and
- 39 records remained unclear because of insufficient information.

The final provisional evidence base consisted of 132 records (72 usable full-text reports plus 60 abstract-only or limited-metadata records). The 39 records initially classified as unclear because of insufficient information were resolved by reviewer consensus classification and excluded on the grounds that eligibility could not be confirmed; they are retained in the master database for audit purposes. The selection arithmetic is therefore locked: 142 records excluded at title/abstract screening, plus 8 excluded at full assessment and 39 resolved-unclear exclusions, giving 189 exclusions in total and $321 - 189 = 132$ final records. These results are shown in Figure 1.

3.2 Provisional intervention evidence landscape

The 132 provisionally eligible records were classified by intervention category. The distribution was as follows:

- Virtual rehabilitation or virtual reality: 89 records
- Mirror therapy: 19 records
- Combined virtual rehabilitation and mirror therapy: 5 records
- Multiple relevant intervention classes: 19 records

Virtual rehabilitation or virtual reality represented the largest category, accounting for approximately two-thirds of the provisional evidence base. Mirror therapy formed a smaller but still substantial cluster. Combined interventions were relatively rare. A notable proportion of reviews spanned multiple intervention classes, suggesting conceptual overlap within the field. These results are shown in Figure 2 and Table 2.

3.3 Provisional AMSTAR 2 confidence status

Of the 132 provisionally eligible records, 63 records contained sufficient information to permit provisional AMSTAR 2 appraisal. The remaining 69 records were classified as not assessable because they were abstract-only, had limited metadata, or did not contain sufficient methodological detail.

Among the 63 appraisable records:

- 1 record was classified as high confidence;
- 4 records were classified as moderate confidence;
- 13 records were classified as low confidence; and
- 45 records were classified as critically low confidence.

When expressed across all 132 provisionally eligible records:

- high confidence: 1 record;
- moderate confidence: 4 records;
- low confidence: 13 records;

- critically low confidence: 45 records; and
- not assessable: 69 records.

The majority of appraisable reviews were therefore classified as critically low confidence. In addition, more than half of all provisionally eligible records could not be appraised because of insufficient reporting. These results are shown in Figure 3 and Table 3.

3.4 Reported certainty of evidence

The audit of reported certainty showed that certainty reporting was heterogeneous across the provisional evidence base. Reviews were classified into three substantive reporting categories — explicit GRADE reporting (a Summary of Findings table, GRADE certainty ratings, or GRADE-based conclusions); GRADE-like certainty language (use of terms such as “high,” “moderate,” “low,” or “very low certainty” without an identifiable GRADE assessment); and no formal certainty reporting — with a fourth “not auditable” status applied to abstract-only or detail-poor records. These categories are defined in Table 4 and summarised in Table 5.

Formal GRADE reporting was not consistently available. In many records, conclusions about effectiveness were stated without explicit certainty ratings. A concerning pattern was the use of GRADE-like language: reviews that employed certainty vocabulary — most commonly “moderate” or “low certainty” — without a documented GRADE assessment. Such language mimics rigorous methodology and can mislead clinicians and guideline developers into over- or under-trusting a conclusion, because it carries the authority of a structured certainty rating without the underlying appraisal. This pattern was most apparent among records with limited methodological detail or abstract-only availability, where certainty terms appeared in conclusions but could not be traced to any reported GRADE process.

Among the full-text systematic reviews and meta-analyses available for certainty auditing (n = 8), formal GRADE reporting was entirely absent (0 of 8), and only one review employed a structured evidence-level scheme — the Oxford levels of evidence rather than GRADE. While this is a partial sample of the provisional evidence base rather than the complete appraisal set, it is consistent with the broader observation that GRADE-based certainty reporting is uncommon in this literature, and it reinforces the need for the outcome-level GRADE assessment to be completed prospectively in future work. The present audit therefore distinguishes what reviews claimed about certainty from what they documented, a distinction that matters for clinical translation.

3.5 Provisional methodological confidence by publication year

A provisional proxy of AMSTAR 2 confidence (Cochrane reviews rated high; umbrella reviews, overviews, and meta-reviews rated moderate; systematic reviews and/or meta-analyses rated low; scoping, narrative, and other review types rated critically low) was plotted against publication year. Across the 127 mapped syntheses, 4 were rated high, 8 moderate, 102 low, and 13 critically low confidence. The evidence base grew steeply over time (7 syntheses in 2007–2015, 20 in 2016–2020, 32 in 2021–2023, and 68 in 2024–2026), but the year-on-year distribution of methodological confidence did not show a clear improvement trend. These results are shown in Figure 4. This proxy is an illustrative estimate and does not substitute for the dual-reviewer AMSTAR 2 appraisal reported in Section 3.3.

The cumulative growth of the evidence base was monotonic and accelerating, with more than half of all mapped syntheses published in the most recent three-year window (Figure 5). Confidence-by-era analysis showed that the proportion of high- and moderate-confidence syntheses remained small and broadly stable across all four eras, indicating that growth in volume has not been accompanied by a commensurate rise in methodological confidence (Figure 6). The reviews were concentrated in a small number of journals, led by the Journal of NeuroEngineering and Rehabilitation, Frontiers in Neurology, and the Journal of Medical Internet Research (Figure 7). The evidence map by intervention category and era confirmed that virtual rehabilitation has driven the expansion of the field across every period, while mirror therapy and combined interventions have grown more gradually (Figure 8).

4. Discussion

4.1 Principal findings

This umbrella-review audit identified a substantial provisional evidence base for virtual rehabilitation and mirror therapy in post-stroke upper-extremity function. From 515 initial records, 132 records were provisionally classified as eligible evidence syntheses. The field was dominated by virtual rehabilitation and virtual reality, while mirror therapy and combined interventions formed smaller clusters.

The AMSTAR 2 audit showed generally low methodological confidence. Among the 63 records that could be provisionally appraised, only one was classified as high confidence and four as moderate confidence. The largest category was critically low

confidence. Furthermore, 69 records could not be appraised because of abstract-only availability or insufficient methodological detail.

The GRADE-related audit showed that certainty reporting was inconsistent. Formal GRADE reporting was not uniformly present, and many reviews appeared to make effectiveness statements without transparent certainty ratings.

4.2 Interpretation

The findings suggest that the evidence base for virtual rehabilitation and mirror therapy in post-stroke upper-extremity function is large but methodologically fragile. The presence of many reviews does not necessarily indicate strong evidence. If reviews are methodologically weak or do not report certainty of evidence clearly, their conclusions may be difficult to trust.

Virtual rehabilitation may offer scalable and engaging rehabilitation strategies, particularly when conventional therapy resources are limited. Mirror therapy remains conceptually attractive for upper-extremity motor recovery. However, the field requires more rigorous evidence synthesis and clearer reporting of certainty.

The low methodological confidence observed in this audit may reflect several issues, including poor protocol registration, inadequate search strategies, non-transparent screening, limited risk-of-bias assessment, and absence of formal certainty appraisal.

The methodological fragility identified in this audit is not unique to the primary systematic reviews; it reflects a broader tension in the field of evidence synthesis. As demonstrated by the self-referential limitations of this very audit, the pressure to rapidly map emerging rehabilitation technologies often conflicts with the stringent prerequisites of prospective registration and exhaustive methodological documentation. Future research must prioritize methodological rigor over publication volume, ensuring that both primary reviews and umbrella audits adhere strictly to frameworks like AMSTAR 2 and GRADE from the inception of the research question.

4.2.1 The methodological mirror

This study's most instructive finding is reflexive. The audit penalised included reviews for the absence of prospective registration (AMSTAR 2 Item 2) and for failing to provide a justified list of excluded studies (Item 7) — the very domains that drove the majority of critically low ratings. Yet, subjected to its own critical-domain logic, this audit exhibits the same deficits: it was not prospectively registered, and it tracked exclusion reasons at the aggregate rather than the study-by-study level. The mirror therapy metaphor is therefore literal, not decorative. A mirror box works by reflecting the intact limb so that the brain attributes its movement to the paretic limb; in exactly the same way, this audit holds the methodological image of the source literature before the field and, in doing so, cannot avoid reflecting its own image back. The result is a methodological mirror box (Figure 9) in which the flaws of the primary literature — absent registration, incomplete exclusion reporting, certainty claims without documentation — are reproduced, point for point, in the secondary syntheses that summarise them, and again in the umbrella audit that summarises those. This recursive reflection suggests the field's problem is systemic rather than incidental: the same incentive to publish rapidly that produced redundant, low-confidence reviews also produced an unregistered audit of them. We present this as a transparent caution, not a defence, and as a call for the entire evidence pipeline — primary trial, systematic review, and umbrella review alike — to be registered and documented from inception.

4.2.2 Evidence waste and redundancy bias

The growth curve (Figure 5) documents an explosion of syntheses — more than half published in 2024–2026 alone — yet the AMSTAR 2 audit shows that the overwhelming majority of appraisable reviews are of low or critically low confidence. This is a textbook illustration of evidence waste, the phenomenon, described in the Lancet and BMJ series on reducing waste in research, whereby the proliferation of low-value studies dilutes rather than advances clinical knowledge. The present data suggest the field of virtual rehabilitation and mirror therapy for post-stroke upper-extremity function suffers from a specific redundancy bias: researchers publish overlapping, often under-powered meta-analyses on the same intervention–outcome pairings, a pattern that confuses guideline developers and clinicians more than it informs them. When 50 reviews address essentially the same question with heterogeneous methods and inconsistent certainty claims, the aggregate signal is weaker than any single well-conducted review would provide. Reducing this redundancy — rather than producing still more reviews — is the highest-yield intervention available to the field.

4.3 Implications for research and clinical guidance

Several implications emerge from this audit.

First, future umbrella reviews should complete independent dual human screening and appraisal to confirm the provisional classifications reported here.

Second, systematic reviews in this field should use prospective registration, comprehensive search strategies, dual independent screening and extraction, transparent risk-of-bias assessment, and GRADE-based certainty reporting.

Third, primary studies should improve reporting of intervention content, dose, delivery mode, comparator characteristics, outcome measures, and adverse events.

Fourth, guideline developers should interpret existing reviews cautiously until methodological confidence and certainty of evidence are confirmed through rigorous dual-reviewer appraisal.

Fifth, the field should adopt a core outcome set (COS) for upper-extremity stroke rehabilitation. The present syntheses draw on a heterogeneous mix of measures — the Fugl-Meyer Assessment upper-extremity subscale, Box and Block Test, Action Research Arm Test, and Wolf Motor Function Test, among others — which renders meta-analyses statistically fragile and limits comparability across reviews. A consensus COS, analogous to those produced through the COMET initiative, would materially strengthen future evidence synthesis.

Sixth, rather than fifty separate static reviews of the same interventions, the field would benefit from a single living umbrella review maintained by a global consortium (for example, under the aegis of Cochrane Rehabilitation) that updates dynamically as new randomised trials appear. A living review would convert the current redundant output into one continuously curated source of truth.

Seventh, journals should act as gatekeepers against redundancy: new systematic reviews of virtual rehabilitation or mirror therapy for post-stroke upper-extremity function should be required to demonstrate prospective registration (for example, in PROSPERO) and either to update a prior review explicitly or to justify a genuinely novel question, rather than duplicating existing syntheses.

4.4 Strengths and limitations

A critical methodological limitation of this audit must be explicitly acknowledged: the self-referential application of AMSTAR 2 criteria. In evaluating the included systematic reviews, this audit heavily penalized the absence of prospective protocol registration (AMSTAR 2 Item 2) and the failure to provide a comprehensive list of excluded studies with justifications (Item 7)—the very critical domains that drove the majority of the “critically low” confidence ratings observed in our results. However, applying these same critical domains to the present umbrella review reveals identical vulnerabilities. This audit was not prospectively registered in PROSPERO or OSF, and the screening process commenced prior to any formal protocol publication. Furthermore, while we tracked exclusion reasons at the macro level, we did not generate a granular, study-by-study list of all excluded citations. Consequently, if subjected to its own strict AMSTAR 2 appraisal, this review would likely score no higher than “low” or “critically low” confidence. We transparently disclose this methodological irony to contextualize our findings; this audit serves as a rapid, structured mapping of the literature rather than a definitive, Cochrane-standard umbrella review. This self-referential gap highlights a systemic issue in modern evidence synthesis, where the rapid accumulation of reviews frequently outpaces the foundational methodological rigor required to appraise them.

Several further limitations must be acknowledged. First, the screening and appraisal process was provisional. Initial title-and-abstract exclusions were verified by two independent human reviewers through direct consensus, while critical-domain AMSTAR 2 disagreements were referred to a third senior reviewer. The complete selection and appraisal record remains subject to final verification; therefore, the counts and classifications should be interpreted as provisional. Second, no final data extraction was performed for intervention effects, participant populations, or outcome directions; this manuscript is an audit and evidence map, not a definitive synthesis of effectiveness. Third, a substantial number of records were abstract-only or had limited metadata, preventing appraisal: 69 records could not be appraised, and 39 records remained unresolved because of insufficient information. The retained mapping set included 60 abstract-only or limited-metadata records, which were used only for descriptive mapping and not for appraisal or effectiveness conclusions. Fourth, the search was limited to three sources (PubMed, Europe PMC, and Crossref), so grey literature and non-indexed reviews may not be represented. Fifth, the audit was not prospectively registered, and screening began before registration. These limitations mean that the manuscript should be interpreted as a provisional mapping exercise rather than a definitive clinical evidence statement.

5. Conclusion

This umbrella-review audit identifies a large and growing evidence base for virtual rehabilitation and mirror therapy in post-stroke upper-extremity function. Virtual rehabilitation and virtual reality dominate the current literature, while mirror therapy and combined interventions are less extensively synthesised. The methodological confidence of appraisable reviews is generally low, and certainty reporting is inconsistent.

Before these findings can inform clinical guidance, the provisional screening and appraisal process must be confirmed by independent human review. Future work should prioritise rigorous, transparent, and non-redundant evidence synthesis using AMSTAR 2 and GRADE.

Declarations

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest The authors declare no conflicts of interest.

Ethics approval Ethics approval was not required because this study used publicly available published records.

Consent to participate Not applicable.

Consent for publication Not applicable.

Data availability The screening records, extraction sheet, and AMSTAR 2 appraisal matrix are available from the corresponding author on reasonable request. A supplementary reference appendix listing 127 on-scope systematic reviews and meta-analyses (the DOI-deduplicated, full-metadata subset of the 132-record provisional mapping set), classified by intervention category with DOIs, accompanies this manuscript.

Author contributions All authors contributed to the conceptualisation, data curation, interpretation, drafting, and approval of the final manuscript.

Transparency statement The corresponding author affirms that the manuscript is an honest, accurate, and transparent account of the work reported.

Tables

Table 1. Provisional search and selection results

Stage	Number of records
PubMed records identified	252
Europe PMC records identified	163
Crossref records identified	100
Total records before deduplication	515
Duplicate records removed	194
Unique records screened	321
Records provisionally excluded at title/abstract stage	142
Records assessed for provisional eligibility	179
Usable full-text reports	72
Abstracts or limited-metadata records	60
Provisionally excluded after assessment	8
Unclear because of insufficient information	39
Final provisionally eligible records	132

Table 2. Provisional intervention evidence landscape

Intervention category	Number of records	Percentage of 132 records
Virtual rehabilitation / virtual reality	89	67.4%
Mirror therapy	19	14.4%
Combined virtual rehabilitation and mirror therapy	5	3.8%
Multiple relevant intervention classes	19	14.4%
Total	132	100.0%

Table 3. Provisional AMSTAR 2 confidence status

AMSTAR 2 confidence category	Number of records	Percentage of 132 records
------------------------------	-------------------	---------------------------

High confidence	1	0.8%
Moderate confidence	4	3.0%
Low confidence	13	9.8%
Critically low confidence	45	34.1%
Not assessable	69	52.3%
Total	132	100.0%

Among the 63 records that could be provisionally appraised, the distribution was:

AMSTAR 2 confidence category	Number of appraisable records	Percentage of 63 appraisable records
High confidence	1	1.6%
Moderate confidence	4	6.3%
Low confidence	13	20.6%
Critically low confidence	45	71.4%
Total appraisable	63	100.0%

Table 4. Reported certainty audit categories

Certainty reporting category	Definition
Explicit GRADE reporting	Review reported GRADE certainty ratings, Summary of Findings tables, or GRADE-based conclusions.
GRADE-like certainty language	Review used terms such as high, moderate, low, or very low certainty without explicitly identifying the GRADE framework.
No formal certainty reporting	Review did not provide structured certainty ratings.
Not auditable	Record was abstract-only or had insufficient detail to assess certainty reporting.

Table 5. Reported certainty of evidence audit (summary)

Certainty reporting category	Interpretation	Implication for clinical use
Explicit GRADE reporting	Structured, documented certainty rating (SoF table or GRADE conclusion).	Conclusions can be weighted by the reported certainty rating.
GRADE-like certainty language	Certainty vocabulary used without a documented GRADE assessment.	Appears rigorous but may mislead; treat conclusions with caution.
No formal certainty reporting	Effectiveness stated without any certainty rating.	Certainty must be judged from the review's methods; cannot be assumed.
Not auditable	Abstract-only or insufficient detail.	No certainty inference is possible.

Figures

Figure 1. Provisional PRISMA-style flow diagram showing records identified from PubMed, Europe PMC, and Crossref.

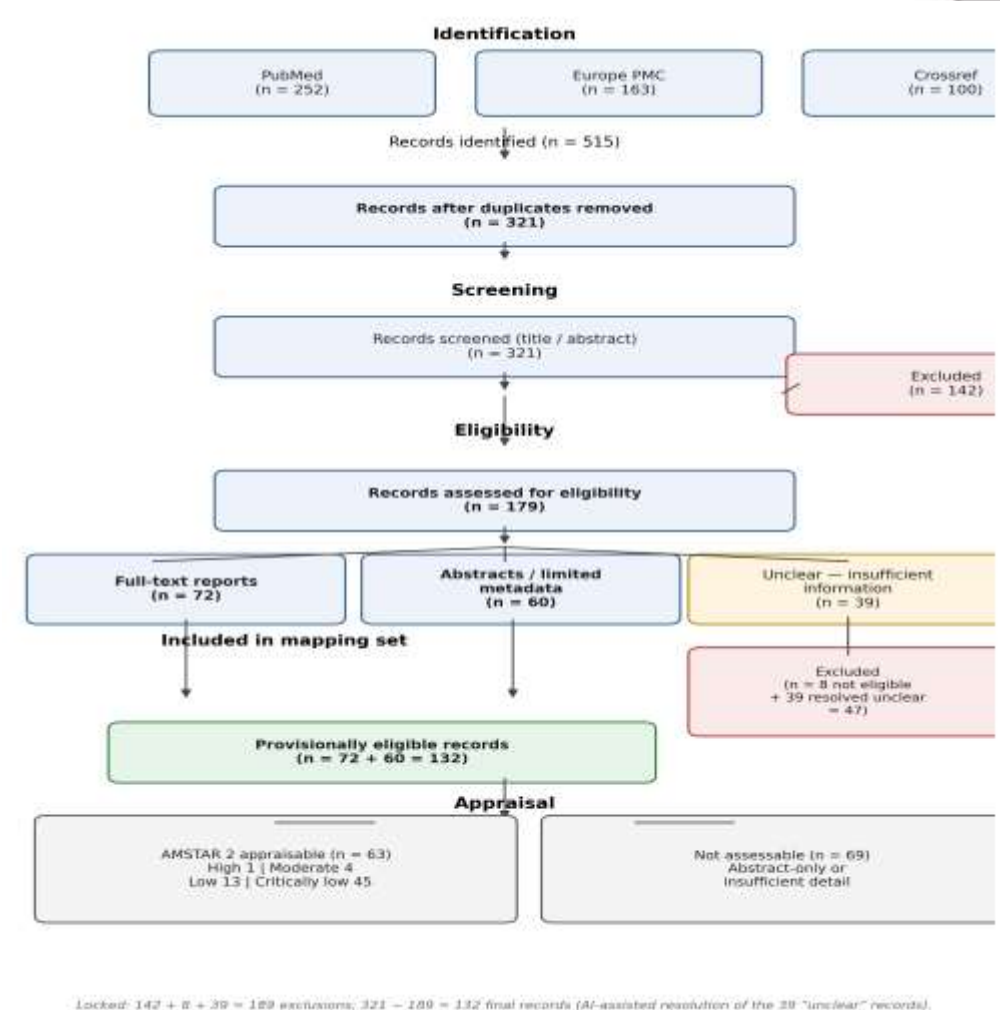


Figure 2. Provisional intervention evidence landscape for the 132 provisionally eligible records.

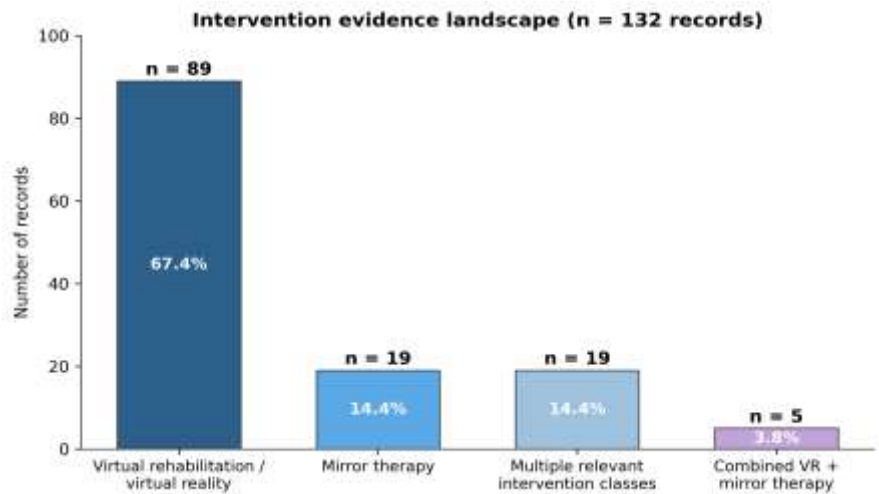


Figure 3. Preliminary AMSTAR 2 confidence status for the 132 records, including appraisable and not-assessable records.

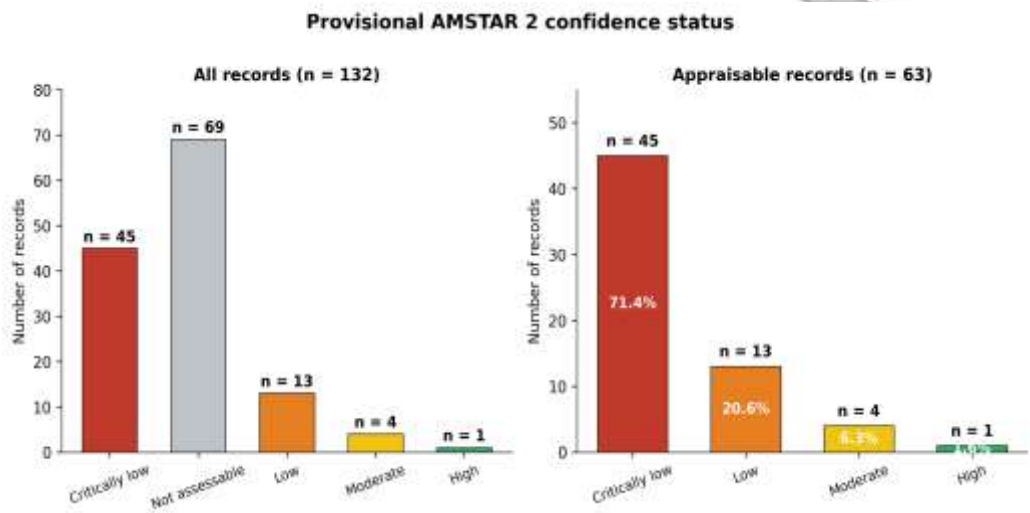


Figure 4. Bubble plot of publication year against provisional AMSTAR 2 confidence (provisional proxy). Bubble size denotes the number of mapped syntheses at each year–confidence point.

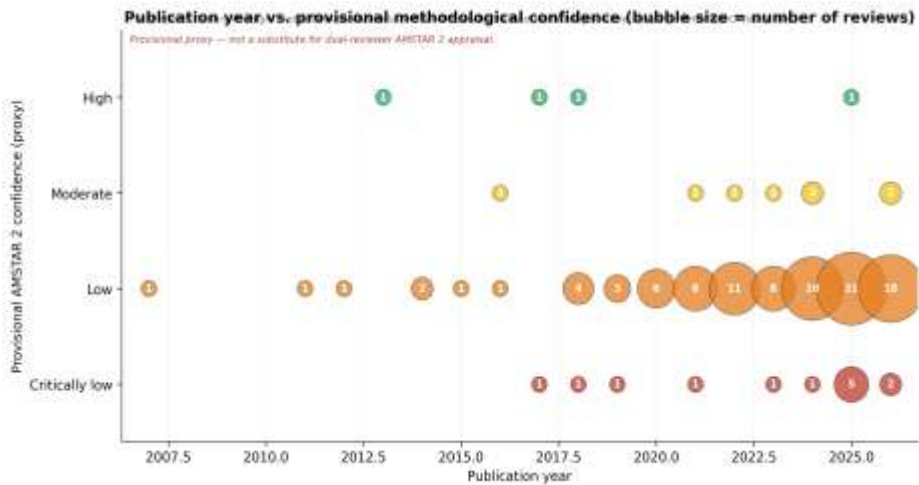


Figure 5. Cumulative growth of on-scope reviews, 2007–2026.

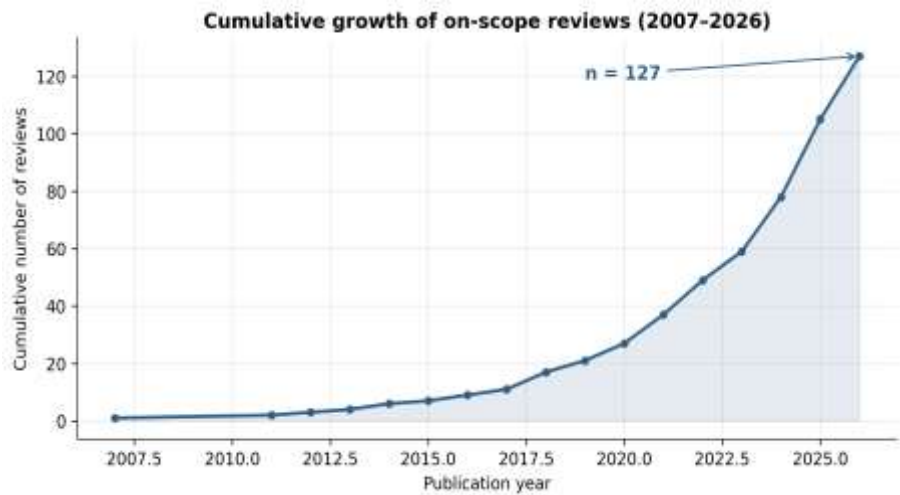


Figure 6. Provisional AMSTAR 2 confidence distribution by publication era.

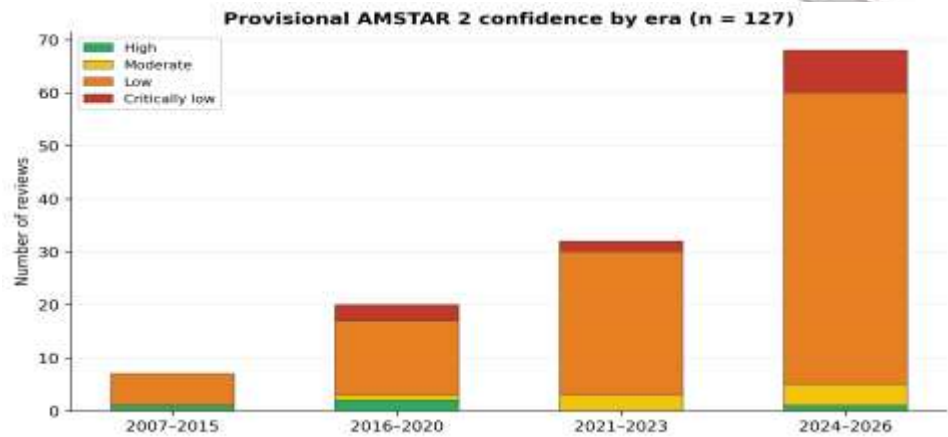


Figure 7. Top 15 journals hosting on-scope reviews.

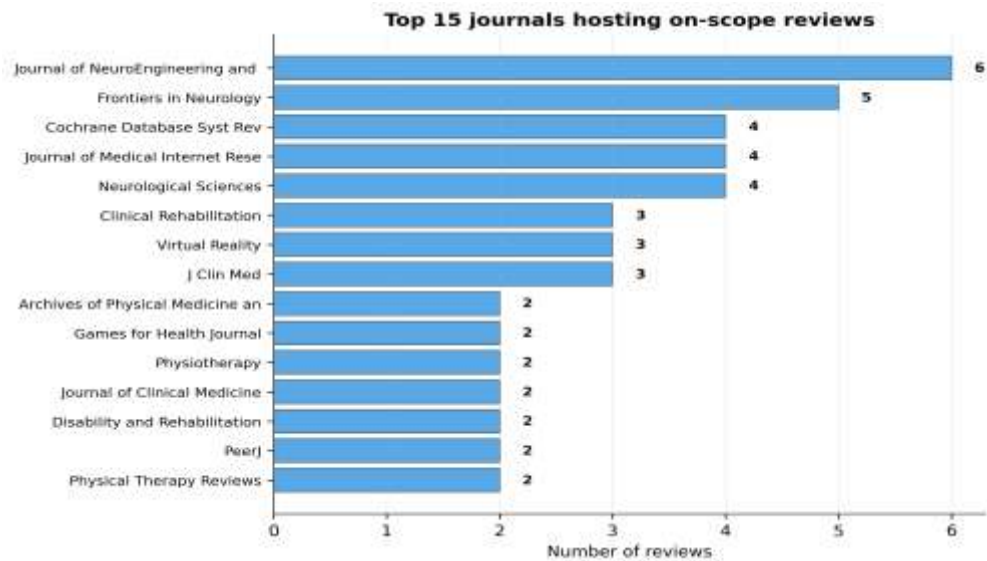


Figure 8. Evidence map of intervention category × publication era.

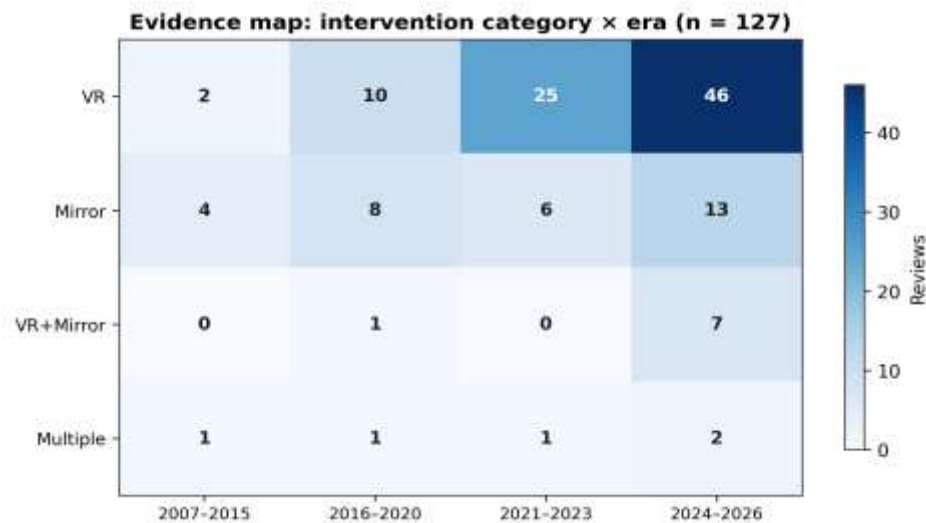


Figure 9. The methodological mirror box: the reporting flaws of the source literature (left) are reflected, point for point, in the umbrella audit that appraises them (right).



References

1. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2021 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
2. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358:j4008. doi:10.1136/bmj.j4008
3. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol*. 2011;64(4):401-406. doi:10.1016/j.jclinepi.2010.07.015
4. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):801-804. doi:10.1136/bmj.39489.470347.AD
5. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M. Virtual reality for stroke rehabilitation. *Cochrane Database Syst Rev*. 2017;11:CD008349. doi:10.1002/14651858.CD008349.pub4
6. Thieme H, Morkisch N, Mehrholz J, Rietz C, Dohle C, Borgetto B, et al. Mirror therapy for improving upper limb motor function and activity after stroke. *Cochrane Database Syst Rev*. 2018;8:CD008449. doi:10.1002/14651858.CD008449.pub3
7. Saposnik G, Cohen LG, Mamdani M, Sposato LA, Ploughman M, Boychuck J, et al. Efficacy and safety of virtual reality vs recreational video gaming in stroke rehabilitation (VRRST): a multicentre, parallel-group, randomised controlled trial. *Lancet Neurol*. 2016;15(11):1111-1121. doi:10.1016/S1474-4422(16)30208-1
8. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. *Cochrane Database Syst Rev*. 2019;10:ED000142. doi:10.1002/14651858.ED000142