

The Umbrella Collaboration®: An Innovative Tertiary Evidence Synthesis Methodology. Validation Study

Beltran Carrillo, Marta Rubinos-Cuadrado, Jazmin Parellada-Martin, Alejandra Palacios-López, Beltran Rubinos-Cuadrado, Fernando Canillas del Rey, Juan Jose Baztán Cortés, Javier Gómez Pavón

Submitted to: JMIR Formative Research
on: March 30, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 51

 TOC/Feature image for homepages 52

 TOC/Feature image for homepage 0 53



The Umbrella Collaboration®: An Innovative Tertiary Evidence Synthesis Methodology. Validation Study

Beltran Carrillo^{1*} MD, PhD; Marta Rubinos-Cuadrado^{2*} RNC; Jazmin Parellada-Martin² BBI; Alejandra Palacios-López² BME; Beltran Rubinos-Cuadrado² BME; Fernando Canillas del Rey³ MD, PhD; Juan Jose Baztán Cortés³ MD, PhD; Javier Gómez Pavón³ MD, PhD

¹ The Umbrella Collaboration Madrid ES

² Hospital Central de la Cruz Roja San José y Santa Adela Madrid ES

*these authors contributed equally

Corresponding Author:

Beltran Carrillo MD, PhD

Abstract

Background: The synthesis of evidence in healthcare is essential for informed decision-making and policy development. This study aims to validate The Umbrella Collaboration® (TU®), an innovative, semi-automatic tertiary evidence synthesis methodology, by comparing it with Traditional Umbrella Reviews (TUR), which are currently the gold standard.

Objective: The primary objective of this study is to evaluate whether TU®, an AI-assisted, software-driven system for tertiary evidence synthesis, can achieve comparable effectiveness to TURs, while offering a more timely, efficient, and comprehensive approach.

Methods: This comparative study evaluated TU® against TURs across eight matched projects in geriatrics. For each selected TUR, a parallel TU® project was conducted using the same research question. Outcomes of interest (OoIs), effect sizes, certainty ratings, and execution times were systematically compared. Effect sizes were assessed both quantitatively, by transforming TUR metrics to Cohen's d and correlating them with TU®'s RTU metric, and qualitatively, through categorical classifications (trivial, small, moderate, large). Certainty levels were compared by mapping GRADE ratings and TU®'s sentiment analysis scores onto a common 0–1 scale. Execution time was measured precisely in TU®, while TUR durations were estimated from literature benchmarks. Statistical analyses included chi-squared tests and Spearman correlations.

Results: Eight TURs in geriatrics were matched with parallel projects using TU®. TU® replicated 84.9% (73/86) of the OoIs identified by TURs and reported an additional 337 OoIs, representing a 4.77-fold increase in outcome identification. In the comparison of effect size classifications, full concordance was observed in 50.0% of cases and consistent concordance (full plus one-level deviation) in 93.8%, with a moderate strength of association (Cramér's V = 0.339). The correlation of transformed certainty values between TU® and GRADE yielded a statistically significant Spearman coefficient ($\rho = 0.446$; $P = .025$). The average execution time per TU® project was 4 hours and 46 minutes, compared to estimated durations of 6–12 months for TURs.

Conclusions: The Umbrella Collaboration® demonstrated high concordance with TURs, replicating 84.9% of the outcomes identified by TURs and identifying nearly five times as many additional outcomes. The experimental effect size metric (RTU) showed moderate agreement with conventional measures, and the certainty ratings derived from sentiment analysis correlated acceptably with GRADE-based assessments. While further validation is needed, TU® appears to be a valid and efficient approach for tertiary evidence synthesis, offering a scalable and time-efficient alternative when rapid results are required.

(JMIR Preprints 30/03/2025:75215)

DOI: <https://doi.org/10.2196/preprints.75215>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#)

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

The Umbrella Collaboration[®]: An Innovative Tertiary Evidence Synthesis Methodology.

Authors:

Beltrán Carrillo^{a,b}, Rubinos-Cuadrado, M^a, Parellada-Martin, J^a, Palacios-López, A^a, Carrillo-Rubinos, B^a, Canillas del Rey, F^{b, c}, Baztán-Cortés, JJ^b, Gómez-Pavón, J^{b, c}.

a: The Umbrella Collaboration[®]

b: Hospital Universitario Cruz Roja, (Madrid, Spain)

c: Universidad Alfonso X El Sabio, Villanueva de la Cañada (Madrid, Spain)

email autor: Beltran Carrillo. (bcm@theumbrellacollaboration.org)

ORCID:

Beltran Carrillo: 0000-0001-5748-9423

Rubinos-Cuadrado, M: 0000-0003-0108-2629

Parellada-Martin, J: 0000-0002-8528-4481

Palacios-López, A: 0000-0002-7263-9167

Carrillo-Rubinos, B: 0000-0002-0796-4414

Canillas del Rey, F: 0000-0002-3326-3494

Baztán-Cortes, JJ: 0009-00099792-0404

Gómez-Pavón, J: 0000-0002-4732-6265

ABSTRACT

BACKGROUND:

The synthesis of evidence in healthcare is essential for informed decision-making and policy development. This study aims to validate The Umbrella Collaboration[®] (TU[®]), an innovative, semi-automatic tertiary evidence synthesis methodology, by comparing it with Traditional Umbrella Reviews (TUR), which are currently the gold standard.

OBJECTIVES:

The primary objective of this study is to evaluate whether TU[®], an AI-assisted, software-driven system for tertiary evidence synthesis, can achieve comparable effectiveness to TURs, while offering a more timely, efficient, and comprehensive approach.

METHODS:

This comparative study evaluated TU[®] against TURs across eight matched projects in geriatrics. For each selected TUR, a parallel TU[®] project was conducted using the same research question.

Outcomes of interest (OoIs), effect sizes, certainty ratings, and execution times were systematically compared. Effect sizes were assessed both quantitatively, by transforming TUR metrics to Cohen's d and correlating them with TU[®]'s R_{TU} metric, and qualitatively, through categorical classifications (trivial, small, moderate, large). Certainty levels were compared by mapping GRADE ratings and TU[®]'s sentiment analysis scores onto a common 0–1 scale. Execution time was measured precisely in TU[®], while TUR durations were estimated from literature benchmarks. Statistical analyses included chi-squared tests and Spearman correlations.

RESULTS:

Eight TURs in geriatrics were matched with parallel projects using TU[®]. TU[®] replicated 84.9% (73/86) of the OoIs identified by TURs and reported an additional 337 OoIs, representing a 4.77-fold increase in outcome identification. In the comparison of effect size classifications, full concordance was observed in 50.0% of cases and consistent concordance (full plus one-level deviation) in 93.8%, with a moderate strength of association (Cramér's $V = 0.339$). The correlation of transformed certainty values between TU[®] and GRADE yielded a statistically significant Spearman coefficient ($\rho = 0.446$; $P = .025$). The average execution time per TU[®] project was 4 hours and 46 minutes, compared to estimated durations of 6–12 months for TURs.

CONCLUSION:

The Umbrella Collaboration[®] demonstrated high concordance with TURs, replicating 84.9% of the outcomes identified by TURs and identifying nearly five times as many additional outcomes. The experimental effect size metric (R_{TU}) showed moderate agreement with conventional measures, and the certainty ratings derived from sentiment analysis correlated acceptably with GRADE-based assessments. While further validation is needed, TU[®] appears to be a valid and efficient approach for tertiary evidence synthesis, offering a scalable and time-efficient alternative when rapid results are required.

FUNDING:

This project has received a research grant from the Fundación Alfonso X el Sabio.

CONFLICTS OF INTEREST:

The authors Beltrán Carrillo (BC), Rubinos-Cuadrado (MR), Parellada-Martin (JP), Palacios-López (AP), and Carrillo-Rubinos (BCR) were involved in the development of The Umbrella

Collaboration[®] (TU[®]) software. Additionally, BC and MR are owners of the software and may have financial or intellectual property interests in its future applications.

KEY WORDS: Tertiary Evidence Synthesis, The Umbrella Collaboration, Umbrella Reviews, Health Research Methodology, AI-assisted Synthesis, Evidence-based Decision Making



The Umbrella Collaboration®: An Innovative Tertiary Evidence Synthesis Methodology. Validation Study

INTRODUCTION

The synthesis of evidence in healthcare is a knowledge acquisition process designed to transform extensive volumes of data into manageable information to support decision-making based on the best possible evidence. It aims to integrate information from multiple sources on complex topics in a comprehensive, precise, transparent, and easily understandable manner. These principles ensure that synthesized evidence is accessible and useful to all stakeholders, including healthcare professionals, policymakers, and patients [1]. Evidence synthesis plays a pivotal role in Knowledge Translation (KT) [2], serving as a bridge between research and healthcare. Consequently, evidence synthesis is essential for the development of health policies and informed decision-making [3].

The role of all stakeholders in health decision-making underscores the importance of their ability to access, understand, and evaluate health information adequately [4]. A persistent challenge in KT is the low level of statistical and health literacy among the general population and health professionals, which significantly complicates effective health management [5,6]. Therefore, it is crucial to democratize access to high-level health information and to promote active participation of all stakeholders in decisions affecting healthcare [7].

The evolution of evidence synthesis methodologies has led to the development of tertiary synthesis, designed to condense knowledge from multiple systematic reviews with or without meta-analyses (SR/MA). This synthesis, often referred to as Umbrella Reviews now referred to as Traditional Umbrella Reviews (TUR) to distinguish them from the experimental methodology under study, The Umbrella Collaboration® (TU®), builds upon the concept of primary studies (individual studies with participant samples) and secondary studies (systematic reviews and analysis of those primary studies). Tertiary synthesis represents a third level, named with other terms such as overviews, meta-epidemiological studies, meta-meta-analyses, meta-synthesis, and meta-reviews also describing this approach [8,9]. Tertiary synthesis has gained prominence in contexts where broad research questions are posed, rapid results are needed, and resources for extensive systematic reviews are limited. TURs follow a structured methodological process that involves several clearly defined stages [9–14]. Although organizations like Cochrane Collaboration [14] and the Joanna Briggs Institute (JBI) [11]

have developed and continually updated detailed methodologies for these reviews, there remains considerable divergence in how TUR authors implement these steps in practice. This methodological freedom leads to significant variations among different TURs in terms of rigor and approach.

The urgency for high-quality, timely information during crises like the SARS-CoV-2 pandemic has highlighted the critical need for faster evidence synthesis methods, even if it means accepting certain limitations in comprehensiveness, detail, and precision [15]. This demand has driven the development of innovative approaches such as The Umbrella Collaboration® (TU®), which leverages AI-assisted software to facilitate tertiary evidence synthesis under human oversight. While AI tools such as PICO Portal, DistillerSR, Covidence, and Rayyan have already improved secondary evidence synthesis by streamlining data management and analysis [16], technologies like large language models (LLMs), including ChatGPT, are beginning to demonstrate potential for conducting systematic reviews autonomously, though human supervision remains essential to mitigate risks such as errors and hallucinations [17]. Despite these developments in secondary synthesis, the application of AI and software engineering in tertiary synthesis is still in its early stages, with no dedicated software currently available. TU® is at the forefront of this field, pioneering the integration of AI and software engineering with human oversight to ensure accuracy and minimize technology-induced errors. As AI continues to evolve, it is likely that fully automated processes for both secondary and tertiary synthesis will emerge, potentially revolutionizing clinical research and practice. However, building confidence in these technologies will require ongoing development and rigorous validation.

The Umbrella Collaboration® (patent pending)

TU® is primarily a software-driven system engineered to streamline tertiary evidence synthesis, relying on programmed algorithms to automate the majority of its functions. The core of the system is built on a software infrastructure that processes and synthesizes data from SR/MA abstracts stored in Medline. While AI plays a crucial role, particularly through the use of LLMs and machine learning (ML), it is utilized selectively within the broader software framework to enhance specific tasks.

LLMs are used in generating related search terms, expanding upon human-generated queries to enhance the comprehensiveness of literature searches. Any LLM can be adapted to TU® software, up to date we have used ChatGPT 4 [18]. This function is crucial in broadening the scope of the searches while ensuring that the results remain relevant and precise. The AI component is designed to support, not replace, human oversight, ensuring that the final selection of literature is both

accurate and comprehensive [19]. To mitigate the risk of AI-generated hallucinations due to insufficient data, TU[®] primarily operates as a stable, auditable software system. The use of AI is limited to the search term expansion phase, where it suggests synonymous terms for the keywords provided by the human reviewer. All AI-generated search terms are subject to human validation before being incorporated into the search strategy. The human reviewer evaluates the relevance and appropriateness of each AI-suggested term, ensuring that only those that align with the research objectives are retained. This manual oversight serves as a critical safeguard against inaccuracies or misleading AI-generated suggestions, maintaining the methodological integrity of the evidence synthesis process.

As the TU[®] database grows, ML will be integrated to further refine and optimize the software's performance. Training the system on an expanding dataset is expected to enhance its ability to select, categorize, and analyze relevant research, thereby improving both efficiency and accuracy over time. This approach allows TU[®] to evolve, continuously improving its utility in evidence synthesis through the iterative learning process [20].

Overall, TU[®] represents a hybrid model where traditional software engineering and targeted AI applications work in tandem. This balance ensures that while the software performs most functions automatically, AI enhances specific tasks, such as search term generation and future predictive analysis, under strict human supervision. This strategic integration of AI elements within a primarily software-driven system ensures the reliability and precision of the evidence synthesis process.

The outcomes generated by TU[®] will be presented through an interactive web application designed to enhance accessibility and comprehension for a broad range of stakeholders, including those without advanced statistical expertise. The use of graphical formats and clear language aims to facilitate the interpretation of findings by diverse audiences.

Additionally, the platform supports continuous updates, automatically integrating new SR/MA data every 24 hours, thereby ensuring the most current and reliable evidence synthesis. This approach aligns with the concept of Living Systematic Reviews (LSRs) [21], which advocate for frequent updates to maintain relevance in rapidly evolving fields. While Cochrane Collaboration recommends updating LSRs monthly [22], TU[®] is designed to surpass this standard by ensuring updates are incorporated daily, providing near real-time evidence synthesis. The daily updates provided by TU[®]

are expected to enhance its efficiency, allowing for a more dynamic and continuously updated evidence synthesis process. If validated by the upcoming research, this advancement could demonstrate TU[®]'s potential to streamline the synthesis process, delivering rapid yet reliable results while upholding the highest standards of accuracy. Should these findings be confirmed, TU[®] may emerge as an invaluable tool for accelerating the pace of evidence-based research.

TU[®] is designed to maximize computational efficiency while maintaining methodological rigor. Unlike TURs, which require extensive manual data extraction and synthesis over months, TU[®] automates critical steps in the tertiary synthesis process, significantly reducing execution time.

A key distinction of TU[®] is that it assists in the entire tertiary review process, rather than being limited to isolated stages. Compared to existing approaches such as DistillerSR, Covidence, and Rayyan, which focus on specific tasks like study screening or data extraction, TU[®] integrates a complete methodology for tertiary evidence synthesis. This includes search term expansion, literature retrieval, synthesis automation, and structured result visualization. While other tools provide assistance in certain steps, TU[®] ensures a fully automated, structured, and reproducible workflow for umbrella reviews.

To our knowledge, there are currently no automated systems specifically designed for tertiary evidence synthesis that comprehensively address and assist in the entire synthesis process. While several tools exist for systematic review automation, such as DistillerSR, Covidence, and Rayyan, these primarily focus on secondary evidence synthesis (i.e., systematic reviews) and are not designed to facilitate tertiary synthesis methodologies like umbrella reviews.

Additionally, no previous tool has undergone a formal assessment or validation for automated tertiary synthesis, as TU[®] is the first system explicitly developed for this purpose. Unlike existing software, TU[®] does not merely automate isolated steps (such as literature screening or data extraction) but provides a structured, end-to-end approach for tertiary evidence synthesis. This distinction highlights the novelty of TU[®] and underscores the need for the present study to formally assess its performance compared to TURs.

A schematic diagram is provided to clearly illustrate TU[®]'s workflow, showing the process from data acquisition to the generation of synthesized results. Abstracts of SR/MA, retrieved from the Medline

database via PubMed, form the foundation of the analysis. The decision to use Medline via PubMed as the sole database for literature retrieval in TU[®] is based on its strong coverage of SR/MA, as well as the feasibility of leveraging LLMs for search term expansion. Recognizing the potential limitation of relying on a single database, we conducted preliminary assessments to estimate the impact of this decision. Specifically, we evaluated the ability of TU[®] to retrieve references from TURs and analyzed the proportion of SR/MA references found in Medline. Our results showed that TU[®] was able to retrieve 86% of TUR references using its AI-assisted search methodology in Medline alone. Additionally, an independent assessment of 511 references from 22 TURs found that only 11 references were not indexed in Medline. While some loss of relevant studies is inevitable, these findings suggest that the methodological approach used in TU[®] remains sufficiently comprehensive for tertiary evidence synthesis, balancing feasibility and completeness (unpublished data). The data obtained is processed through a range of techniques, including Natural Language Processing (NLP), sentiment analysis (SA), web scraping (WS) and ML. The expected results comprise synthesized evidence on intervention effectiveness and risk exposures, presented in a graphical and visual format. These results are conveyed in plain language, making them easily understandable by all stakeholders, regardless of their statistical literacy.

The project is continuously updated through automated and on-demand searches, with data from new studies seamlessly integrated into the existing body of evidence. Each inclusion restarts the synthesis process, creating a dynamic, cyclical workflow that ensures the results of the project remain up-to-date (Figure 1).

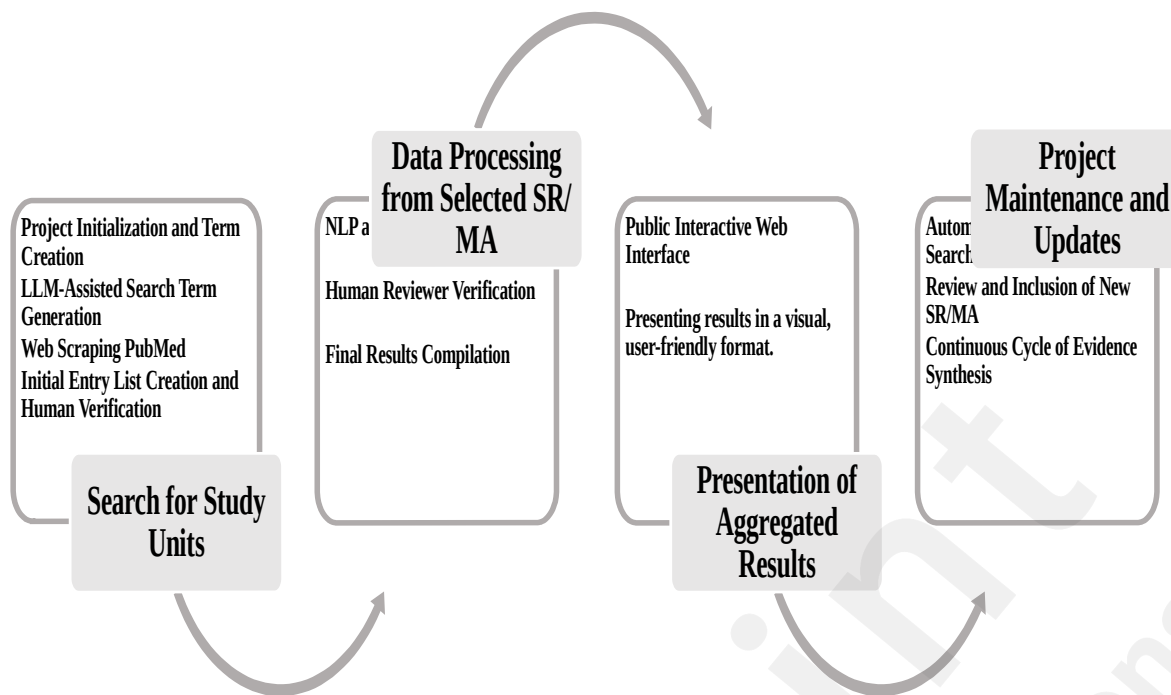


Figure 1 The Umbrella Collaboration® Workflow

The implementation of new methodologies in the scientific field requires a comparative validation process with established methods to ensure their reliability and effectiveness. TU[®], being an innovative methodology still in its theoretical-conceptual stage, must be evaluated against established methodologies. Therefore, the aim of this study is to validate TU[®] by comparing its performance and outcomes with the gold standard, TURs, to establish its credibility and potential superiority.

Objectives

The primary objective of the present study is to assess whether a software-driven AI-assisted system of evidence synthesis, TU[®], can match the effectiveness of traditional methods of tertiary synthesis, providing a potentially more timely, efficient, and comprehensive approach while remaining open to findings that could demonstrate superior performance.

METHODS AND MATERIALS

Study Design

Figure 2 provides an overview of the study design. The study follows a structured comparative approach in which selected TURs in geriatrics serve as the gold standard for validation. Using the

same research questions, projects are conducted in parallel with TU® to assess its performance. Data from both methodologies are systematically collected and compared, focusing on key variables such as the identification of outcomes of interest, effect size, certainty of evidence, and execution time. *The figure illustrates the step-by-step workflow of the study, from the selection of umbrella reviews in geriatrics as reference models to the comparison of results obtained using TU® and TURs.*

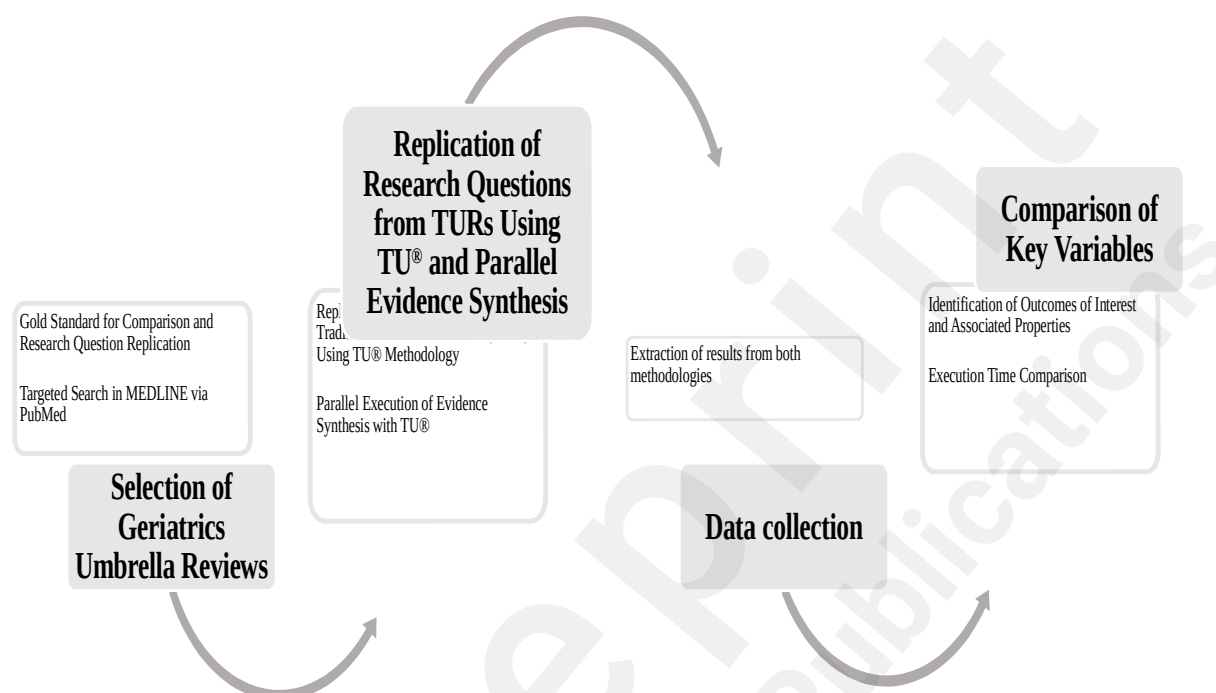


Figure 2
Study workflow.

The study focuses on a quantitative comparison between the two tertiary synthesis methodologies. To facilitate this comparison, a targeted search in PubMed identified relevant TURs in geriatrics, focusing on representative reviews rather than an exhaustive search. Our approach involved a focused search in PubMed, using specific terms relevant to geriatrics, to find suitable TURs that serve as a benchmark for this comparative analysis. This targeted search is considered sufficient for our methodological comparison and does not require the comprehensive search strategy typical of systematic reviews, as our goal is not to cover the entire scope of available literature but to enable a parallel evaluation of synthesis methodologies. Therefore, while the search strategy may appear basic, it is intentionally designed to fulfill the specific needs of our project without aiming for exhaustive literature retrieval, which is beyond the scope of this project.

Study Variables

In the quantitative comparison, several critical variables were analyzed. Key among these is the identification and evaluation of Outcomes of Interest (OoIs). It is essential to define the concept of an "outcome of interest" within the context of tertiary evidence synthesis, as an outcome of interest refers to specific aspects identified and evaluated by systematic reviews with meta-analyses (SR/MA) that examine the same research question. These outcomes are critical for understanding the overall impact of various interventions on health conditions or the effects of exposures to risks. The analysis included assessing the degree of concordance between the methodologies in identifying OoIs, using a concordance matrix to document and compare the OoIs identified by TU[®] and TURs. Additionally, we analyzed the total number of outcomes identified by each methodology, providing a descriptive and statistical comparison.

A crucial aspect of this analysis involves the comparison of effect sizes (ES) for the identified outcomes. TU[®] employs an automated approach for standardizing ES using a custom-designed metric (R_{TU}). This metric transforms all commonly used ES statistics in evidence synthesis (e.g., standardized mean difference (SMD), mean difference (MD), relative risk (RR), odds ratio (OR), hazard ratio (HR), and others) into a single weighted composite measure. The R_{TU} metric enables the aggregation of heterogeneous ES measures across different systematic reviews supporting a given OoI in tertiary evidence synthesis.

Since this process is fully automated, accuracy is assessed through comparisons with TURs. By replicating research questions from TURs in TU[®], we compare whether the R_{TU} -derived ES align with those obtained using traditional methodologies. This validation step ensures that the automated transformation process does not introduce systematic distortions and maintains consistency with established ES estimation methods. Given the diverse metrics used in TURs, such as SMD, MD, RR, OR, and HR, and the unique metric employed by TU[®], we standardized all effect sizes, used in TURs, to Cohen's d . This standardization facilitates a direct comparison, ensuring consistency in the interpretation of results.

Effect Size Comparison: Quantitative and Qualitative Approaches

ES for matched OoIs were compared using two complementary strategies: a quantitative comparison based on the numerical values of the ES, and a qualitative comparison based on categorical classification.

For the quantitative analysis, ES derived from TURs were transformed into a common metric (Cohen's d) to enable direct comparison with the continuous R_{TU} values calculated by TU° . Spearman correlation was used to assess the strength and direction of association between the two continuous measures.

For the qualitative analysis, ES were categorized into four standard interpretive levels, trivial, small, moderate, and large, based on conventional thresholds. A contingency matrix was used to evaluate the degree of concordance between TU° and TUR classifications, applying a structured system to distinguish between full concordance, partial concordance, discordance and consistent concordance, as described below.

Concordance between the two methodologies was assessed using a contingency table comparing these categorical classifications (Table 1). Based on the degree of agreement, each matched OoI was assigned to one of the following concordance levels:

Full concordance: Both TU° and TUR classified the outcome in the same category (e.g., trivial–trivial).

Partial concordance: The classifications differed by only one level (e.g., trivial–small or moderate–large). These differences were considered acceptable, as they are unlikely to lead to substantial changes in interpretation or clinical decision-making.

Discordance: The classifications differed by two or more levels (e.g., trivial–moderate or small–large), representing a more substantial methodological discrepancy.

Consistent concordance: A combined category including both full and partial concordance, reflecting a general alignment between methodologies even when minor categorical differences were present.

This categorization framework acknowledges that qualitative ES categories are artificial boundaries imposed on a continuous metric. Thus, small shifts between adjacent categories may not reflect meaningful disagreement. The concept of partial concordance accounts for this inherent imprecision in categorization and the methodological variability between synthesis approaches. Only discrepancies exceeding one categorical level were interpreted as meaningful discordances.

Table
Concordance matrix for effect size classification between TU[®] and TURs

1

TUR - ES	TU [®] - ES			
	Trivial	Small	Moderate	Large
Trivial				
Small				
Moderate				
Large				

Consistent	Full Concordance	Partial Concordance (discrepancy of one category)	Discordance (discrepancy > one category)
Concordance			

Furthermore, the certainty of the evidence (LoE) associated with each OoI was evaluated. To assess the LoE within TU[®], we employ sentiment analysis (SA) as a natural language processing (NLP) technique. This approach allows for automated classification of certainty indicators extracted from systematic review abstracts. Regarding the concern about the training data for sentiment analysis, at this stage, TU[®] employs a SA model initially trained on Twitter data. While we acknowledge that this is not highly optimized for medical texts, it provides a cost-effective starting point for sentiment classification without incurring additional expenses. Cloud-based solutions, such as Azure Cognitive Services, offer more specialized SA models for healthcare, but these are paid services that exceed our current budget.

We fully recognize the limitations of using general SA for LoE, as traditional GRADE-based approaches consider multiple factors such as study limitations, inconsistency, indirectness, publication bias, and confounders. However, our primary aim is to determine whether a simplified approach, using abstract-level sentiment classification, can yield reasonable agreement with traditional methodologies. TU[®] does not claim to replace the rigor of full-text GRADE assessments but instead seeks to evaluate whether an alternative automated method can provide valuable insights with lower resource demands. We are actively working towards developing our own custom SA algorithm tailored to medical literature, trained on validated medical datasets. This will enhance precision and improve TU[®]'s ability to evaluate LoE more effectively in the future.

For TURs, this was done using the GRADE system (Grading of Recommendations Assessment, Development and Evaluation) [23], which categorizes evidence into very low, low, moderate, and high levels. TU[®] applies a SA-based scoring system on a scale from -1 to +1. Both scales were normalized to a similar quantitative range (0-1) to facilitate comparison. To align the two systems, the SA scores are normalized using the following transformation formula: $X = (SA \text{ score} + 1) / 2$.

Where X represents the normalized TU[®] certainty score, ensuring that SA values originally in the range of -1 to 1 are mapped to a 0 to 1 scale. The GRADE ordinal levels are mapped onto the 0-1 scale as follows: Very Low = 0.00 – 0.25, Low = 0.26 – 0.50, Moderate = 0.51 – 0.75 and High = 0.76 – 1.00. This process allows TU[®] and TUR certainty ratings to be compared in a standardized manner, enabling statistical concordance analyses between the two methodologies.

Finally, the execution time of each methodology was assessed, with TU[®] providing exact time measurements and TURs relying on an estimated timeframe of 6 to 12 months based on existing literature.

Data Collection and Research Question Replication

Data collection began with a targeted search in PubMed to identify TURs in geriatrics, using the search terms "umbrella" AND "geriatric". The identified TURs served as benchmarks for our comparative analysis. The research questions from these selected umbrella reviews were directly replicated in TU[®] without modification, ensuring a precise comparison of outcomes generated by each methodology. TU[®] was configured to replicate these questions, employing automated searches and synthesis through NLP, WS, SA and ML, with human reviewer verifying and extracting data as necessary. This approach allows us to assess the comparative effectiveness and efficiency of TU[®] relative to traditional methods, particularly in identifying and analyzing outcomes critical to evaluating health interventions and exposure risks. Data from both TURs and TU[®] were systematically collected and recorded in a database to facilitate precise comparisons of outcomes, effect sizes, and other critical variables, ensuring a thorough evaluation of both methodologies.

The decision to rely solely on abstracts for data extraction in TU[®] was driven by practical and methodological considerations. Firstly, this approach was chosen due to budgetary constraints, as full-text access to all systematic reviews and meta-analyses would require extensive licensing fees or institutional subscriptions, which are beyond the scope of this project. Secondly, abstracts help mitigate language bias, since all systematic review abstracts indexed in Medline are available in English, regardless of the original language of publication. This ensures a broader and more internationally representative evidence base.

While it is acknowledged that abstracts often contain limited methodological details and may lack comprehensive information on certainty of evidence, outcome effect sizes, or risks of bias, the core

objective of TU[®] methodology is to assess whether robust conclusions can still be drawn based on abstracts alone. Recognizing the inherent limitations of abstracts, TU[®] incorporates structured extraction criteria to capture the most relevant information while acknowledging the potential risks of missing key methodological details. The validation process of TU[®] explicitly includes a comparison with full-text TURs to evaluate whether synthesis based on abstracts alone yields comparable conclusions.

Statistical Analysis: Data Analysis and Statistical Methods

The quantitative analysis compared the results obtained by both methodologies. Contingency tables were constructed to contrast the identification of outcomes and the ES, ensuring statistical congruence between the methods. For the evaluation of certainty levels, certainty scores from both methodologies were transformed to a common 0–1 scale to allow direct comparison.

The chi-squared test was used to determine significant differences between the methodologies. Additionally, Pearson or Spearman correlation analyses was conducted to quantify the relationship between TU[®] certainty levels and GRADE certainty levels of the TURs. Pearson correlation is useful for quantifying the strength and direction of a linear relationship between two continuous variables. Moreover, the analysis will be complemented with Spearman correlation due to potential violations of normality assumptions in the data.

Statistical analysis will be performed using IBM SPSS Statistics 26, with an alpha level of 0.05 to determine statistical significance.

RESULTS

Identification of Traditional Umbrella Reviews in Geriatrics as Reference Models

This section outlines the process used to identify and select traditional umbrella reviews (TURs) in the field of geriatrics, which served as the gold standard for validating the experimental methodology, The Umbrella Collaboration[®] (TU[®]).

To enable a meaningful comparison, we conducted a targeted search in PubMed aimed at identifying representative TURs in geriatrics. Rather than aiming for a comprehensive retrieval, the search strategy was deliberately focused to select TURs that would serve as valid comparators in the methodological analysis. We used the terms “umbrella” AND “geriatric”, seeking TURs that were

methodologically sound and clinically relevant for this validation study. Given the scope and objective of our analysis, a narrowly focused search was sufficient and did not require the exhaustive approach typically used in systematic reviews. Although simple in design, this strategy was intentionally crafted to meet the specific needs of the project and should not be interpreted as a limitation.

The search was performed in PubMed on March 5, 2023, using the following query:

<https://pubmed.ncbi.nlm.nih.gov/?term=umbrella+geriatric&sort=date>

The search yielded 111 references. After a detailed screening process, 75 articles were excluded for not meeting the inclusion criteria, such as being unrelated to geriatrics, not being umbrella reviews, being study protocols, or not addressing interventions or risk exposures.

From the remaining 36 eligible TURs, we randomly selected 8 for detailed comparison [24–31], giving preference to the most recent publications, which were more likely to follow current methodological standards in tertiary evidence synthesis. This selection process, summarized in Figure 3 and Table 2, ensured transparency and reproducibility in establishing TURs as reference models.

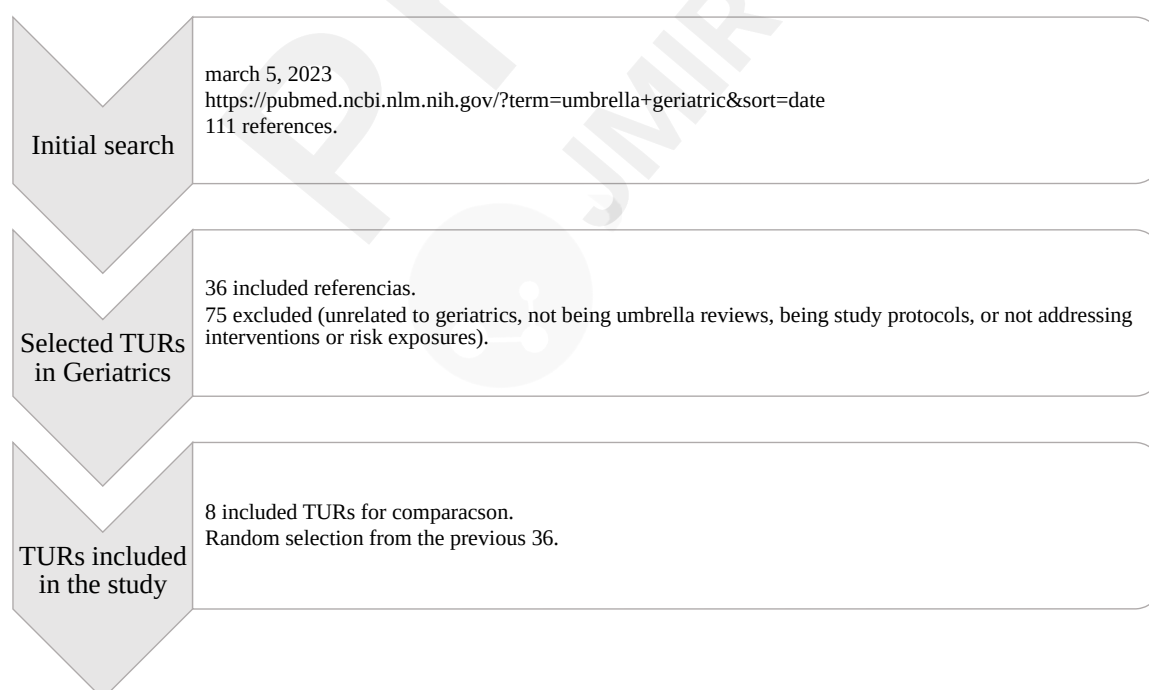


Figure 3
Flow diagram of the search and selection process for TURs in geriatrics.

Initial TUR search 36/111	TURs included for comparison 8/36 (n° project)	Author (año)	Title TUR
1	4	Conneely, 2022 [24]	The effectiveness of interventions to reduce adverse outcomes among older adults following Emergency Department discharge: umbrella review.
2		Visser, 2022 [32]	Effectiveness and characteristics of physical fitness training on aerobic fitness in vulnerable older adults: an umbrella review of systematic reviews.
3		Yuan, 2022 [33]	Modifiable predictive factors and all-cause mortality in the non-hospitalized elderly population: An umbrella review of meta-analyses.
4		Travica, 2023 [34]	Peri-Operative Risk Factors Associated with Post-Operative Cognitive Dysfunction (POCD): An Umbrella Review of Meta-Analyses of Observational Studies.
5	1	Veronese, 2023 [25]	Knee osteoarthritis and adverse health outcomes: an umbrella review of meta-analyses of observational studies.
6		Li, 2022 [35]	Zinc Intakes and Health Outcomes: An Umbrella Review.
7	3	Shen, 2022 [26]	Effects of Exercise on Patients Important Outcomes in Older People With Sarcopenia: An Umbrella Review of Meta-Analyses of Randomized Controlled Trials.
8		Beck Jepsen, 2022 [36]	Predicting falls in older adults: an umbrella review of instruments assessing gait, balance, and functional mobility.
9	5	Gazzaniga, 2023 [31]	Effect of digoxin on all-cause and cardiovascular mortality in patients with atrial fibrillation with and without heart failure: an umbrella review of systematic reviews and 12 meta-analyses.
10	6	Musazadeh, 2023 [27]	Effect of vitamin D supplementation on type 2 diabetes biomarkers: an umbrella of interventional meta-analyses.
11	7	Veronese, 2021 (1) [28]	Interventions for reducing loneliness: An umbrella review of intervention studies.
12	8	Veronese, 2021 (2) [29]	Efficacy of conservative treatments for hand osteoarthritis: An umbrella review of interventional studies.
13		Veronese, 2020 [37]	Effect of low-dose aspirin on health outcomes: An umbrella review of systematic reviews and meta-analyses.
14		Soysal, 2023 [38]	The impact of urinary incontinence on multiple health outcomes: an umbrella review of meta-analysis of observational studies.
15		Zhang, 2021 [39]	Menopausal hormone therapy and women's health: An umbrella review.
16	2	Marx, 2021 [30]	The Dietary Inflammatory Index and Human Health: An Umbrella Review of Meta-Analyses of Observational Studies.
17		Trott, 2022 [40]	Eye disease and mortality, cognition, disease, and modifiable risk factors: an umbrella review of meta-analyses of observational studies.
18		Celotto, 2021 [41]	An umbrella review of systematic reviews with meta-analyses evaluating positive and negative outcomes of Hydroxychloroquine and chloroquine therapy.
19		Demurtas, 2021 [42]	Stem cells for treatment of cardiovascular diseases: An umbrella review of randomized controlled trials.
20		Demurtas, 2020 [43]	Physical Activity and Exercise in Mild Cognitive Impairment and Dementia: An Umbrella Review of Intervention and Observational Studies.
21		Obaid, 2020	Can we prevent poststroke cognitive impairment? An umbrella review of risk factors

		[44]	and treatments.
22		Veronese, 2020	Glucosamine sulphate: an umbrella review of health outcomes.
		[45]	
23		Trott, 2021	Hearing impairment and diverse health outcomes: An umbrella review of meta-analyses of observational studies.
		[46]	
24		Neil-Sztramko, 2022	Community-based group physical activity and/or nutrition interventions to promote mobility in older adults: an umbrella review.
		[47]	
25		Antonazzo, 2022	Impact of arterial hypertension and its management strategies on cognitive function and dementia: a comprehensive umbrella review.
		[48]	
26		Demurtas, 2020	The efficacy and safety of influenza vaccination in older people: An umbrella review of evidence from meta-analyses of both observational and randomized controlled studies.
		[49]	
27		Conroy, 2020	In-hospital interventions for reducing readmissions to acute care for adults aged 65 and over: An umbrella review.
		[50]	
28		Gielen, 2021	Nutritional interventions to improve muscle mass, muscle strength, and physical performance in older people: an umbrella review of systematic reviews and meta-analyses.
		[51]	
29		Biondi-Zoccai, 2020	Oral antiplatelet therapy in the elderly undergoing percutaneous coronary intervention: an umbrella review.
		[52]	
30		Ofori-Asenso, 2020	Frailty Confers High Mortality Risk across Different Populations: Evidence from an Overview of Systematic Reviews and Meta-Analyses.
		[53]	
31		Veronese, 2019	Sarcopenia and health-related outcomes: an umbrella review of observational studies.
		[54]	
32		Beckwée, 2019	Exercise Interventions for the Prevention and Treatment of Sarcopenia. A Systematic Umbrella Review.
		[55]	
33		De Spiegeleer, 2018	Pharmacological Interventions to Improve Muscle Mass, Muscle Strength and Physical Performance in Older People: An Umbrella Review of Systematic Reviews and Meta-analyses.
		[56]	
34		Jadczak, 2018	Effectiveness of exercise interventions on physical function in community-dwelling frail older people: an umbrella review of systematic reviews.
		[57]	
35		Poscia, 2018	Effectiveness of nutritional interventions addressed to elderly persons: umbrella systematic review with meta-analysis.
		[58]	
36		Stubbs, 2015	What works to prevent falls in older adults dwelling in long term care facilities and hospitals? An umbrella review of meta-analyses of randomised controlled trials.
		[59]	

Table

TURs identified in the initial search and TURs selected for comparative analysis [24–31]

2

Identification of Outcomes of Interest

This section outlines the outcomes of interest (OoIs) identified across the eight comparative projects included in this analysis [24–31]. In the context of evidence synthesis, an OoI denotes a specific variable assessed within the included primary studies, reflecting the effect of an intervention or risk exposure across comparison groups. These outcomes may be quantitatively pooled through meta-analysis or reported individually, depending on the synthesis strategy employed. We assessed the performance of two methodologies in identifying OoIs: TURs, which served as the reference standard, and the experimental approach implemented through TU[®]. The results are presented in

three stages: first, an overview of the outcomes identified by each method; second, an analysis of concordance in OoIs between TURs and TU[®]; and finally, the identification gain factor observed with TU[®] in comparison to TURs.

The OoIs identified by each methodology across the eight comparative projects are summarized, along with the level of concordance and discordance between methods (Table 3). In total, the eight TURs included in the study identified 86 OoIs. For the same research questions, the corresponding TU[®] projects identified 410 OoIs. Of the 86 OoIs reported by TURs, TU[®] identified 73 of them (84.88%), missing 13 (15.11%) (Figure 4). Conversely, of the 410 OoIs identified by TU[®], only 73 (17.80%) were also reported by TURs, meaning 337 OoIs (82.19%) were not captured by the traditional reviews (Figure 5).

At project level, concordance rates (i.e., the percentage of TUR OoIs also identified by TU[®]) ranged from 70% in Project 4 [24] to 100% in Projects 1 [25], 5 [31], 6 [27], 7 [28], and 8 [29]. The TURs from Projects 2 [30], 3 [26], and 4 [24] showed the highest proportion of OoIs not identified by TU[®], with 21.05% (n=8), 28.57% (n=2), and 30.00% (n=3) of missed outcomes, respectively.

Table
Global summary table of OoIs identified by TURs and TU[®]: totals, matches, and discrepancies.

3

Project	1		2		3		4		5		6		7		8		Total	
	TUR	TU [®]	TUR	TU [®]	TUR	TU [®]	TUR	TU [®]	TUR	TU [®]	TUR	TU [®]	TUR	TU [®]	TUR	TU [®]	TUR	TU [®]
OoI Identified	5	12	38	71	7	34	10	22	4	25	8	149	6	49	8	48	86	410
TUR OoI identified by TU[®] (N) (%)	5	12	30	71	5	34	7	22	4	25	8	149	6	49	8	48	73	307
	(100%)		(78.94%)		(71.42%)		(70.0%)		(100%)		(100%)		(100%)		(100%)		(84.88%)	
TUR OoI NOT identified by TU[®] (N) (%)	-	-	8	-	2	-	3	-	-	-	-	-	-	-	-	-	13	103
			(21.05%)		(28.57%)		(30.00%)										(15.11%)	
TU[®] OoI identified by TUR (N) (%)	5	33	30	339	5	29	7	35	4	21	8	141	6	43	8	40	73	337
	(41.66%)		(42.25%)		(14.70%)		(31.81%)		(16.00%)		(5.36%)		(12.24%)		(16.66%)		(17.80%)	
TU[®] OoI NOT identified by TUR (N) (%)	7	377	41	368	29	315	15	313	21	23	141	140	43	45	40	397	337	73
	(58.33%)		(57.74%)		(85.29%)		(68.18%)		(84.00%)		(94.63%)		(87.75%)		(83.33%)		(82.19%)	

Figure
OoIs identified by TURs (N=86), matched/not-matched by TU®

4

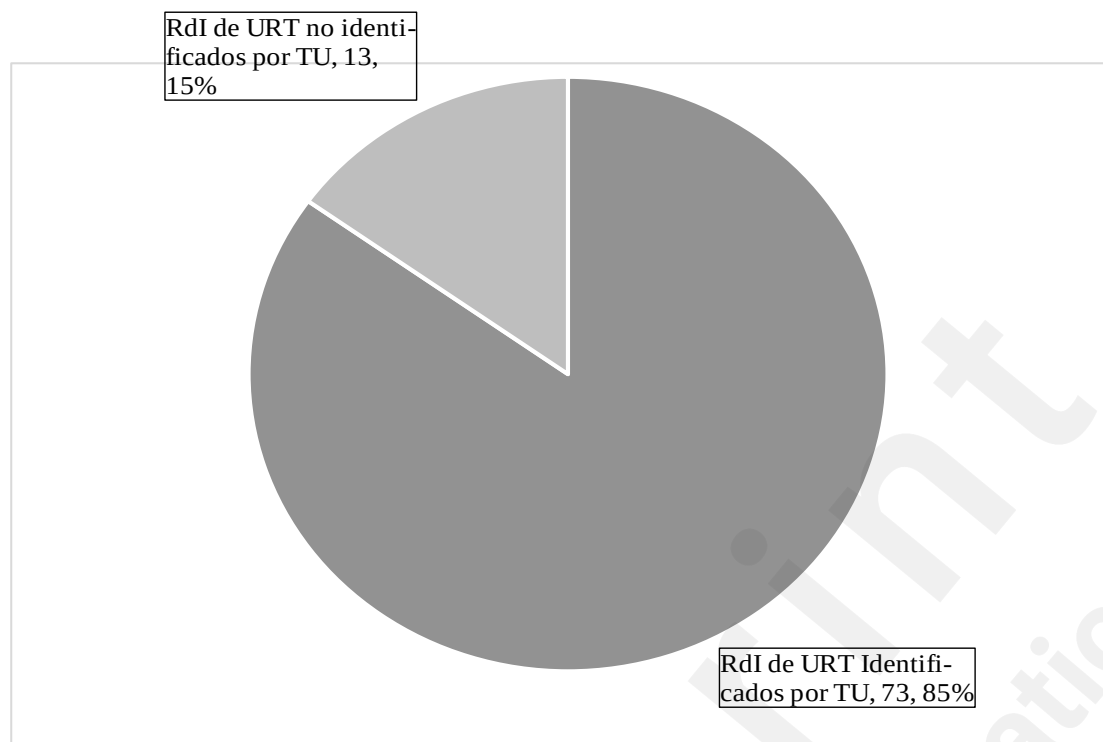
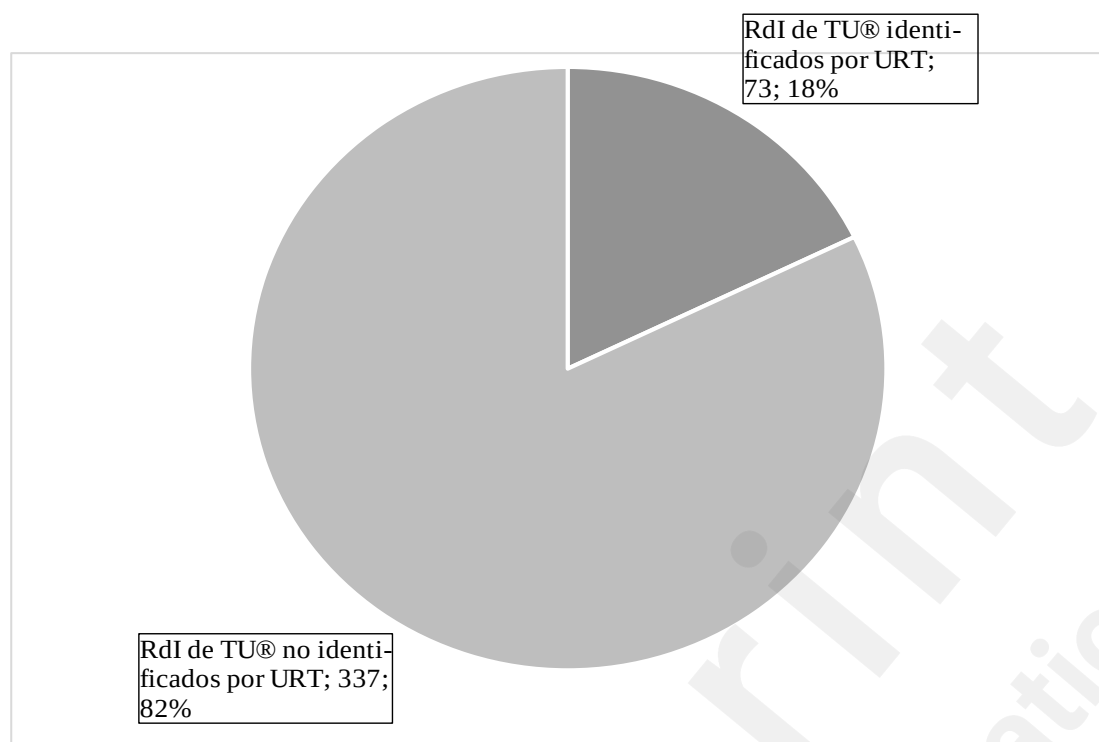


Figure
OoIs identified by TU® (410), matched/not-matched by TURs

5



The identification gain factor quantifies the extent to which TU® identifies a greater number of OoIs compared to TURs. As shown in the Table 4, this multiplier is calculated by dividing the number of OoIs identified by TU® by the number identified by the corresponding TUR.

Across all eight projects, TU® consistently outperformed TURs in the number of OoIs identified. In total, TU® identified 324 additional OoIs that were not captured by the TURs. This corresponds to an overall identification gain factor of 4.77, indicating that TU® identified nearly five times as many outcomes as TURs on average. At the project level, the gain factor ranged from 1.87 in Project 2 [30] to 18.63 in Project 6 [27].

Table
OoIs identified by TU® and TURs, additional OoIs identified by TU®, and identification gain factor

4

Project	N. ° OoI identified by TU®	N. ° OoI identified by URT	Additional OoI identified by TU®	Identification Gain Factor of OoI by TU® compared to URT
P. 1 Veronese, 2023	12	5	7	2,40
P. 2 Marx, 2021	71	38	33	1,87

P. 3 Shen, 2022	34	7	27	4,86
P. 4 Conneely, 2022	22	10	12	2,20
P. 5 Gazzaniga, 2023	25	4	21	6,25
P. 6 Musazadeh, 2023	149	8	141	18,63
P. 7 Veronese, 2021 (1)	49	6	43	8,17
P. 8 Veronese, 2021 (2)	48	8	40	6,00
TOTAL	410	86	324	4,77

Concordance in Effect Size Classification of Outcomes of Interest

To assess agreement between the two methodologies, we conducted a comparative analysis of effect sizes (ES) for OoIs identified by both TURs and TU[®]. The analysis included only those OoIs for which both methodologies reported ES using comparable standardized metrics.

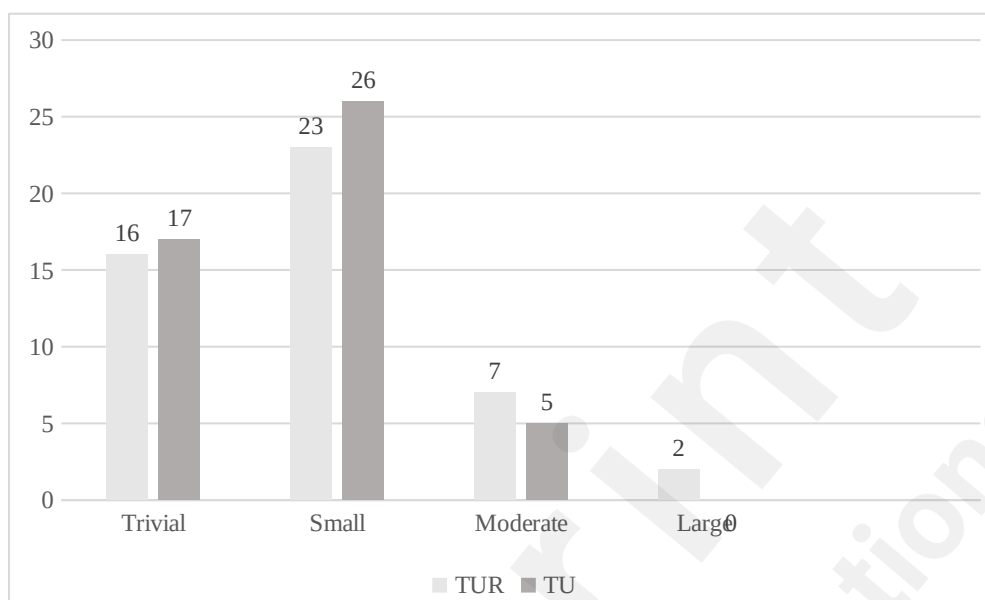
Effect sizes reported in TURs were either expressed directly as standardized mean differences (SMD) or transformed into Cohen's *d*, in order to harmonize the metrics and enable direct comparison with the corresponding ES metric generated by TU[®], R_{TU} . When TURs reported an outcome using SMD, this value was retained as-is for analysis. When other metrics were used (e.g., odds ratios or correlation coefficients), conversion to Cohen's *d* was performed using standard formulas.

Outcomes reported in TURs using mean differences (MD) or weighted mean differences (WMD) were excluded from the analysis, as sample size data were insufficient to allow proper standardization. After applying these criteria, a total of 48 matched OoIs with comparable effect sizes were included in the concordance analysis.

The Figure 6 illustrates the distribution of ES classifications by both TURs and TU[®] across the 48 OoIs included in the concordance analysis. TURs classified 33.3% ($n = 16$) of the OoIs as trivial, 47.9% ($n = 23$) as small, 14.6% ($n = 7$) as moderate, and 4.2% ($n = 2$) as large. In comparison, TU[®] classified 35.4% ($n = 17$) as trivial, 54.2% ($n = 26$) as small, 10.4% ($n = 5$) as moderate, and none as

large 0.0% (n = 0).

Figure 6
Distribution of effect size classifications by TURs and TU®



The contingency Table 5 summarizes the qualitative classification of ES for OoIs identified by both TU® and TURs, allowing for categorical-level concordance assessment between the two methodologies. Focusing on concordance from the perspective of TU® classification: TU® categorized 17 OoIs (35.4%) as having a trivial ES. Of these, 10 (58.8%) were also classified as trivial by TURs, while the remaining 7 (41.2%) were considered small. TU® classified 26 OoIs (54.2%) as having a small ES. Among these, 13 (50.0%) were also classified as small by TURs, 5 (19.2%) as trivial, 6 (23.1%) as moderate, and 2 (7.7%) as large. For the 5 OoIs (10.4%) that TU® classified as moderate, TURs concurred on only 1 (20.0%), while 3 (60.0%) were considered small and 1 (20.0%) trivial.

Table
Contingency table of effect size classification: TUR vs TU® (qualitative categories)

5

TUR - ES	TU® - ES				Total
	1. Trivial	2. Small	3. Moderate	4. Large	
1. Trivial	58.8% (10)	19.2% (5)	20.0% (1)	-	33.3% (16)
2. Small	41.2% (7)	50.0% (13)	60.0% (3)	-	47.9% (23)
3. Moderate	-	23.1% (6)	20.0% (1)	-	14.6% (7)
4. Large	-	7.7% (2)	-	-	4.1% (2)
Total	35.4% (17)	54.2% (26)	10.4% (5)	-	(48)
$\chi^2 = 11.03^+$ (p = 0.088), V = 0.339 (Moderate)					
Notes.					
^{NS} Not significant, ⁺ Marginally significant (p<.10), * Significant (p<.05) y ** Highly significant (p<.01)					
Statistical test: χ^2 - Pearson's chi-squared. Effect size measure: Cramer's V					
Full Concordance		Partial Concordance		Discordance	

	(Discrepancy of one category)	(Discrepancy of more than one category)
--	-------------------------------	---

Of the 48 OoIs analyzed, full concordance in effect size classification between TU[®] and TURs was observed in 24 cases (50.0%). Partial concordance, defined as a one-level difference in qualitative ES category, was observed in 21 cases (43.8%). Only 3 cases (6.3%) showed discordance, defined as a discrepancy of more than one category (Table 6) (Figure 7). The total number of consistent classifications (full or partial concordance) was 45 out of 48 cases, representing 93.8% of the sample.

Table

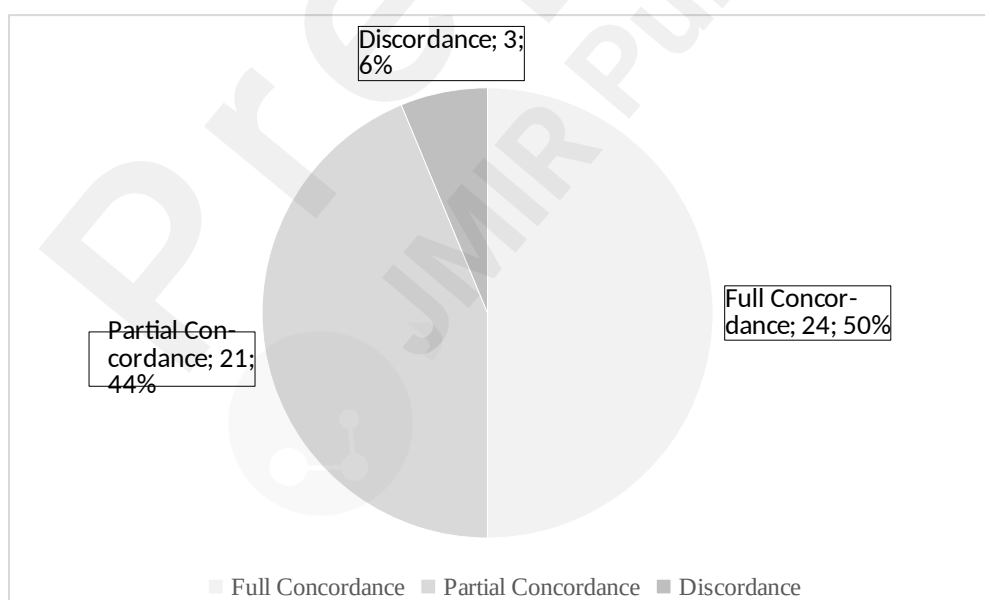
Concordance levels in effect size classification: frequency and cumulative distribution

6

	Absolute Frequency	Cumulative Absolute Frequency	Relative Frequency	Cumulative Relative Frequency
Different ES	24	24	50.0%	50.0%
Same ES	24	48	50.0%	100%
Same ES	24	24	50.0%	50.0%
One-Level	21	45	43.8%	93.8%
Discrepancy				
Discrepancy > 1 Level	3	48	6.3%	100%

Figure 7

Levels of concordance in effect size classification between TU[®] and TURs across 48 matched outcomes of interest



Correlation of Quantitative Effect Sizes Between TU[®] and TURs

This section presents the correlation analysis between the ES reported by TURs and those generated by TU[®], using continuous, quantitative values. The goal is to assess the strength and direction of the relationship between both metrics, that is, whether increases in one are associated with increases

(positive correlation) or decreases (negative correlation) in the other.

A nonparametric Spearman correlation was used to quantify the association (Table 8). The resulting coefficient ($\rho = 0.399$, $P = .005$) indicates a statistically significant positive correlation of low to moderate strength. Given the nonparametric nature of the test, the analysis is robust to outliers and non-normal distributions, which enhances the reliability of the observed association (Table 7). This supports the consistency of the TU[®] ES metric (R_{TU}) relative to the standardized Cohen's d and SMD.

To complement the analysis, a scatter plot was generated (Figure 8), displaying TU[®] ES on the x-axis and TUR ES on the y-axis. A linear regression line was added to visualize the expected relationship between both measures. Although the coefficient of determination ($R^2 = 0.111$) indicates modest explanatory power, the overall trend supports a positive association between the two methodologies.

Table
Normality assessment of TU[®] and TUR effect size distributions using the Shapiro–Wilk test

7

	TU [®] - ES (R_{TU})	TUR - ES (Cohen's d)
Shapiro-Wilk W	.874	.847
Shapiro-Wilk P value	<.001**	<.001**

^{NS} Not significant, [†] Marginally significant ($p < .10$), * Significant ($p < .05$),
** Highly significant ($p < .01$)

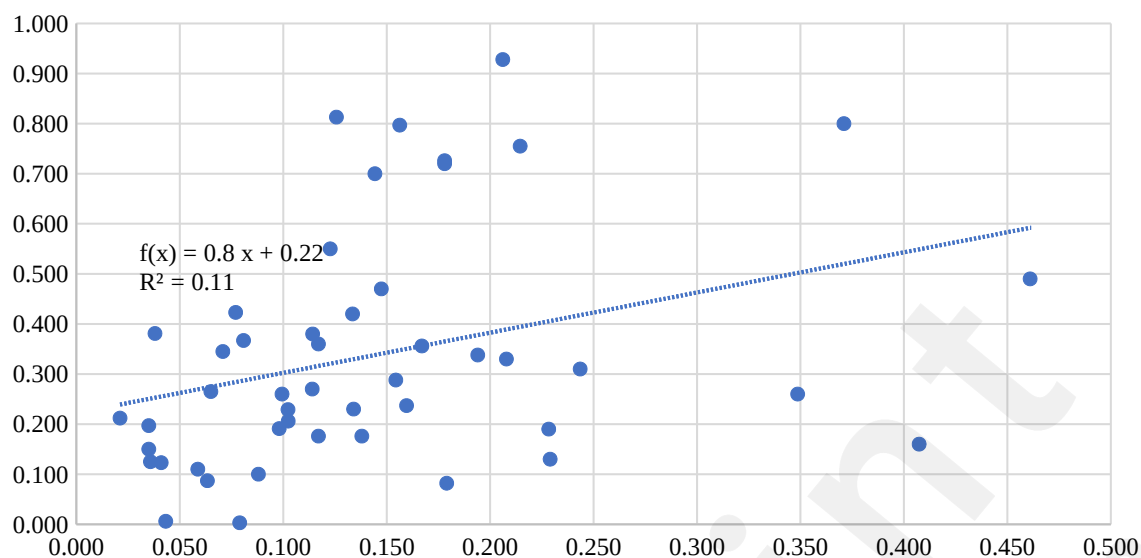
Table
Spearman correlation between TU[®] and TUR quantitative effect sizes

8

Statistic	Value (Statistic P value)	Significance	Strength
Spearman's Rho	.399** (.005)	Highly Significant	Low-Moderate

^{NS} Not significant, [†] Marginally significant ($p < .10$), * Significant ($p < .05$), ** Highly significant ($p < .01$)

Figure

Correlation of quantitative effect sizes between TU[®] and TURs

Concordance in Transformed Certainty Scores Between TU[®] and TURs

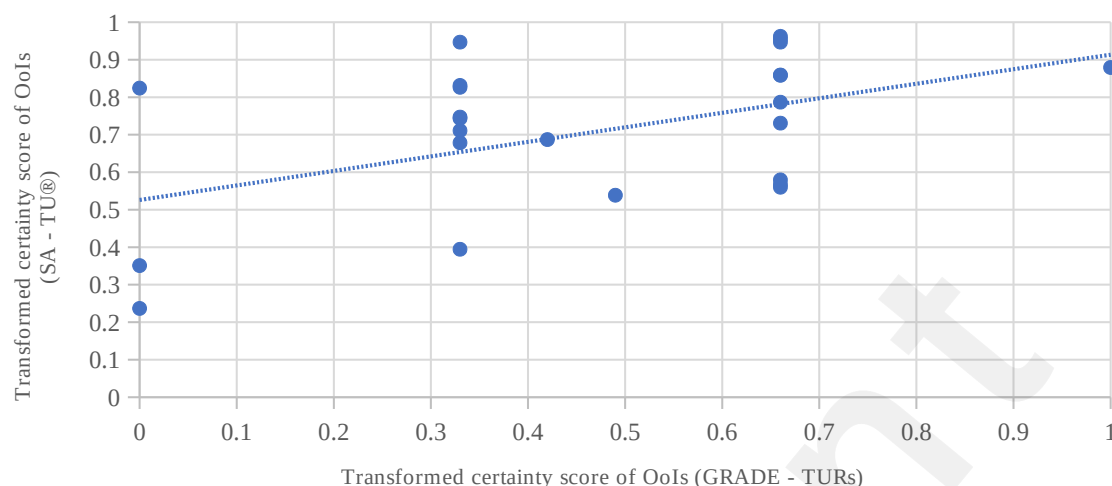
This analysis aims to assess the concordance between two distinct approaches for evaluating the certainty of OoIs in tertiary evidence synthesis. TURs rely on the GRADE framework to assess the certainty of evidence, whereas TU[®] employs an automated SA applied to the abstracts of the included SR/MA as a proxy for certainty. To enable a valid comparison between the two methodologies, both GRADE ratings and TU[®] sentiment scores were transformed to a common numerical scale ranging from 0 to 1. All values used in this analysis are referred to as transformed certainty scores. Only the TURs that explicitly applied the GRADE approach were eligible for comparison. These included five of the eight TURs selected for this study: Shen (2023) [26], Conneely (2022) [24], Gazzaniga (2023) [31], and both reviews by Veronese (2021) [28,29]. From these, a total of 25 transformed certainty scores were extracted for comparative analysis.

To visualize the relationship between the two certainty estimation methods, a scatter plot was generated (Figure 9). This figure displays the transformed certainty scores for OoIs as assessed by TURs using the GRADE framework (x-axis) and by TU[®] using sentiment analysis (y-axis). Each point represents a matched OoI, allowing for direct comparison between methodologies across a standardized 0-to-1 scale.

Figure

9

Scatter plot of transformed certainty scores for outcomes of interest ($N = 25$) assessed by TURs (GRADE) and TU[®] (sentiment analysis)



To assess the association between the certainty scores provided by the two methodologies, a Spearman correlation analysis was conducted using the transformed certainty scores for OoIs derived from TURs (GRADE) and TU[®] (sentiment analysis). Given the results of the normality test (Table 9), which showed non-normal distributions for both sets of scores (Shapiro–Wilk $P = .005$ for TURs; $P = .049$ for TU[®]), a nonparametric approach was applied. The Spearman correlation coefficient was 0.446 ($P = .025$), indicating a statistically significant positive correlation of moderate strength (Table 10). These results suggest a consistent and meaningful relationship between the transformed certainty scores obtained through GRADE and those estimated by TU[®], reinforcing the overall concordance observed between the two approaches.

Table

9

Shapiro–Wilk test for normality of transformed certainty scores (GRADE vs TU[®])

	Transformed GRADE Certainty Value (TUR)	Transformed Certainty Value (TU [®])
Shapiro-Wilk W Statistic	0.874	0.928
Shapiro-Wilk P value	0.005	0.049

Table

10

Spearman correlation between transformed certainty scores assigned by TURs and TU[®]

Spearman Correlation			
N	Coefficient	P Value	Result
25	0.446	.025	Statistically significant positive correlation of moderate strength

25	0.446	0.025	Significant
----	-------	-------	-------------

Execution Time of TU[®] Projects

This section reports the execution time required to complete tertiary evidence synthesis projects using TU[®] methodology. Execution time could not be evaluated for TURs, as this information is not reported in any of the selected reviews, nor is it systematically documented in the umbrella review literature. Although expert consensus and international guidelines suggest a typical timeframe of 6 to 12 months for completing a TUR, this information is insufficient for direct comparison.

As a result, our analysis focuses exclusively on the recorded execution times for the eight TU[®] projects included in this study. A summary is presented in the Table 11. The mean execution time was 4 hours and 46 minutes (± 2 hours and 30 minutes), with a median of 4 hours and 36 minutes. There was moderate variability across projects, with the shortest time recorded for the Gazzaniga, 2023 [31] project at 1 hour and 34 minutes, and the longest for Musazadeh, 2023 [27] at 10 hours and 8 minutes.

Table
Project execution time, search parameters, and reference overlap between TU[®] and TUR methodologies

11

Project	Search Date	TU [®] Execution Time (hh:mm:ss)	N [°] . of TU [®] Search Terms	N [°] of References in TUR	N [°] of References in TU [®]	Reference Overlap (%)
P. 1 Veronese, 2023	01 Aug. 2023	2:37:27	16	5	8	100
P. 2 Marx, 2021	01 May 2023	4:54:57	60	15	54	100
P. 3 Shen 2022	01 May 2023	5:29:22	19	6	18	100
P. 4 Conneely, 2022	01 May 2023	5:37:14	6	16	23	100
P. 5 Gazzaniga, 2023	01 May 2023	1:34:37	15	11	19	90.9
P. 6 Musazadeh, 2023	12 Nov. 2023	10:08:24	41	37	90	97.2
P. 7 Veronese, 2021 (1)	12 Dec. 2023	4:17:43	53	7	35	100
P. 8 Veronese 2021 (2)	23 Dec. 2023	3:33:54	16	8	22	88.8

ETHICAL CONSIDERATIONS

This study does not involve human participants, personal data, or animals. All data will be sourced from published materials, and the analyses will be conducted in accordance with established ethical standards for secondary data analysis.

DISCUSSION

The findings of this study support the validity of The Umbrella Collaboration[®] (TU[®]) as a robust and efficient methodology for tertiary evidence synthesis. TU[®] demonstrated a high level of concordance with traditional umbrella reviews (TURs) in identifying outcomes of interest (OoIs), successfully replicating the majority of those reported in TURs while also identifying a substantial number of additional outcomes. The experimental ES metric employed by TU[®] (R_{TU}) showed a statistically significant correlation with standardized measures such as Cohen's *d* and SMD, reinforcing its methodological soundness for estimating the magnitude of interventions or exposures. In addition, TU[®]'s automated sentiment-based certainty assessment exhibited an acceptable level of agreement with GRADE, the most widely adopted system for evaluating certainty in evidence synthesis. One of the most notable advantages of TU[®] lies in its execution time. While TURs typically require several months to complete, TU[®] was able to generate comprehensive results within hours. This efficiency, combined with its demonstrated consistency and methodological coherence, highlights its potential value in settings where rapid evidence synthesis is essential or where resources are limited.

The development of TU[®] aligns with the growing interest in AI-assisted tools for evidence synthesis, an emerging field that is likely to shape the future of scientific research. Rather than replacing TURs, TU[®] is intended to complement them by offering a methodologically sound alternative for contexts where conventional approaches may be impractical or untimely. As a semi-automated platform, TU[®] leverages natural language processing, sentiment analysis, and machine learning to streamline tertiary synthesis while maintaining methodological rigor. Despite its potential, TU[®] represents a paradigm shift that may require time and educational efforts for widespread understanding and adoption. During the initial dissemination of this project, a general unfamiliarity with automated tertiary synthesis was observed, highlighting the need for clearer conceptual frameworks and greater engagement with the research community.

The identification of OoIs represents a central axis in the comparison between TU[®] and TURs. In the context of tertiary evidence synthesis, an OoI reflects a specific finding previously evaluated in one or more systematic reviews with or without meta-analysis (SR/MA) and integrated to answer a defined research question. The validity of a synthesis model depends not only on aggregating relevant data but also on faithfully capturing the main conclusions emerging from the primary

literature.

Due to the limited literature on tertiary synthesis, this analysis draws from well-established principles in secondary synthesis, where many of the same strengths and limitations apply. These include challenges in outcome selection, variability in inclusion criteria, and the balance between breadth and precision.

The comparative analysis revealed substantial differences in the number and breadth of OoIs identified. Across the eight matched projects, TURs identified a total of 86 OoIs, whereas TU[®] detected 410, yielding an identification gain factor of 4.77. TU[®] replicated 84.88% of the OoIs reported by TURs and contributed 337 additional OoIs. This difference does not appear to stem solely from broader search strategies or inclusion criteria. According to the Cochrane Handbook [60], umbrella reviews can pursue either broad or narrow questions [61]. While both methodologies support both approaches, several TURs in this study applied restrictive filters. For example, Project 6 [27] limited both its search strategy and inclusion criteria, while Projects 3 [26] and 4 [24] applied inclusion restrictions. Nevertheless, even when only unrestricted TURs are considered, TU[®] maintained a gain factor of 4.93, suggesting an intrinsic property of TU[®] in capturing a greater volume of outcomes.

One explanation lies in TU[®]'s automated and systematic extraction process, which minimizes human bias and ensures that any OoI mentioned in an abstract is captured, regardless of its perceived importance. In contrast, traditional reviews often apply subjective judgment in prioritizing outcomes, potentially omitting secondary findings. Moreover, TU[®] is not constrained by human workload, allowing it to process larger volumes of data and extract multiple OoIs per review without sacrificing comprehensiveness.

A second consideration is the differing approach to quality assessment. TURs often prioritize high-quality SRs/MAs, limiting scope but potentially increasing rigor. TU[®], by relying on abstracts, does not perform formal quality appraisal. While this design choice increases the number of included studies, it should not be interpreted as a methodological flaw. Notably, up to 30% of SRs/MAs fail to report study quality assessments, and only 12.9% set quality thresholds for meta-analysis inclusion [62,63]. Furthermore, tools such as PRISMA-Abstract could be integrated in future versions of TU[®] to address this limitation, though prior research has shown that abstracts often score low on PRISMA

quality indicators [64–67]. Nonetheless, TU[®] consistently identifies more OoIs based solely on abstracts than TURs do using full-text reviews.

Another factor is the role of cognitive bias in outcome selection. Human reviewers may unconsciously favor familiar or hypothesis-confirming outcomes, limiting variability. While TU[®] still involves human oversight in final output selection, its core processes are software-driven, applying uniform parameters across all included studies and reducing cognitive distortions. For instance, whereas TURs tended to report outcomes only in one direction, TU[®] captured both positive and negative effects, aligning with best-practice guidelines from the Joanna Briggs Institute and Cochrane Collaboration, which emphasize the need to present balanced results for informed decision-making [11,60].

Finally, while it is possible that TU[®] identifies outcomes of limited clinical relevance, this risk appears minimal. All OoIs extracted by TU[®] are derived from published SRs/MAs indexed in MEDLINE, which are assumed to meet minimum standards of relevance and methodological quality. According to the Cochrane Handbook [61], reviews should focus on critical and important outcomes, excluding trivial ones. Future research could further assess the clinical importance of the additional OoIs identified by TU[®] to strengthen its practical applicability.

The study examined the concordance and correlation of ES of the OoIs metrics between TU[®] and TURs, using a unified set of outcomes assessed by both methodologies. TU[®] incorporates a novel metric, R_{TU} , specifically developed to standardize and synthesize effect sizes from SRs/MAs within a tertiary synthesis framework. Unlike conventional metrics such as Cohen's d or the SMD, R_{TU} was designed to allow direct aggregation of heterogeneous ES measures and requires empirical validation. This analysis represents the first formal evaluation of R_{TU} against traditional metrics.

In the qualitative comparison of ES categories (trivial, small, moderate, large), TU[®] demonstrated a high level of agreement with TURs. Full concordance, defined as exact categorical match between the two methods, was observed in 50% of cases, while consistent concordance (full plus one-level deviation) reached 93.8%. Only 6.3% of cases showed discordance, meaning the categorical classifications differed by more than one level. These findings suggest that TU[®] approximates the classification logic of TURs with a high degree of reliability, reinforcing the potential of R_{TU} to align with established interpretive frameworks.

The strength of the association between ES categorizations assigned by each methodology was evaluated using Cramér's V, which yielded a value of 0.339. This indicates a moderate association and supports the notion that TU[®] and TURs tend to classify the strength of associations in a comparable manner. Although the chi-square test did not reach statistical significance, likely due to the limited sample size and sparse contingency cells, Cramér's V remains a valid effect size measure and is less sensitive to these limitations [68]. In applied research, the strength of association can often be more informative than statistical significance, particularly when evaluating the practical utility of an experimental method [69]. It is worth noting that the statistical power of the chi-square test may have been reduced by the small number of observations and the presence of empty cells in the contingency table, both of which are known to limit test sensitivity. In line with recent critiques of overreliance on P values in biomedical research [70–73], the observed concordance rates and effect size association offer more meaningful insights into the comparability of methods than the statistical significance alone.

The quantitative comparison of ES values further supports the validity of TU[®]. The Spearman correlation coefficient between TU[®] and TUR values was positive and statistically significant, indicating a consistent directional relationship. Additionally, the slope of the regression line approached 1.0 ($\beta = 0.8$), suggesting that the R_{TU} metric yields ES of similar magnitude to those reported in TURs using standard metrics. This represents a marked improvement in proportionality compared to earlier regression slopes reported in other validation contexts.

However, the coefficient of determination (R^2) remained modest ($R^2 = 0.11$), indicating that a substantial portion of the variance in TUR ES is not explained by TU[®] estimates. This implies that while TU[®] captures general trends in effect magnitude, additional factors, such as differences in underlying metric types, weighting procedures, or reviewer judgments, may influence the exact values. It also highlights the limitations of linear models in fully explaining the relationship between these two methods.

These findings provide preliminary but encouraging evidence for the validity of R_{TU} as a useful metric for tertiary synthesis. Despite its experimental nature, R_{TU} appears to closely approximate established metrics such as Cohen's d and the SMD in both categorical classification and magnitude. Future studies will be needed to further evaluate the sensitivity, specificity, and contextual

performance of R_{TU} across diverse synthesis domains. Nonetheless, this initial validation suggests that $TU^{\circledR}$ can offer a consistent and efficient alternative for evaluating and categorizing ES in large-scale evidence synthesis.

In this study, certainty assessments were compared exclusively among those TURs that applied the GRADE framework, a well-established and widely endorsed method for evaluating the certainty of evidence [74]. Although other appraisal tools exist, such as the statistical grading criteria proposed by Papatheodorou [75] or the methodological assessment framework AMSTAR-2 [76], these differ conceptually from the certainty model implemented in $TU^{\circledR}$. Papatheodorou's system, despite its structure, has been criticized for relying on arbitrary statistical thresholds and lacking sensitivity to clinical relevance and risk of bias [77]. AMSTAR-2, in turn, focuses on the methodological quality of the SRs/MAs themselves, rather than the certainty of individual outcomes. In contrast, $TU^{\circledR}$'s approach aligns more closely with GRADE principles by estimating certainty at the outcome level, albeit through an innovative method based on automated sentiment analysis of SR/MA abstracts. Originally, the GRADE system referred to "quality of evidence," which denoted the degree of confidence in the validity of study findings to inform clinical decisions. Over time, this evolved to "confidence in estimates," and more recently, to "certainty of evidence", a term now widely accepted as it better reflects both the trust in effect estimates and their applicability in practice [23,78].

The comparison between transformed certainty scores standardized to a 0–1 scale for outcomes evaluated by $TU^{\circledR}$ and those rated with GRADE provides important insights into the validity of $TU^{\circledR}$'s automated approach. The observed positive and statistically significant correlation suggests that $TU^{\circledR}$ is capable of approximating GRADE-based judgments, thereby reinforcing the conceptual alignment between the two methodologies. However, given the limited number of comparable data points, these results should be interpreted cautiously. The small sample size restricts generalizability and highlights the need for further studies to validate this relationship across broader contexts and larger datasets. Despite this limitation, the findings support increased confidence in $TU^{\circledR}$ as a tool for researchers and end users. While not intended to replace GRADE, $TU^{\circledR}$ offers a complementary approach that leverages automation and processing speed without substantially compromising the quality of certainty evaluation. This alignment with GRADE principles strengthens $TU^{\circledR}$'s position as an innovative yet methodologically grounded tool in the evolving landscape of evidence synthesis. More broadly, the correlation observed invites further exploration into the potential integration of automated text-based analyses with established appraisal frameworks such as GRADE. Such

integration may offer a promising avenue for optimizing both the efficiency and reliability of certainty assessments. As these technologies evolve, ongoing validation will be essential to define their appropriate role in scientific practice.

Execution time is a critical but often overlooked dimension in the evaluation of evidence synthesis methodologies. TURs rarely document their execution timelines, and this omission was consistent among the TURs included in this study. In contrast, TU[®] methodology records execution time precisely, from the initiation of the first search term to the completion of the first version of the synthesis project, excluding any subsequent update. This level of detail in time tracking is unprecedented in both secondary and tertiary synthesis and represents a methodological innovation that may become increasingly relevant with the integration of automation tools. As a result, traditional timelines measured in months or years may soon be replaced by more accurate, time-stamped process metrics.

Due to the lack of concrete time data from TURs, direct comparisons were not feasible. However, the execution times recorded for TU[®] are sufficiently notable to highlight its efficiency. TU[®] projects were completed within a range of 1.5 to 10 hours, substantially shorter than the typical duration for TURs, which is estimated between 6 and 12 months. This is also well below the average timeframes reported for SRs/MAs, which commonly take 1 to 2 years [79], or for accelerated alternatives such as rapid reviews, ranging from 5 to 12 weeks [80], and even the Two-Week SR approach, which targets completion in 11 working days [81]. The remarkable speed of TU[®] is a key strength, particularly in light of growing international demand for agile evidence synthesis methods. Organizations such as the WHO [82], Cochrane Collaboration [83], and Joanna Briggs Institute [15] have emphasized the need for rapid yet valid synthesis approaches, an objective that TU[®] appears to fulfill effectively in the context of tertiary reviews.

It is important to note that the time measured in this study refers solely to the initial execution phase of TU[®] projects and does not account for later updates. However, TU[®] includes a novel updating system that keeps projects active beyond their initial completion. The platform automatically checks daily for new SRs/MAs matching the predefined search criteria. When new studies are identified, a human reviewer is notified and can assess their eligibility. If approved, the system integrates the new results into the existing synthesis following revision and confirmation. While no equivalent system currently exists in traditional tertiary synthesis, the need for regular updates is well documented in

evidence-based practice. In SRs/MAs, updates are typically recommended every two years or in response to emerging evidence, yet compliance with this standard is often poor [84–87]. TU[®]'s automated updating mechanism represents a major advantage in maintaining the relevance and accuracy of results in fast-evolving fields.

Notably, projects with a higher number of included references required longer execution times, which is expected given the additional workload involved in processing more SRs/MAs. Although TU[®] significantly streamlines the synthesis process, the complexity and volume of information remain factors that influence total execution time, particularly when human oversight is still involved. Looking ahead, TU[®] is being developed to support full automation of SR/MA data analysis, eliminating the need for human intervention. This transition will depend on training algorithms through semi-automated projects to progressively improve accuracy via machine learning.

Interestingly, the number of search terms used did not appear to correlate with execution time (unpublished data), suggesting that TU[®]'s algorithms are capable of managing variable search complexity efficiently. In traditional methodologies, the search phase is often complex and performed by specialized information professionals using advanced Boolean strategies across multiple databases [88–90]. In contrast, TU[®] simplifies the process by using basic search terms only in PubMed, automatically generated and refined through generative AI. This shift towards simplified, automated search strategies opens new avenues for exploring the adequacy and efficiency of non-complex methods in evidence synthesis. It also raises questions about the continued necessity of highly complex search protocols, suggesting that AI-assisted models may offer comparable effectiveness in certain contexts [91–93].

Ultimately, the short execution times observed with TU[®] not only improve the efficiency of tertiary synthesis, but also enhance the feasibility of rapidly delivering high-level evidence. In settings where timeliness is essential, this temporal advantage may outweigh minor trade-offs in other methodological dimensions. This capability holds particular promise for public health and healthcare decision-making, where fast access to up-to-date, high-quality evidence is crucial. TU[®] therefore emerges as a promising tool in scenarios where speed is a priority, offering a pragmatic balance between rapid synthesis and analytical depth, while still requiring further validation against established gold standards.

LIMITATIONS

One of the primary limitations of this comparative methodological study lies in the relatively small number of umbrella reviews (TURs) included in the analysis. The comparison was based on only eight TURs, randomly selected from an initial pool of 36 eligible reviews. Although the selection aimed to obtain a representative sample of TURs within the field of geriatrics, the randomization procedure was not formally documented, introducing the possibility of selection bias. Furthermore, the limited number of TURs may not fully capture the methodological diversity and topic variability typically observed across umbrella reviews, nor does it guarantee comprehensive coverage of the range of clinical areas within geriatrics. Nevertheless, it is worth noting that all selected TURs were recently published, between 2021 and 2023. This recency represents a significant advantage, as older umbrella reviews, particularly in a relatively new field like tertiary synthesis, tend to show greater methodological heterogeneity and reduced alignment with current standards of evidence synthesis practice.

Another notable limitation is the absence of blinding of the lead investigator to the selected TURs. As the same researcher who conducted the comparative analysis was also responsible for selecting and extracting data from the TURs, prior knowledge of their findings may have inadvertently influenced the outcomes generated using TU[®]. Specifically, awareness of the TUR results could have affected decisions made during the TU[®] analysis, potentially introducing bias toward known outcomes. However, while this lack of blinding represents a methodological weakness, several mitigating factors should be considered. The TURs included in this study comprised multiple OoIs with diverse metrics, effect sizes, statistical significance levels, directions of effect, and certainty ratings. This high degree of heterogeneity makes it unlikely that prior knowledge could have systematically influenced all aspects of the TU[®] analyses. Moreover, TU[®] relies heavily on algorithm-guided, semi-automated processes for data extraction and interpretation, which introduces an additional layer of neutrality. Although the investigator had access to the TUR designs and outcomes, the selection and presentation of results in TU[®] were primarily determined by predefined algorithmic rules, limiting the extent to which cognitive bias could have meaningfully distorted the findings.

The decision to utilize only one database, Medline via PubMed, in TU[®] is both a recognized limitation and a deliberate choice shaped by resource constraints and technical considerations. While

systematic reviews typically require searching multiple databases to capture all relevant literature [94], our approach focuses on testing whether TU[®] can achieve outcomes comparable to TURs despite these limitations. While a pilot study provided preliminary evidence of PubMed's strong coverage, further validation through TU[®] is necessary to confirm its applicability across different domains. Furthermore, while searching multiple databases is often recommended to avoid language and indexing biases, especially those related to non-English literature [95,96]. TU[®] mitigates some of these biases by focusing on abstracts in English, as all PubMed abstracts are provided in this language regardless of the original publication's language. However, the absence of Chinese databases in our approach is a notable limitation, given that only a small proportion of Chinese journals are indexed in Medline.

The decision to rely exclusively on abstracts rather than full texts in TU[®] is a deliberate methodological choice aligned with the core objective of this study: to evaluate whether tertiary evidence synthesis can be conducted efficiently, with fewer resources, and without requiring extensive methodological expertise from the reviewer, while still producing results comparable to TURs. We acknowledge that abstracts may lack key methodological details, including certainty of evidence assessments, effect size calculations, and risk of bias evaluations. However, the fundamental hypothesis of TU[®] is that an automated, structured approach to abstract-based synthesis may still yield clinically useful conclusions, particularly when applied under standardized and reproducible conditions. Moreover, relying on abstracts offers two key advantages:

1. Minimization of language bias, all systematic review abstracts in Medline are available in English, regardless of the original publication language. This ensures that non-English studies are not automatically excluded due to language barriers, a common issue in traditional systematic reviews.
2. Feasibility and accessibility, full-text access to all systematic reviews requires extensive licensing fees and institutional subscriptions, which may not always be feasible. Abstracts provide a universally accessible data source, allowing for broader implementation of evidence synthesis methodologies.

While we acknowledge the potential benefits of testing TU[®] across multiple medical domains, the current study is a pilot study designed to assess the feasibility of our methodology in a single, well

defined field, geriatrics. At this stage, limiting the scope to geriatrics is a strategic choice, ensuring that the study maintains clarity, feasibility, and methodological rigor in its initial validation phase. This focused approach allows for a controlled evaluation of TU[®] against TURs, ensuring that the initial validation is conducted under clearly defined conditions. Future studies should expand its application to other medical fields, such as cardiovascular medicine and psychiatry, as well as non-medical domains, including education, sociology, and other disciplines with abundant systematic reviews. This stepwise approach ensures a methodologically sound progression, allowing TU[®] to be tested and refined incrementally before broader implementation.

While TU[®] does not seek to replace the depth and comprehensiveness of full-text review, the present study demonstrates that a structured synthesis based solely on abstracts can yield results that are sufficiently coherent and robust to serve as a complementary or alternative approach. The findings support the feasibility of this method in contexts where full-text access is limited or where rapid evidence synthesis is required, highlighting TU[®] as a potentially valuable tool in such scenarios.

CONCLUSION

This study provides preliminary evidence supporting the validity of The Umbrella Collaboration[®] (TU[®]) as a complementary semi-automated tool for tertiary evidence synthesis. In a set of comparative projects, TU[®] demonstrated a high level of concordance with traditional umbrella reviews (TURs) in identifying outcomes of interest, estimating effect sizes, and assessing certainty, while offering substantially shorter execution times.

Although the methodology remains experimental and further validation is required across broader contexts, the results suggest that semi-automated approaches such as TU[®] may represent a promising step toward more efficient, scalable, and continuously updatable models of evidence synthesis. TU[®] does not aim to replace traditional methods but to provide a practical alternative in settings where time, resources, or responsiveness are critical.

Future studies with larger datasets, enhanced blinding procedures, and expanded topic areas will be essential to confirm these findings and further explore the potential role of TU[®] in the evolving landscape of evidence-based research.

FUNDING

This project has received a research grant from the Fundación Alfonso X el Sabio.

CONFLICTS OF INTEREST

The authors Beltrán Carrillo (BC), Rubinos-Cuadrado (MR), Parellada-Martin (JP), Palacios-López (AP), and Carrillo-Rubinos (BCR) were involved in the development of The Umbrella Collaboration[®] (TU[®]) software. Additionally, BC and MR are owners of the software and may have financial or intellectual property interests in its future applications.

While this study was conducted with the aim of validating TU[®] as an evidence synthesis tool, the authors have taken all necessary precautions to ensure that the study design, data analysis, and interpretation of results remain as objective and methodologically rigorous as possible. To mitigate potential biases, comparisons were conducted against gold-standard TURs, and statistical methodologies were employed to ensure transparency and replicability of findings.

Furthermore, the affiliations of the authors did not influence the study design, execution, data analysis, or interpretation of results. All research activities were conducted following rigorous methodological standards to ensure impartiality and scientific integrity.

REFERENCES:

- [1] Evidence synthesis for policy A STATEMENT OF PRINCIPLES 2018.
- [2] World Health Organization, 2005. bridging-the-know-do-gap.pdf n.d.
- [3] Cottrell E, Whitlock E, Kato E, Uhl S, Belinson S, Chang C, et al. Defining the Benefits of Stakeholder Engagement in Systematic Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014.
- [4] Sørensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: A systematic review and integration of definitions and models. *BMC Public Health* 2012;12:80. <https://doi.org/10.1186/1471-2458-12-80>.
- [5] Martínez-García JJ, Canizalez-Román A, Velázquez-Román JA, Flores-Villaseñor HM, León-Sicaños NM. Evaluación del conocimiento de métodos básicos de epidemiología e investigación en médicos residentes. *Rev Médica Univ Autónoma Sinaloa REVMEDUAS* 2022;11:115–23.
- [6] Baccolini V, Rosso A, Di Paolo C, Isonne C, Salerno C, Migliara G, et al. What is the Prevalence of Low Health Literacy in European Union Member States? A Systematic Review and Meta-analysis. *J Gen Intern Med* 2021;36:753–61. <https://doi.org/10.1007/s11606-020-06407-8>.
- [7] Baumann LA, Reinhold AK, Brütt AL. Public and patient involvement in health policy decision-making on the health system level – A scoping review. *Health Policy* 2022;126:1023–38. <https://doi.org/10.1016/j.healthpol.2022.07.007>.
- [8] Biondi-Zoccai G, editor. Umbrella Reviews: Evidence Synthesis with Overviews of Reviews and Meta-Epidemiologic Studies. Cham: Springer International Publishing; 2016. <https://doi.org/10.1007/978-3-319-25655-9>.
- [9] Choi GJ, Kang H. Introduction to Umbrella Reviews as a Useful Evidence-Based Practice. *J Lipid Atheroscler* 2023;12:3–11. <https://doi.org/10.12997/jla.2023.12.1.3>.
- [10] Fusar-Poli P, Radua J. Ten simple rules for conducting umbrella reviews. *BMJ Ment Health* 2018;21:95–100. <https://doi.org/10.1136/ebmental-2018-300014>.
- [11] Aromataris E, Lockwood C, Porritt K, Pilla B. JBI Manual for Evidence Synthesis - JBI Global Wiki n.d. <https://jbi-global-wiki.refined.site/space/MANUAL> (accessed June 5, 2024).
- [12] Cant R, Ryan C, Kelly MA. A nine-step pathway to conduct an umbrella review of literature. *Nurse Author Ed* 2022;32:31–4. <https://doi.org/10.1111/nae2.12039>.
- [13] Belbasis L, Brooker RD, Zavalis E, Pezzullo AM, Axfors C, Ioannidis JP. Mapping and systematic appraisal of umbrella reviews in epidemiological research: a protocol for a meta-epidemiological study. *Syst Rev* 2023;12:123. <https://doi.org/10.1186/s13643-023-02265-7>.
- [14] Pollock M. Chapter V: Overviews of Reviews 2023. <https://training.cochrane.org/handbook/current/chapter-v> (accessed May 26, 2024).
- [15] Tricco AC, Straus SE, Ghaffar A, Langlois EV. Rapid reviews for health policy and systems decision-making: more important than ever before. *Syst Rev* 2022;11:153. <https://doi.org/10.1186/s13643-022-01887-7>.
- [16] Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Syst Rev* 2023;12:72. <https://doi.org/10.1186/s13643-023-02243-z>.
- [17] Teperikidis L, Boulmpou A, Papadopoulos C, Biondi-Zoccai G. Using ChatGPT to perform a systematic review: a tutorial. *Minerva Cardiol Angiol* 2024. <https://doi.org/10.23736/S2724-5683.24.06568-2>.
- [18] ChatGPT n.d. <https://chatgpt.com/c/e4d561fe-1745-4b09-8196-d2a99b2f2661> (accessed August 15, 2024).
- [19] Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, et al. The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. *JMIR Med Inform* 2024;12:51187. <https://doi.org/10.2196/51187>.
- [20] Kothamali PR, Karne VK, Dandyala SSM. Integrating AI and Machine Learning in Quality

Assurance for Automation Engineering. *Int J Res Publ Semin* 2024;15:93–102. <https://doi.org/10.36676/jrps.v15.i3.1445>.

[21] Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction-the why, what, when, and how. *J Clin Epidemiol* 2017;91:23–30. <https://doi.org/10.1016/j.jclinepi.2017.08.010>.

[22] 201912_LSR_Revised_Guidance.pdf n.d.

[23] 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;336:924–6. <https://doi.org/10.1136/bmj.39489.470347.AD>.

[24] Conneely M, Leahy S, Dore L, Trépel D, Robinson K, Jordan F, et al. The effectiveness of interventions to reduce adverse outcomes among older adults following Emergency Department discharge: umbrella review. *BMC Geriatr* 2022;22:462. <https://doi.org/10.1186/s12877-022-03007-5>.

[25] Veronese N, Honvo G, Bruyère O, Rizzoli R, Barbagallo M, Maggi S, et al. Knee osteoarthritis and adverse health outcomes: an umbrella review of meta-analyses of observational studies. *Aging Clin Exp Res* 2023;35:245–52. <https://doi.org/10.1007/s40520-022-02289-4>.

[26] Shen Y, Liu D, Li S, He Y, Tan F, Sun X, et al. Effects of Exercise on Patients Important Outcomes in Older People With Sarcopenia: An Umbrella Review of Meta-Analyses of Randomized Controlled Trials. *Front Med* 2022;9. <https://doi.org/10.3389/fmed.2022.811746>.

[27] Musazadeh V, Kavyani Z, Mirhosseini N, Dehghan P, Vajdi M. Effect of vitamin D supplementation on type 2 diabetes biomarkers: an umbrella of interventional meta-analyses. *Diabetol Metab Syndr* 2023;15:76. <https://doi.org/10.1186/s13098-023-01010-3>.

[28] Veronese N, Galvano D, D'Antiga F, Vecchiato C, Furegon E, Allocco R, et al. Interventions for reducing loneliness: An umbrella review of intervention studies. *Health Soc Care Community* 2021;29:e89–96. <https://doi.org/10.1111/hsc.13248>.

[29] Veronese N, Smith L, Bolzetta F, Cester A, Demurtas J, Punzi L. Efficacy of conservative treatments for hand osteoarthritis. *Wien Klin Wochenschr* 2021;133:234–40. <https://doi.org/10.1007/s00508-020-01702-0>.

[30] Marx W, Veronese N, Kelly JT, Smith L, Hockey M, Collins S, et al. The Dietary Inflammatory Index and Human Health: An Umbrella Review of Meta-Analyses of Observational Studies. *Adv Nutr* 2021;12:1681–90. <https://doi.org/10.1093/advances/nmab037>.

[31] Gazzaniga G, Menichelli D, Scaglione F, Farcomeni A, Pani A, Pastori D. Effect of digoxin on all-cause and cardiovascular mortality in patients with atrial fibrillation with and without heart failure: an umbrella review of systematic reviews and 12 meta-analyses. *Eur J Clin Pharmacol* 2023;79:473. <https://doi.org/10.1007/s00228-023-03470-y>.

[32] Visser D, Wattel EM, Gerrits KHL, Wouden JC van der, Meiland FJM, Groot AJ de, et al. Effectiveness and characteristics of physical fitness training on aerobic fitness in vulnerable older adults: an umbrella review of systematic reviews. *BMJ Open* 2022;12:e058056. <https://doi.org/10.1136/bmjopen-2021-058056>.

[33] Yuan Y, Lin S, Lin W, Huang F, Zhu P. Modifiable predictive factors and all-cause mortality in the non-hospitalized elderly population: An umbrella review of meta-analyses. *Exp Gerontol* 2022;163:111792. <https://doi.org/10.1016/j.exger.2022.111792>.

[34] Travica N, Lotfaliany M, Marriott A, Safavynia SA, Lane MM, Gray L, et al. Peri-Operative Risk Factors Associated with Post-Operative Cognitive Dysfunction (POCD): An Umbrella Review of Meta-Analyses of Observational Studies. *J Clin Med* 2023;12:1610. <https://doi.org/10.3390/jcm12041610>.

[35] Li J, Cao D, Huang Y, Chen B, Chen Z, Wang R, et al. Zinc Intakes and Health Outcomes: An Umbrella Review. *Front Nutr* 2022;9. <https://doi.org/10.3389/fnut.2022.798078>.

[36] Beck Jepsen D, Robinson K, Ogliari G, Montero-Odasso M, Kamkar N, Ryg J, et al. Predicting falls in older adults: an umbrella review of instruments assessing gait, balance, and functional mobility. *BMC Geriatr* 2022;22:615. <https://doi.org/10.1186/s12877-022-03271-5>.

- [37] Veronese N, Demurtas J, Thompson T, Solmi M, Pesolillo G, Celotto S, et al. Effect of low-dose aspirin on health outcomes: An umbrella review of systematic reviews and meta-analyses. *Br J Clin Pharmacol* 2020;86:1465–75. <https://doi.org/10.1111/bcp.14310>.
- [38] Soysal P, Veronese N, Ippoliti S, Pizzol D, Carrie AM, Stefanescu S, et al. The impact of urinary incontinence on multiple health outcomes:: an umbrella review of meta-analysis of observational studies. *Aging Clin Exp Res* 2023;35:479–95. <https://doi.org/10.1007/s40520-022-02336-0>.
- [39] Zhang G-Q, Chen J-L, Luo Y, Mathur MB, Anagnostis P, Nurmatov U, et al. Menopausal hormone therapy and women's health: An umbrella review. *PLOS Med* 2021;18:e1003731. <https://doi.org/10.1371/journal.pmed.1003731>.
- [40] Trott M, Smith L, Veronese N, Pizzol D, Barnett Y, Gorely T, et al. Eye disease and mortality, cognition, disease, and modifiable risk factors: an umbrella review of meta-analyses of observational studies. *Eye* 2022;36:369–78. <https://doi.org/10.1038/s41433-021-01684-x>.
- [41] Celotto S, Veronese N, Barbagallo M, Ometto F, Smith L, Pardhan S, et al. An umbrella review of systematic reviews with meta-analyses evaluating positive and negative outcomes of Hydroxychloroquine and chloroquine therapy. *Int J Infect Dis* 2021;103:599–606. <https://doi.org/10.1016/j.ijid.2020.12.018>.
- [42] Demurtas J, Fanelli GN, Romano SL, Solari M, Yang L, Soysal P, et al. Stem cells for treatment of cardiovascular diseases: An umbrella review of randomized controlled trials. *Ageing Res Rev* 2021;67:101257. <https://doi.org/10.1016/j.arr.2021.101257>.
- [43] Demurtas J, Schoene D, Torbahn G, Marengoni A, Grande G, Zou L, et al. Physical Activity and Exercise in Mild Cognitive Impairment and Dementia: An Umbrella Review of Intervention and Observational Studies. *J Am Med Dir Assoc* 2020;21:1415-1422.e6. <https://doi.org/10.1016/j.jamda.2020.08.031>.
- [44] Obaid M, Douiri A, Flach C, Prasad V, Marshall I. Can we prevent poststroke cognitive impairment? An umbrella review of risk factors and treatments. *BMJ Open* 2020;10:e037982. <https://doi.org/10.1136/bmjopen-2020-037982>.
- [45] Veronese N, Demurtas J, Smith L, Reginster J-Y, Bruyère O, Beaudart C, et al. Glucosamine sulphate: an umbrella review of health outcomes. *Ther Adv Musculoskelet Dis* 2020;12:1759720X20975927. <https://doi.org/10.1177/1759720X20975927>.
- [46] Trott M, Smith L, Xiao T, Veronese N, Koyanagi A, Jacob L, et al. Hearing impairment and diverse health outcomes. *Wien Klin Wochenschr* 2021;133:1028–41. <https://doi.org/10.1007/s00508-021-01893-0>.
- [47] Neil-Sztramko SE, Teggart K, Moore C, Sherifali D, Fitzpatrick-Lewis D, Coletta G, et al. Community-based group physical activity and/or nutrition interventions to promote mobility in older adults: an umbrella review. *BMC Geriatr* 2022;22:539. <https://doi.org/10.1186/s12877-022-03170-9>.
- [48] Antonazzo B, Marano G, Romagnoli E, Ronzoni S, Frati G, Sani G, et al. Impact of arterial hypertension and its management strategies on cognitive function and dementia: a comprehensive umbrella review. *Minerva Cardiol Angiol* 2022;70:285–97. <https://doi.org/10.23736/S2724-5683.20.05452-3>.
- [49] Demurtas J, Celotto S, Beaudart C, Sanchez-Rodriguez D, Balci C, Soysal P, et al. The efficacy and safety of influenza vaccination in older people: An umbrella review of evidence from meta-analyses of both observational and randomized controlled studies. *Ageing Res Rev* 2020;62:101118. <https://doi.org/10.1016/j.arr.2020.101118>.
- [50] Conroy T, Heuzenroeder L, Feo R. In-hospital interventions for reducing readmissions to acute care for adults aged 65 and over: An umbrella review. *Int J Qual Health Care* 2020;32:414–30. <https://doi.org/10.1093/intqhc/mzaa064>.
- [51] Gielen E, Beckwée D, Delaere A, De Breucker S, Vandewoude M, Bautmans I, et al. Nutritional interventions to improve muscle mass, muscle strength, and physical performance in older people: an umbrella review of systematic reviews and meta-analyses. *Nutr Rev* 2021;79:121–

47. <https://doi.org/10.1093/nutrit/nuaa011>.

[52] Biondi-Zoccai G, Antonazzo B, Giordano A, Versaci F, Frati G, Ronzoni S, et al. Oral antiplatelet therapy in the elderly undergoing percutaneous coronary intervention: an umbrella review. *J Thorac Dis* 2020;12. <https://doi.org/10.21037/jtd.2019.12.87>.

[53] Ofori-Asenso R, Chin KL, Sahle BW, Mazidi M, Zullo AR, Liew D. Frailty Confers High Mortality Risk across Different Populations: Evidence from an Overview of Systematic Reviews and Meta-Analyses. *Geriatrics* 2020;5:17. <https://doi.org/10.3390/geriatrics5010017>.

[54] Veronese N, Demurtas J, Soysal P, Smith L, Torbahn G, Schoene D, et al. Sarcopenia and health-related outcomes: an umbrella review of observational studies. *Eur Geriatr Med* 2019;10:853–62. <https://doi.org/10.1007/s41999-019-00233-w>.

[55] Beckwée D, Delaere A, Aelbrecht S, Baert V, Beaudart C, Bruyere O, et al. Exercise Interventions for the Prevention and Treatment of Sarcopenia. A Systematic Umbrella Review. *J Nutr Health Aging* 2019;23:494–502. <https://doi.org/10.1007/s12603-019-1196-8>.

[56] De Spiegeleer A, Beckwée D, Bautmans I, Petrovic M, Bautmans I, Beaudart C, et al. Pharmacological Interventions to Improve Muscle Mass, Muscle Strength and Physical Performance in Older People: An Umbrella Review of Systematic Reviews and Meta-analyses. *Drugs Aging* 2018;35:719–34. <https://doi.org/10.1007/s40266-018-0566-y>.

[57] Jadczyk AD, Makwana N, Luscombe-Marsh N, Visvanathan R, Schultz TJ. Effectiveness of exercise interventions on physical function in community-dwelling frail older people: an umbrella review of systematic reviews. *JBIS Evid Synth* 2018;16:752. <https://doi.org/10.11124/JBISRIR-2017-003551>.

[58] Poscia A, Milovanovic S, La Milia DI, Duplaga M, Grysztar M, Landi F, et al. Effectiveness of nutritional interventions addressed to elderly persons: umbrella systematic review with meta-analysis. *Eur J Public Health* 2018;28:275–83. <https://doi.org/10.1093/eurpub/ckx199>.

[59] Stubbs B, Denking MD, Brefka S, Dallmeier D. What works to prevent falls in older adults dwelling in long term care facilities and hospitals? An umbrella review of meta-analyses of randomised controlled trials. *Maturitas* 2015;81:335–42. <https://doi.org/10.1016/j.maturitas.2015.03.026>.

[60] Higgins J. *Cochrane Handbook for Systematic Reviews of Interventions* 2023. <https://training.cochrane.org/handbook> (accessed May 26, 2024).

[61] McKenzie JE, Brennan SE. Overviews of systematic reviews: great promise, greater challenge. *Syst Rev* 2017;6:185. <https://doi.org/10.1186/s13643-017-0582-8>.

[62] Luchini C, Veronese N, Nottegar A, Shin JJ, Gentile G, Granzio U, et al. Assessing the quality of studies in meta-research: Review/guidelines on the most important quality assessment tools. *Pharm Stat* 2021;20:185–95. <https://doi.org/10.1002/pst.2068>.

[63] Seehra J, Pandis N, Koletsi D, Fleming PS. Use of quality assessment tools in systematic reviews was varied and inconsistent. *J Clin Epidemiol* 2016;69:179–184.e5. <https://doi.org/10.1016/j.jclinepi.2015.06.023>.

[64] Bigna JJR, Um LN, Nansseu JRN. A comparison of quality of abstracts of systematic reviews including meta-analysis of randomized controlled trials in high-impact general medicine journals before and after the publication of PRISMA extension for abstracts: a systematic review and meta-analysis. *Syst Rev* 2016;5:174. <https://doi.org/10.1186/s13643-016-0356-8>.

[65] Li T, Hua F, Dan S, Zhong Y, Levey C, Song Y. Reporting quality of systematic review abstracts in operative dentistry: An assessment using the PRISMA for Abstracts guidelines. *J Dent* 2020;102:103471. <https://doi.org/10.1016/j.jdent.2020.103471>.

[66] El Ansari W, AlRumaihi K, El-Ansari K, Arafa M, Elbardsi H, Majzoub A, et al. Reporting quality of abstracts of systematic reviews/meta-analyses: An appraisal of Arab Journal of Urology across 12 years: the PRISMA-Abstracts checklist. *Arab J Urol* 2023;21:52–65. <https://doi.org/10.1080/2090598X.2022.2113127>.

[67] Zhong Y, Wang Y, Dan S, Zhao T, Li T, Qin D, et al. THE REPORTING QUALITY OF

SYSTEMATIC REVIEW ABSTRACTS IN LEADING GENERAL DENTAL JOURNALS: A METHODOLOGICAL STUDY. *J Evid-Based Dent Pract* 2023;23:101831. <https://doi.org/10.1016/j.jebdp.2022.101831>.

[68] Brzezińska J. The Problem of Zero Cells in the Analysis of Contingency Tables. *Zesz Nauk Univ Ekon W Krakowie* 2015;49–61. <https://doi.org/10.15678/ZNUEK.2015.0941.0504>.

[69] Sapra RL, Saluja S. Understanding statistical association and correlation. *Curr Med Res Pract* 2021;11:31. https://doi.org/10.4103/cmrip.cmrip_62_20.

[70] Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. *J Grad Med Educ* 2012;4:279–82. <https://doi.org/10.4300/JGME-D-12-00156.1>.

[71] Demidenko E. The p-Value You Can't Buy. *Am Stat* 2016;70:33–8. <https://doi.org/10.1080/00031305.2015.1069760>.

[72] Greenland S. Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values. *Am Stat* 2019;73:106–14. <https://doi.org/10.1080/00031305.2018.1529625>.

[73] LU Y, BELITSKAYA-LEVY I. The debate about p-values. *Shanghai Arch Psychiatry* 2015;27:381–5. <https://doi.org/10.11919/j.issn.1002-0829.216027>.

[74] Clinical practice guidelines: Concepts, limitations and challenges 2020. <https://www.medwave.cl/revisiones/metodoinvestreport/7887.html> (accessed October 20, 2024).

[75] Papatheodorou S. Umbrella reviews: what they are and why we need them. *Eur J Epidemiol* 2019;34:543–6. <https://doi.org/10.1007/s10654-019-00505-6>.

[76] 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. <https://doi.org/10.1136/bmj.j4008>.

[77] Schlesinger S, Schwingshackl L, Neuenschwander M, Barbaresko J. A critical reflection on the grading of the certainty of evidence in umbrella reviews. *Eur J Epidemiol* 2019;34:889–90. <https://doi.org/10.1007/s10654-019-00531-4>.

[78] Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;87:4. <https://doi.org/10.1016/j.jclinepi.2017.05.006>.

[79] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017;7:e012545. <https://doi.org/10.1136/bmjopen-2016-012545>.

[80] Hamel C, Michaud A, Thuku M, Skidmore B, Stevens A, Nussbaumer-Streit B, et al. Defining Rapid Reviews: a systematic scoping review and thematic analysis of definitions and defining characteristics of rapid reviews. *J Clin Epidemiol* 2021;129:74–85. <https://doi.org/10.1016/j.jclinepi.2020.09.041>.

[81] Clark J, Glasziou P, Mar CD, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. *J Clin Epidemiol* 2020;121:81–90. <https://doi.org/10.1016/j.jclinepi.2020.01.008>.

[82] Tricco AC, Langlois Etienne V, Straus SE, Alliance for Health Policy and Systems Research, World Health Organization. Rapid reviews to strengthen health policy and systems: a practical guide. Geneva: World Health Organization; 2017.

[83] Garritty C, Gartlehner G, Nussbaumer-Streit B, King VJ, Hamel C, Kamel C, et al. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. *J Clin Epidemiol* 2021;130:13–22. <https://doi.org/10.1016/j.jclinepi.2020.10.007>.

[84] Garner P, Hopewell S, Chandler J, MacLehose H, Akl EA, Beyene J, et al. When and how to update systematic reviews: consensus and checklist. *BMJ* 2016;354:i3507. <https://doi.org/10.1136/bmj.i3507>.

[85] Hoffmeyer B, Fonnes S, Andresen K, Rosenberg J. Use of inactive Cochrane reviews in academia: A citation analysis. *Scientometrics* 2023;128:2923–34. <https://doi.org/10.1007/s11192->

023-04691-9.

- [86] Beller EM, Chen JK-H, Wang UL-H, Glasziou PP. Are systematic reviews up-to-date at the time of publication? *Syst Rev* 2013;2:36. <https://doi.org/10.1186/2046-4053-2-36>.
- [87] Bashir R, Surian D, Dunn AG. Time-to-update of systematic reviews relative to the availability of new evidence. *Syst Rev* 2018;7:195. <https://doi.org/10.1186/s13643-018-0856-9>.
- [88] Heinen L, Goossen K, Lunny C, Hirt J, Puljak L, Pieper D. The optimal approach for retrieving systematic reviews was achieved when searching MEDLINE and Epistemonikos in addition to reference checking: a methodological validation study. *BMC Med Res Methodol* 2024;24:271. <https://doi.org/10.1186/s12874-024-02384-2>.
- [89] Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst Rev* 2017;6:245. <https://doi.org/10.1186/s13643-017-0644-y>.
- [90] Aromataris E, Riitano D. Systematic Reviews: Constructing a Search Strategy and Searching for Evidence. *AJN Am J Nurs* 2014;114:49. <https://doi.org/10.1097/01.NAJ.0000446779.99522.f6>.
- [91] Yu X, Wu S, Sun Y, Wang P, Wang L, Su R, et al. Exploring the diverse definitions of 'evidence': a scoping review. *BMJ Evid-Based Med* 2024;29:37–43. <https://doi.org/10.1136/bmjebm-2023-112355>.
- [92] Guimarães NS, Joviano-Santos JV, Reis MG, Chaves RRM, Observatory of Epidemiology N Health Research (OPENS). Development of search strategies for systematic reviews in health using ChatGPT: a critical analysis. *J Transl Med* 2024;22:1. <https://doi.org/10.1186/s12967-023-04371-5>.
- [93] Ge L, Agrawal R, Singer M, Kannapiran P, De Castro Molina JA, Teow KL, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. *Syst Rev* 2024;13:269. <https://doi.org/10.1186/s13643-024-02682-2>.
- [94] Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res Synth Methods* 2020;11:181–217. <https://doi.org/10.1002/jrsm.1378>.
- [95] Jia Y, Huang D, Wen J, Wang Y, Rosman L, Chen Q, et al. Assessment of Language and Indexing Biases Among Chinese-Sponsored Randomized Clinical Trials. *JAMA Netw Open* 2020;3:e205894. <https://doi.org/10.1001/jamanetworkopen.2020.5894>.
- [96] Mao C, Li M. Language Bias Among Chinese-Sponsored Randomized Clinical Trials in Systematic Reviews and Meta-analyses—Can Anything Be Done? *JAMA Netw Open* 2020;3:e206370. <https://doi.org/10.1001/jamanetworkopen.2020.6370>.

Supplementary Files

TOC/Feature image for homepages

The Umbrella Collaboration.

