Amount and certainty of evidence in Cochrane systematic reviews of interventions: a large-scale meta-research study

Thomas Starck, PhD[1, 2], Philippe Ravaud, MD, PhD[1, 2], Isabelle Boutron, MD, PhD[1, 2, 3]

December 3, 2025

1 Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Centre for Research in Epidemiology and Statistics (CRESS), Paris, France

2 Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, Paris, France

3 Cochrane France

Corresponding authors

Thomas Starck, PhD

Centre for Research in Epidemiology and Statistics (CRESS)

Hôpital Hôtel Dieu, 1 place du Parvis Notre-Dame, 75004 Paris, France

Email: thomas.starck@polytechnique.org

Summary

What is already known on this topic

- High quality up-to-date evidence synthesis is essential for decision-makers.
- Confidence in the evidence informing decision-making can be limited by the amount and quality of primary research on a specific research question

What this study adds

- This large-scale meta-research study analyzed all Cochrane "Summary of findings" tables (i.e., 64,849 population-intervention-comparison-outcome – PICO – questions), and found that about two thirds of the PICO questions were informed by two or fewer studies, with a median [Q1-Q3] of 123 [0-557] participants per PICO; the associated certainty of evidence was rated as high in only 4% of the cases.
- After an update of the review (i.e., 7,461 PICOs), 63% PICOs did not include additional studies, and 71% showed no change in certainty of evidence; upgrades and downgrades of certainty occurred at similar frequencies.
- These results question the efficiency of the research ecosystem.

ABSTRACT

Objectives: To quantify the amount and certainty of evidence in Cochrane systematic reviews of interventions, and to describe how this evidence has evolved over time.

Design: Large-scale meta-research study

Data source: Cochrane Database of Systematic Reviews (search date April 8, 2025)

Eligibility criteria: Cochrane systematic reviews assessing interventions reporting "Summary of findings" tables.

Data extraction: Data were automatically extracted using web scraping and a large language model, with quality control performed by humans on a random sample.

Analysis: We describe the certainty of evidence for each population-intervention-comparison-outcome (PICO) question reported in all Cochrane "Summary of findings" tables. When available, we compared the certainty of evidence between the initial version and the latest update.

Results: We identified 5,116 reviews that reported a "Summary of findings" table, containing 64,849 PICO questions.

Overall, 24% (n = 15,768) of PICOs had no study included, 31% (n = 20,390) included only 1 study, 14% (n = 8,796) 2 studies, and 31% (n = 19,895) more than 2 studies. Nearly all PICOs (97%) only included randomized trials. The median [Q1-Q3] number of included participants was 123 [0-557]. The certainty of evidence was rated as high for 4% (n = 2,852), moderate for 16% (n = 10,574), low for 27% (n = 17,409), very low for 26% (n = 17,012), and not assessed for 26% (n = 17,002).

Of the 7,461 PICO questions with an update (median time to update of 4.3 years [Q1-Q3: 2.6-

6.4]), the number of included studies in the latest update remained the same for 63%; the certainty of evidence was unchanged for 71%; upgraded for 13% and downgraded for 15%.

Conclusion: The amount and certainty of evidence is low and has not improved over time with review updates. These results question the efficiency of the research ecosystem.

Abbreviations: LLM: large language model; AI: artificial intelligence; PICO: population, intervention, comparison, outcome; RCT: randomized controlled trial

INTRODUCTION

Clinical research generates an ever-increasing volume of scientific knowledge. It is estimated that in 2020 around 35,000 new reports of clinical trials were indexed in PubMed [1]. Given the scale of this information overload, evidence-based decision making requires access to rigorous, reliable, and up-to-date syntheses of research findings. As stated by Donnelly et al. in Nature, "An accurate, concise and unbiased synthesis of the available evidence is arguably one of the most valuable contributions a research community can offer decision-makers" [2].

Cochrane, an international independent, non-profit collaboration created in 1993, produces high-quality evidence syntheses through rigorous methods and quality controls to support decision making. Cochrane has a priority setting process to ensure the relevance of Cochrane reviews to the needs of patients, the public, healthcare providers and policymakers. The collaboration has published more than 9,000 reviews which are widely used for decision making, with 66% of new World Health Organization guidelines in the past five years citing at least one Cochrane review [3].

However, confidence in the results of a systematic review and meta-analysis can be limited by the availability and quality of primary research on a specific research question [4-7]. The Grading of Recommendations Assessment, Development and Evaluation framework (GRADE) has become the reference standard for communicating how much certainty decision makers can place in evidence syntheses [8]. For each outcome, it classifies the evidence into four levels of certainty: high, moderate, low, and very low. First presented in 2004 [9], then adopted by Cochrane in 2008 [10,11], evaluation of the certainty of evidence through GRADE has been progressively incorporated into Cochrane systematic

reviews, and has been mandatory since the mid-2010s as part of a “Summary of findings” table [12].

This study aimed to quantify the amount and certainty of evidence in Cochrane systematic reviews of interventions, and to describe how this evidence has evolved over time.

METHODS

Study design

We conducted a large-scale meta-research study to describe the evidence in Cochrane systematic reviews of interventions.

Eligibility Criteria

We included all Cochrane systematic reviews of interventions (i.e., reviews assessing the effectiveness and/or safety of a treatment, vaccine, device, preventive measure, procedure or policy) that reported a "Summary of findings" table. We excluded other types of reviews such as diagnostic test accuracy reviews, prognosis study reviews, qualitative evidence syntheses, methodology reviews, overviews of reviews, rapid reviews and prototype reviews. We also excluded network meta-analyses.

Search strategy

We developed an R-based pipeline to automatically scrape all the versions (initial version and subsequent updates where available) of all the systematic reviews indexed in the Cochrane Database of Systematic Reviews [13], their associated metadata and, where available, their "Summary of findings" tables (Figure 1, Supplementary Figures S11 and S12).

The scraping was launched on April 8, 2025, and reflects the state of the Cochrane Database of Systematic Reviews at this date.

Data retrieved through web scraping

For each review version, the following data were automatically extracted during our web scraping:

- metadata: identifier, version, title, URL, official publication date, and dates of revision events (e.g., "new search performed", "conclusion changed");
- "Search Methods" text from the abstracts: Cochrane abstracts are always structured with the same subsections, which include a "Search Methods" section. We retrieved this section during our web scraping. This was then processed by a large language model (LLM) to extract the literature search dates.

Data extraction assisted by a large language model (LLM)

We extracted the following information, along with a record of its source, using an LLM: 1) the search date of each version of each systematic review, 2) information reported in the "Summary of findings" tables (population, intervention, comparison, outcome [PICO] questions; certainty of evidence; number of studies; and number of participants). In addition, when previous review versions reporting a "Summary of findings" table were identified, we used an LLM to match the PICO questions between the most recent update and the initial version with a "Summary of findings" table. The specific procedures for each task are detailed below.

<u>Search date of the systematic reviews</u>

The publication date of the reviews can differ substantially from the search date (Supplementary Figures S13 and S14). We extracted the search date from the "Search Methods" section of each

abstract, with GPT-o4-mini-high. If the LLM extraction returned an empty value (for instance, because a search date was not mentioned in the abstract), we relied on the review version metadata (i.e., “new search” date) when available; as a last resort, we relied by default on the official publication date as an approximation.

We optimized prompts to automatically extract the search month and year with GPT-o4-mini-high on a training set of 600 review "Search Methods" sections manually extracted by two specifically trained medical students, each record being extracted by a single reviewer. We used few-shot prompting with 3 representative examples provided to guide the model (see Supplementary Section B.3, C.2).

This optimized prompt was then validated on an independent testing set of 600 additional search dates, manually extracted by the two reviewers with each record extracted by a single reviewer. Overall, the LLM-assisted extraction exactly matched the manually extracted month–year in 84% of the 600 cases; 98% were within ±1 year, and errors ranged from −3.0 to +1.7 years.

<u>“Summary of findings” table: PICO question, and quantity and certainty of evidence for each outcome</u>

In Cochrane reviews, a “Summary of findings” table is dedicated to a single combination of population-intervention-comparison (PIC); within the table, each line reports the number of studies and certainty of evidence for a single outcome (meaning that each line of the table represents a unique combination of the PICO elements that defines a clinical question). Several “Summary of findings” tables can be reported within a single review. When a “Summary of

findings” table was reported, we extracted the following information (see Supplementary Figure S12 for the structure of a “Summary of findings” table on the Cochrane website):

- PICO questions: population, settings, intervention, comparison, and outcome.
- The number of studies and participants included for the evaluation of each PICO reported in each “Summary of findings” table.
- Whether the number of studies refers to randomized controlled trials (RCTs) or non-randomized studies; this information could also be extracted from the abstracts section "Selection Criteria" if needed.
- The certainty of the evidence for each outcome reported in each “Summary of findings” table. The certainty of evidence was rated as high, moderate, low, very low, or unassessed when insufficient evidence was found.

Because “Summary of findings” tables vary slightly in layout and structure, a deterministic scraping extraction was not appropriate. Consequently, we used an LLM to extract this information.

We provided the raw HTML code displaying the “Summary of findings” tables to an LLM (GPT-o3-mini-high, OpenAI's most recent mini-reasoning model available in April 2025). We developed prompts that were iteratively optimized on 15 reviews (the final prompt is available in Supplementary Section C.1). As the “Summary of findings” tables comprise highly structured tabular data, we used zero-shot prompting (i.e., no example given in the prompt).

The accuracy of the LLM data extraction was assessed in a random sample of 245 review versions (92 unique reviews, containing 2,660 PICO questions) manually checked by two specifically trained medical students.

They classified the extraction generated by the language model into four categories:

- **Correct**: the extracted text was identical to the text reported in the Cochrane "Summary of findings" table.
- **Minor error**: the extracted text differed slightly but without altering the interpretation. For example, for review CD006715, the outcome was "Pulmonary complications (0 to 30 days): Respiratory depression" on the Cochrane website, whereas the LLM extracted "Respiratory depression: (0 to 30 days)".
- **Missing line**: an entire line (PICO question) of the "Summary of findings" table was omitted, with all its elements. This typically occurred in reviews with particularly long tables, or with many tables.
- **Major error**: the extracted text differed from the original in a way that affected interpretation.

As quality control, the two reviewers independently checked in duplicate a subset of 309 PICO questions retrieved from 25 reviews. The raw agreement between the two reviewers was high: 99% to 100% depending on the elements (we did not use Cohen's kappa to judge agreement because the error rates were very low [14]; details in Supplementary Section B.2). Considering this high agreement, the remaining 2,351 PICO questions were divided between the two reviewers for single assessment.

Overall, among 2,660 PICO questions that were manually verified, 98% were judged to be correctly extracted, 1% were missed, and 1% were extracted but with major errors (see Supplementary Section B.2). Most major errors came from a single review (CD013756); when this single review was excluded, the major error rate dropped to 0.2%.

Automatic matching of similar PICO questions with previous versions of the review

To assess the evolution of evidence, we matched similar PICO questions present in the most recent update of a review and its initial versions that reported a "Summary of findings" table, when available. Because the wording of the PICO elements for the same clinical question may vary between versions (e.g., "participants with unexplained subfertility" vs "people with unexplained subfertility" for the population in review CD001838), we used LLM-assisted matching (gpt-o4-mini-high).

We provided the LLM with the review version identifiers and the PICO elements previously extracted from the "Summary of findings" tables into an Excel sheet. The associated certainty of evidence was not provided to the LLM, to avoid influencing the model's decision.

We optimized a prompt on 100 reviews. If a PICO question in one review version had no corresponding question in other versions, it was left unmatched. The final prompt is available in Supplementary Section C.3 (few-shot prompting, with one representative example).

For quality control, two reviewers independently assessed whether the matches identified by the LLM were consistent on a random sample of 181 reviews (425 distinct versions) addressing 4,853 PICO questions. The reviewers rated the level of agreement in four categories: complete agreement, mostly agree, mostly disagree, and complete disagreement. Any discrepancies were

resolved by a third reviewer. The PICO question was classified as similar between review versions only if all elements of the PICO (population-intervention-comparison-outcome) were similar.

Overall, the human reviewers agreed with the LLM's decisions in 98% of cases. The inter-reviewer raw agreement was 99%. The more granular assessment of the similarity between matches was as follows: complete agreement (82%), mostly agree (16%), mostly disagree (1%), and complete disagreement (0.6%). Inter-reviewer reliability was high (Cohen's κ=86%). More details for each PICO element are available in Supplementary Section B.4.

Analyses

Given the low error rates of the LLM-assisted tasks, the analysis was carried out on the entire raw database.

First, we described the amount (number of included studies and participants) and certainty (high, moderate, low, very low, unassessed) of evidence across all Cochrane reviews. If a review had multiple versions, we included only the most recent update.

Second, for reviews with at least two versions that reported a "Summary of findings" table, we described the evolution of the evidence between the initial and latest version.

RESULTS

Identification and characteristics of Cochrane Systematic Reviews

The flow chart is reported in Figure 1.

We identified 9,015 systematic reviews on interventions. Overall, 56% (n = 5,022) were not updated (i.e., one version), 27% (n = 2,416) had one update (i.e., two versions); 11% (n = 978) had two updates (i.e., three versions), and 7% (n = 599) had three or more updates. The median [Q1–Q3] delay to update was 4.5 [2.9–6.8] years.

We identified 5,116 reviews with at least one version reporting a "Summary of findings" table (Supplementary Figure S1), yielding a total of 64,849 PICO questions; for reviews with multiple versions, only the most recent version was analyzed. On average, these reviews had 2.3 "Summary of findings" tables. The median number of PICO questions assessed per review was 8 [Q1–Q3: 6–16].

Evidence available for each PICO question

Amount of evidence

Of the 64,849 PICO questions, 24% (n = 15,768) had no studies reporting the outcome; 31% (n = 20,390) had only one study included, 14% (n = 8,796) had two studies, and 31% (n = 19,895) had three or more studies (Table 1). Half (50%, n = 2,575) of the reviews had at least one PICO for which no study was identified. Nearly all PICOs (97%) with at least one included study relied

exclusively on RCTs. The median [Q1–Q3] number of included participants was 123 [0–557].

The amount of evidence included per PICO in reviews was stable over time (Figures 2 and Supplementary Figure S2).

Table 1. Evidence included (number of studies and participants) in assessments of the certainty of evidence for PICO questions in Cochrane reviews, and evolution of the included evidence when reviews were updated.

Evidence included in assessments of the certainty of evidence for PICO questions. 64,849 PICO questions from 5,116 Cochrane reviews with a "Summary of findings" table (only the most recent review update was analyzed if multiple versions existed).			
Number of included studies	% of PICO questions	Number of included participants	% of PICO questions
0	24%	0	25%
1	31%	1–100	21%
2	14%	101–500	27%
>2	31%	>500	27%
Evolution of included evidence after a review update (change from the first to the most recent review version) 7,461 PICO questions from 986 Cochrane reviews			
Number of included studies	% of PICO questions	Number of included participants	% of PICO questions
identical	63%	identical	59%
increase	33%	increase	35%
decrease	4%	decrease	6%

Certainty of evidence

At the PICO question level

Among the 64,849 PICO questions reported in the "Summary of findings" tables, the certainty of evidence was rated as high for 4% (n = 2,852), moderate for 16% (n = 10,574), low for 27% (n = 17,409), very low for 26% (n = 17,012), and not assessed for 26% (n = 17,002) because of insufficient data.

Over the 15-year period studied, there was no overall improvement in the certainty of evidence (Figure 3). Supplementary Figures S3 and S4 show the certainty of evidence according to the detailed distribution of the number of included studies (Supplementary Figure S3) and participants (Supplementary Figure S4) per PICO, with the PICO-associated certainty.

At the review level

Of the 5,116 systematic reviews with a "Summary of findings" table, 17% (n = 884) included at least one PICO question for which the evidence was rated as high certainty and 55% (n = 2,829) included at least one PICO question for which the evidence was rated as moderate or high. In contrast, 45% (n = 2,287) of reviews included only PICO questions for which the evidence was rated as low, very low, or unassessed. Only 5% (n = 269) of reviews included only PICO questions with evidence exclusively rated as "moderate" or "high".

Evolution of evidence after an update

Amount of evidence

Of the 64,849 PICOs (5,116 reviews), 52,127 PICOs (4,130 reviews) could not be matched to a previous assessment; 33,914 (2,761 reviews) had a unique version, and 18,213 (1,369 reviews) had several versions but only one version reported a "Summary of findings" table. In addition, among the reviews that were ultimately included, 5,261 PICOs could not be matched to a previous assessment.

This left 7,461 PICOs (986 reviews) included in the analysis for the effect of updates. The number of studies included in the most recent update remained identical to the number included in the original version for 63% (n = 4,721) of PICO questions, increased for 33% (n = 2,426), and decreased for 4% (n = 314) (Table 1). For PICOs with an increase in the number of included studies, the median [Q1–Q3] number of additional studies was 2 [1–4]. For PICOs with an increase in the number of included participants, the median [Q1–Q3] number of additional participants was 315 [117–1,073].

Supplementary Figures S5 and S6 show the number of new studies and participants included for each PICO question according to the delay between updates; after five years, at least one additional study was identified for about half of the updated PICOs.

Certainty of evidence

Figure 4 displays the evolution in the PICO certainty of evidence between the initial review version and the most recent update. Overall, the certainty of evidence did not improve: the certainty of evidence for 8% (n = 564) of PICOs in the initial versions of the reviews was rated as high, compared with 7% (n = 529) in the most recent updates. The transition table in Figure

4b highlights that out of the initial 7,461 assessed certainties, 71% (n = 5,329) remained similar; 13% (n = 980) were upgraded to a higher certainty, while 15% (n = 1,152) were downgraded to a lower certainty. Extreme transitions (e.g., from high to very low) were very uncommon.

Figure 5 shows the evolution of the certainty of evidence between the first version and the most recent update according to the delay between each version. The figure shows that the longer the duration between updates, the less likely the certainty was to remain unchanged. After 5 years, about half of the certainties of evidence remained unchanged (45% for high, 52% for moderate and 56% for low, 64% for very low certainty).

Sensitivity analysis

To explore why the certainty of evidence was sometimes downgraded, we performed a post-hoc subgroup analysis restricted to PICOs for which no new evidence was identified (Supplementary Figure S10). When no new evidence was included, certainty was downgraded in 10% of updated PICOs and upgraded in 4%. Moreover, the longer the interval before an update, the higher the probability of downgrading, increasing from <5% within the first two years to 20–30% after seven years. Since no additional evidence was incorporated, these downgrades probably reflect changes over time in how certainty of evidence is assessed

DISCUSSION

Statement of principal findings

We analyzed the evidence synthesized in 5,116 Cochrane reviews reporting a "Summary of findings" table for a total of 64,849 PICO questions. Overall, most PICO questions relied on little evidence: 68% included 2 studies or fewer, and 46% included 100 participants or fewer. Further, the certainty of evidence was most often low or very low, with high certainty very uncommon (i.e., 4% of PICOs). Updating a review seldom changed the amount of evidence: after 5 years, for about half of the PICOs no additional studies were included.

Strengths and weaknesses of the study

Strengths

To our knowledge, this is the first study to describe the certainty of evidence and its evolution on such a large sample. Furthermore, we included all Cochrane reviews of interventions as they are recognized for their methodological rigor and the quality of their evidence synthesis. They include a wide diversity of PICO questions, and apply a process to select research questions and outcomes considered most relevant to clinical practice and public health.

Limitations

This study has several limitations. First, data were extracted exclusively from the "Summary of findings" tables of systematic reviews of interventions published in the Cochrane Library, which limits the generalizability of the results; Further data extraction was automatically performed using an LLM, which may have introduced classification or extraction errors; however, a manual

quality control process was applied and confirmed high accuracy rates, which supports the robustness of the extracted data.

Finally, some changes over time in the certainty of evidence could be related to changes in how certainty of evidence is being assessed. For example, inter-rater reliability of GRADE assessment [15] as well as the move from the original Risk of Bias [16] tool to RoB 2 [17] could impact GRADE assessment and explain that even when no evidence is identified, the certainty of evidence was downgraded in 10% of updated PICOs and upgraded in 4%. This limitation only concerns the certainty of evidence, but not the amount of evidence.

Strengths and weaknesses in relation to other studies, discussing important differences in results

To our knowledge, this is the first study to systematically assess the amount of evidence associated with each PICO in Cochrane Summary of Findings tables. Previous studies that quantified the number of studies and participants were restricted to the review level [18] and did not assess evidence at the outcome level.

Our results are consistent with previous cross-sectional studies assessing the certainty of evidence in smaller samples of Cochrane reviews [19–24], or in specific medical fields such as anesthesia, critical care and emergency medicine [19], surgical procedures for fractures [25], neonatology [26], oral health [27,28] and physical therapy [29].

We identified only two studies assessing the evolution of the certainty after an update on a limited sample of Cochrane reviews. They also found that most updates (50 to 70%) did not change the certainty of evidence [20,21].

Finally, our results are in line with the results of studies evaluating the level of evidence of clinical practice guidelines recommendations for various organizations in different fields such as cardiology [30-33], infectious disease [34,35], pediatrics [36], oncology [37,38], urology [39,40], gastroenterology [41] or more generally by the WHO [42]. Overall, only 4% to 30% of clinical practice guideline are supported by a high level of evidence.

Meaning of the study: possible explanations and implications for clinicians and policymakers

Our results question the efficiency of the current evidence ecosystem, which is unable to adequately support decision making [7]. Most Cochrane systematic reviews of intervention included only randomized controlled trials, which are considered the gold standard for therapeutic evaluation. Randomization with allocation concealment is the best approach to limit bias in the selection of participants and confounding biases. In addition, RCTs are subject to strict methodological and regulatory requirements, including prior approval by an ethics committee, prospective trial registration, and standardized procedures for data collection. These requirements are accompanied by quality control and data verification processes, which ensure the reliability of the findings.

Nevertheless, considering only randomized evidence to support decision making also has important limitations. Indeed, it is very unlikely that randomized evidence alone will be able to support all clinical decisions. RCTs are costly (tens to hundreds of millions of US dollars) [43], raise recruitment challenges (from 1997 to 2020, 4 out of 10 NIHR-funded RCTs did not meet their recruitment target sample size) [44], and have low external validity. In addition, RCTs

are not always feasible. Further, the current production of RCTs is associated with considerable avoidable research waste [45]. The production of trials is fragmented. For instance, Yordanov et al. [46] showed that 78% of trials included in a Cochrane review were excluded from at least one of the review meta-analyses because trials' important outcomes were not appropriately planned, or were incompletely reported.

To adequately support decision, there is a need to reconsider the whole system and bridge the gap between evidence generation, evidence synthesis and guidelines development to develop a learning system [6,46]. Actions should be put in place to reduce avoidable waste in trials planning, conduct and reporting and to explore other source of evidence [4]. The increasing access to high quality routinely collected data [47] as well as novel framework such as target trial emulation to limit bias should be considered [48].

Unanswered questions and future research

Future work could investigate how to expand the primary evidence base and strengthen the certainty of evidence in systematic reviews.

A first possible axis is to consider systematic reviews not only as syntheses of existing evidence but also as tools to guide future research, by identifying important questions that are under-studied. Future research could determine under which concrete mechanisms systematic reviews could provide feedback to trial funders and investigators, by signaling evidence gaps and potential areas for improvement in outcome reporting. A key outcome would be whether such

feedback loops increase the volume and certainty of evidence over subsequent years.

A second possible axis is to establish robust methods for conducting high quality non-randomized studies of interventions and incorporating non-randomized evidence in evidence synthesis to support decision making.

CONCLUSION

In this large-scale assessment of Cochrane systematic reviews of interventions, we found that the amount of available evidence was frequently limited: one quarter of all PICO questions had no included studies, and more than half were informed by either one or two studies. High-certainty evidence was uncommon, and most certainties were low, very low, or unassessed. Updates brought no new evidence for most PICOs, and seldom improved certainty; when changes in certainty occurred, upgrades and downgrades were balanced.

Overall, these findings point to the importance of improving the primary research ecosystem.

Statements

Public and Patient Involvement statement: No patients or members of the public were involved in this research.

Ethics approval: Ethics approval was not required according to French regulations as this study did not involve human participants, their data, or animals.

Scraping and copyrighted content: This work complies with the terms and conditions of Wiley, the publisher of the Cochrane Library (see https://www.wiley.com/en-gb/terms-of-use/ai-principles/wiley-statement-illegal-scraping-ai-copyright). We used information from Cochrane review abstracts (publicly available) and "Summary of findings" tables (public or licensed access). No copyrighted content was reproduced or shared; only aggregate statistics are reported. No AI training or commercial reuse was involved.

Author contributorship: TS and IB conceived the study. TS carried out the data extraction, coding, and analyses. TS, IB, and PR contributed to the interpretation of the results. TS and IB wrote the first draft of the manuscript. IB and PR provided senior guidance throughout the project, and revised the manuscript. TS is the guarantor, had full access to the data, and made the final decision to submit for publication. All authors approved the final version of the manuscript.

Transparency statement: The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported and that no important aspects of the study have been omitted.

Role of the funding source: No funding was required for this study.

Competing interest: Isabelle Boutron is director of Cochrane France and a member of the Cochrane editorial board.

Data sharing: Data can be made available upon request after submission of a protocol which will be evaluated by authors in relation to other ongoing projects using these data.

Acknowledgements We thank Victoire Perrichon and Mazal Sultan, the two medical students who manually checked the accuracy of the AI data extraction and matching of PICO questions.

References

1 Marshall IJ, Nye B, Kuiper J, *et al*. Trialstreamer: A living, automatically updated database of clinical trial reports. *J Am Med Inform Assoc*. 2020;27:1903–12. doi: 10.1093/jamia/ocaa163

2 Donnelly CA, Boyd I, Campbell P, *et al*. Four principles to make evidence synthesis more useful for policy. *Nature*. 2018;558:361–4. doi: 10.1038/d41586-018-05414-4

3 Cochrane. Our impact [online]. 2025. https://www.cochrane.org/about-us/our-impact (accessed 25 September 2025)

4 Yordanov Y, Dechartres A, Porcher R, *et al*. Avoidable waste of research related to inadequate methods in clinical trials. *BMJ*. 2015;350:h809. doi: 10.1136/bmj.h809

5 Créquit P, Boutron I, Meerpohl J, *et al*. Future of evidence ecosystem series: 2. current opportunities and need for better tools and methods. *J Clin Epidemiol*. 2020;123:143–52. doi: 10.1016/j.jclinepi.2020.01.023

6 Ravaud P, Créquit P, Williams HC, *et al*. Future of evidence ecosystem series: 3. From an evidence synthesis ecosystem to an evidence ecosystem. *J Clin Epidemiol*. 2020;123:153–61. doi: 10.1016/j.jclinepi.2020.01.027

7 Boutron I, Créquit P, Williams H, *et al*. Future of evidence ecosystem series: 1. Introduction Evidence synthesis ecosystem needs dramatic change. *J Clin Epidemiol*. 2020;123:135–42. doi: 10.1016/j.jclinepi.2020.01.024

8 Guyatt G, Hultcrantz M, Agoritsas T, *et al*. Why Core GRADE is needed: introduction to a new series in The BMJ. *BMJ*. 2025;389:e081902. doi: 10.1136/bmj-2024-081902

9 GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328:1490. doi: 10.1136/bmj.328.7454.1490

10 Guyatt GH, Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924–6. doi: 10.1136/bmj.39489.470347.AD

11 Guyatt G, Oxman AD, Akl EA, *et al*. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64:383–94. doi: 10.1016/j.jclinepi.2010.04.026

12 Schünemann HJ, Higgins JP, Vist GE, *et al*. Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane 2024.

13 Cochrane Library. Search [online]. 2025. https://www.cochranelibrary.com/cdsr/reviews (accessed 25 September 2025)

14 McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22:276–82. doi: 10.11613/BM.2012.031

15 Mustafa RA, Santesso N, Brozek J, *et al*. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *J Clin Epidemiol*. 2013;66:736-742.e5. doi: 10.1016/j.jclinepi.2013.02.004

16 Higgins JPT, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928. doi: 10.1136/bmj.d5928

17 Sterne JAC, Savović J, Page MJ, *et al*. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi: 10.1136/bmj.l4898

18 Mallett S, Clarke M. The Typical Cochrane Review: How Many Trials? How Many Participants? *Int J Technol Assess Health Care*. 2002;18:820–3. doi: 10.1017/S0266462302000624

19 Conway A, Conway Z, Soalheira K, *et al*. High quality of evidence is uncommon in Cochrane systematic reviews in Anaesthesia, Critical Care and Emergency Medicine. *Eur J Anaesthesiol*. 2017;34:808–13. doi: 10.1097/EJA.0000000000000691

20 Djulbegovic B, Ahmed MM, Hozo I, *et al*. High quality (certainty) evidence changes less often than low-quality evidence, but the magnitude of effect size does not systematically differ between studies with low versus high-quality evidence. *J Eval Clin Pract*. 2022;28:353–62. doi: 10.1111/jep.13657

21 Howick J, Koletsi D, Pandis N, *et al*. The quality of evidence for medical interventions does not improve or worsen: a metaepidemiological study of Cochrane reviews. *J Clin Epidemiol*. 2020;126:154–9. doi: 10.1016/j.jclinepi.2020.08.005

22 Yordanov Y, Dechartres A, Ravaud P. Patient-important outcomes in systematic reviews: Poor quality of evidence. *PLoS One*. 2018;13:e0195460. doi: 10.1371/journal.pone.0195460

23 Howick J, Koletsi D, Ioannidis JPA, *et al*. Most healthcare interventions tested in Cochrane Reviews are not effective according to high quality evidence: a systematic review and meta-analysis. *J Clin Epidemiol*. 2022;148:160–9. doi: 10.1016/j.jclinepi.2022.04.017

24 Fleming PS, Koletsi D, Ioannidis JPA, *et al*. High quality of the evidence for medical and other health-related interventions was uncommon in Cochrane systematic reviews. *J Clin Epidemiol*. 2016;78:34–42. doi: 10.1016/j.jclinepi.2016.03.012

25 Daou JP, Riera R, Pacheco RL. Reasons for Downgrading the Certainty of Evidence for Indirectness in Synthesis of Surgical Procedures for Patients With Fractures: A Meta-Research Analysis. *J Eval Clin Pract*. 2025;31:e70091. doi: 10.1111/jep.70091

26 Varrio T, De Luca D, Kuitunen I. Evidence certainty in neonatology—a meta-epidemiological analysis of Cochrane reviews. *Eur J Pediatr*. 2025;184:191. doi: 10.1007/s00431-025-06023-w

27 Seehra J, Bertl K, Faggion CM, *et al*. The certainty of the evidence in oral health has not improved according to GRADE: a meta-epidemiological study. *J Clin Epidemiol*. 2022;142:29–37. doi: 10.1016/j.jclinepi.2021.10.016

28 Pandis N, Fleming PS, Worthington H, *et al*. The Quality of the Evidence According to GRADE Is Predominantly Low or Very Low in Oral Health Systematic Reviews. *PLoS One*. 2015;10:e0131644. doi: 10.1371/journal.pone.0131644

29 Freire APCF, Eustis KJ, Lemes ÍR, *et al*. Conclusiveness of Cochrane systematic reviews is low but accumulating across time in physical therapy: A meta-research study. *Braz J Phys Ther*. 2025;29:101190. doi: 10.1016/j.bjpt.2025.101190

30 Tricoci P, Allen JM, Kramer JM, *et al*. Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines. *JAMA*. 2009;301:831–41. doi: 10.1001/jama.2009.205

31 Fanaroff AC, Califf RM, Windecker S, *et al*. Levels of Evidence Supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018. *JAMA*. 2019;321:1069–80. doi: 10.1001/jama.2019.1122

32 van Dijk WB, Grobbee DE, de Vries MC, *et al*. A systematic breakdown of the levels of evidence supporting the European Society of Cardiology guidelines. *Eur J Prev Cardiol*. 2019;26:1944–52. doi: 10.1177/2047487319868540

33 Boriani G, Venturelli A, Imberti JF, *et al*. Comparative analysis of level of evidence and class of recommendation for 50 clinical practice guidelines released by the European Society of Cardiology from 2011 to 2022. *Eur J Intern Med*. 2023;114:1–14. doi: 10.1016/j.ejim.2023.04.020

34 Miles KE, Rodriguez R, Gross AE, *et al*. Strength of Recommendation and Quality of Evidence for Recommendations in Current Infectious Diseases Society of America Guidelines. *Open Forum Infect Dis*. 2021;8:ofab033. doi: 10.1093/ofid/ofab033

35 Lee DH, Vielemeyer O. Analysis of Overall Level of Evidence Behind Infectious Diseases Society of America Practice Guidelines. *Arch Intern Med*. 2011;171:18–22. doi: 10.1001/archinternmed.2010.482

36 Matheny Antommaria AH, Kelleher M, Peterson RJ. Quality of Evidence and Strength of Recommendations in American Academy of Pediatrics' Guidelines. *Pediatrics*. 2025;155:e2024067836. doi: 10.1542/peds.2024-067836

37 Cliff ERS, Rome RS, Kesselheim AS, *et al*. National Comprehensive Cancer Network Guideline Recommendations of Cancer Drugs With Accelerated Approval. *JAMA Netw Open*. 2023;6:e2343285. doi: 10.1001/jamanetworkopen.2023.43285

38 Poonacha TK, Go RS. Level of Scientific Evidence Underlying Recommendations Arising From the National Comprehensive Cancer Network Clinical Practice Guidelines. *J Clin Oncol*. 2011;29:186–91. doi: 10.1200/JCO.2010.31.6414

39 Du C, Harandi AA, Hwang K, *et al*. Analysis of the AUA Guidelines: Strength of Evidence Correlates With Recommendation Levels. *Urol Pract*. 2023;10:116–21. doi: 10.1097/UPJ.0000000000000373

40 Antoine SG, Small AC, McKiernan JM, *et al*. Analysis of evidence within the AUA's clinical practice

guidelines. *Can J Urol*. 2018;25:9168–78.

41 Correia C, Almeida N, Figueiredo PN. Guidelines in Gastroenterology: Careful Interpretation Is Essential. *GE Port J Gastroenterol*. 2021;29:240–6. doi: 10.1159/000518322

42 Alexander PE, Brito JP, Neumann I, *et al*. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. *J Clin Epidemiol*. 2016;72:98–106. doi: 10.1016/j.jclinepi.2014.10.011

43 Speich B, von Niederhäusern B, Schur N, *et al*. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *J Clin Epidemiol*. 2018;96:1–11. doi: 10.1016/j.jclinepi.2017.12.018

44 Jacques RM, Ahmed R, Harper J, *et al*. Recruitment, consent and retention of participants in randomised controlled trials: a review of trials published in the National Institute for Health Research (NIHR) Journals Library (1997–2020). *BMJ Open*. 2022;12:e059230. doi: 10.1136/bmjopen-2021-059230

45 Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet*. 2009;374:86–9. doi: 10.1016/S0140-6736(09)60329-9

46 Yordanov Y, Dechartres A, Atal I, *et al*. Avoidable waste of research related to outcome planning and reporting in clinical trials. *BMC Medicine*. 2018;16:87. doi: 10.1186/s12916-018-1083-x

47 Franklin JB, Marra C, Abebe KZ, *et al*. Modernizing the Data Infrastructure for Clinical Research to Meet Evolving Demands for Evidence. *JAMA*. 2024;332:1378–85. doi: 10.1001/jama.2024.0268

48 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183:758–64. doi: 10.1093/aje/kwv254

Figure Legends

Figure 1. Flow chart of the included PICO questions (combinations of population-intervention-comparison-outcome). Numbers in blue disks indicate an LLM-assisted task. 1: extraction of the review literature search dates from abstracts; 2: extraction of the "Summary of findings" table contents (PICO elements, certainty of evidence, number of included participants, number of included studies). 3: matching of PICOs in the most recent review update with a "Summary of findings" table to PICOs in previous review versions.

Figure 2. Top: number of PICO questions assessed each year in Cochrane reviews with a "Summary of findings" table, with the number of studies included for the PICOs. Bottom: relative share of each category of number of included studies, for each year.

Figure 3: Top: number of PICO questions assessed each year in Cochrane reviews with a "Summary of findings" table, for the different certainties of evidence. Bottom: relative share of each certainty of evidence, for each year.

Figure 4: Evolution of the certainty of evidence for the 7,461 pairs of PICO questions (assessed in 986 reviews) for which more than one review version with a "Summary

of findings" table was available. **Top**: Sankey diagram displaying the evolution from the first to the most recent version. **Bottom**: full table of transitions between the first and the most recent version. As an example of how to read this table, the information indicated by the arrow in the top right corner would be interpreted as follows: *out of the 564 certainties graded high in the first review version, 334 (59%) remained high in the most recent version of the same review.* On the diagonal are unchanged certainties; in the bottom right, downgraded certainties; and in the top left, upgraded certainties.

Figure 5: Evolution of the certainty of evidence, with increasing interval between the first version (t=0) and the most recent version (x axis). Results are displayed separately for each level of certainty in the first version (panels from top to bottom). The y axis displays the proportion of PICO questions for which evidence was categorized as each of the four certainties. The absolute number of PICO questions and systematic reviews included for each year are detailed in Supplementary Figure S9.

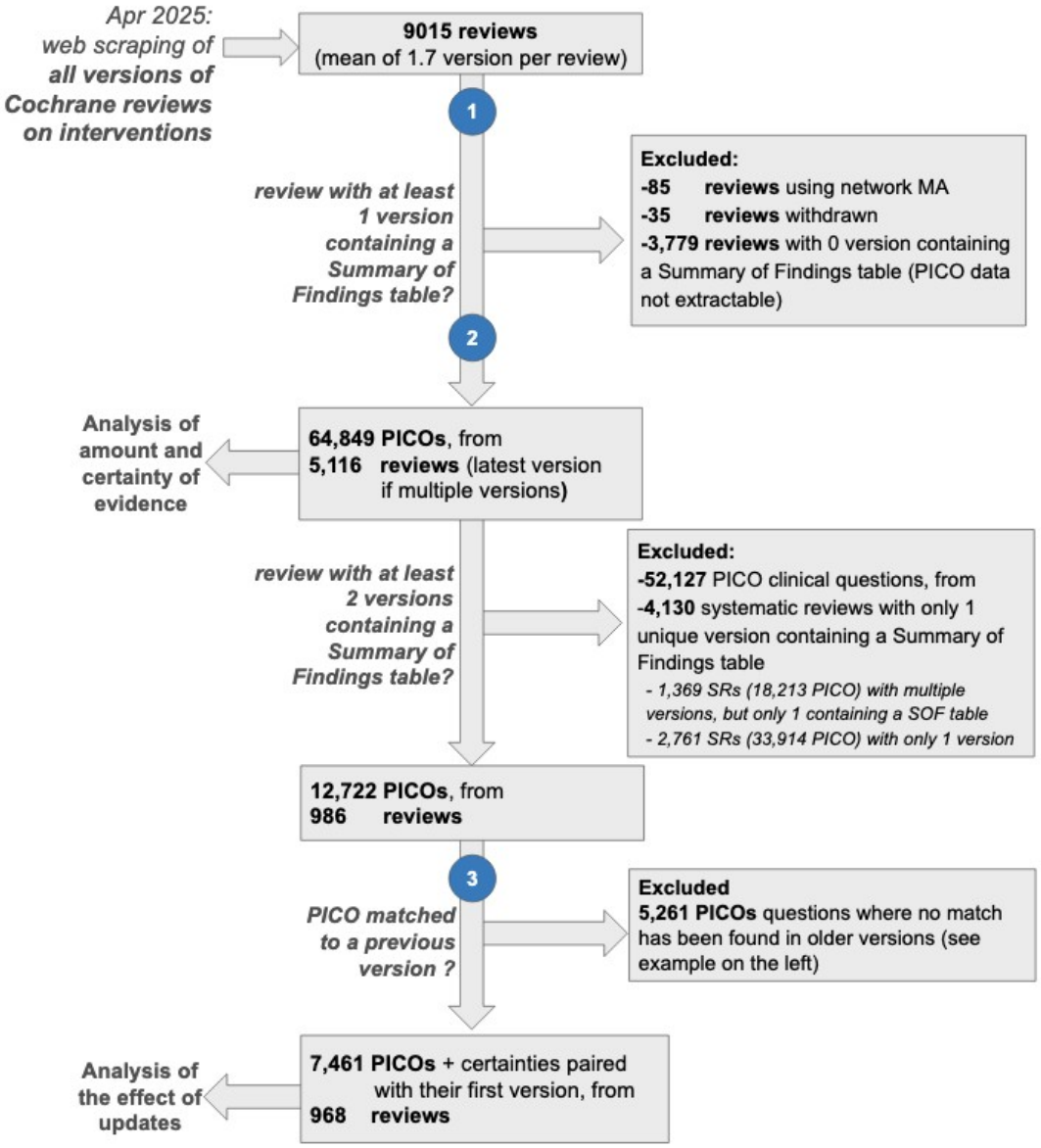

Figure 1. Flow chart of the included PICO questions (combinations of population-intervention-comparison-outcome). Numbers in blue disks indicate an LLM-assisted task. 1: extraction of the review literature search dates from abstracts; 2: extraction of the "Summary of findings" table contents (PICO elements, certainty of evidence, number of included participants, number of included studies). 3: matching of PICOs in the most recent review update to PICOs in previous review

versions.

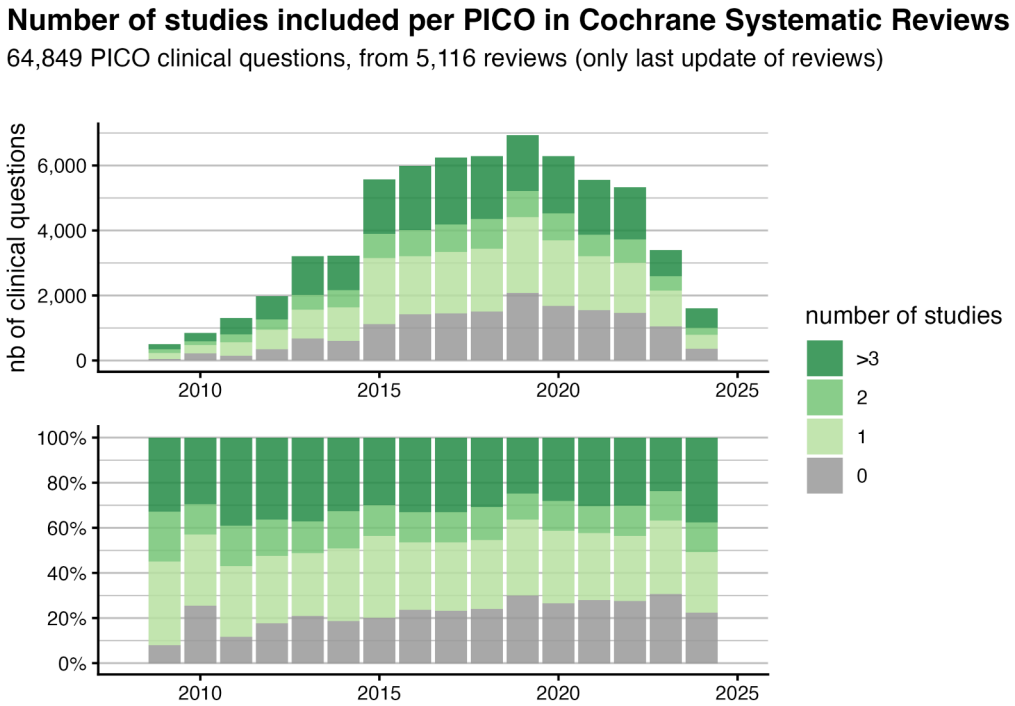

Figure 2. Top: number of PICO questions assessed each year in Cochrane reviews with a "Summary of findings" table, with the number of studies included for the PICOs. Bottom: relative share of each category of number of included studies, for each year.

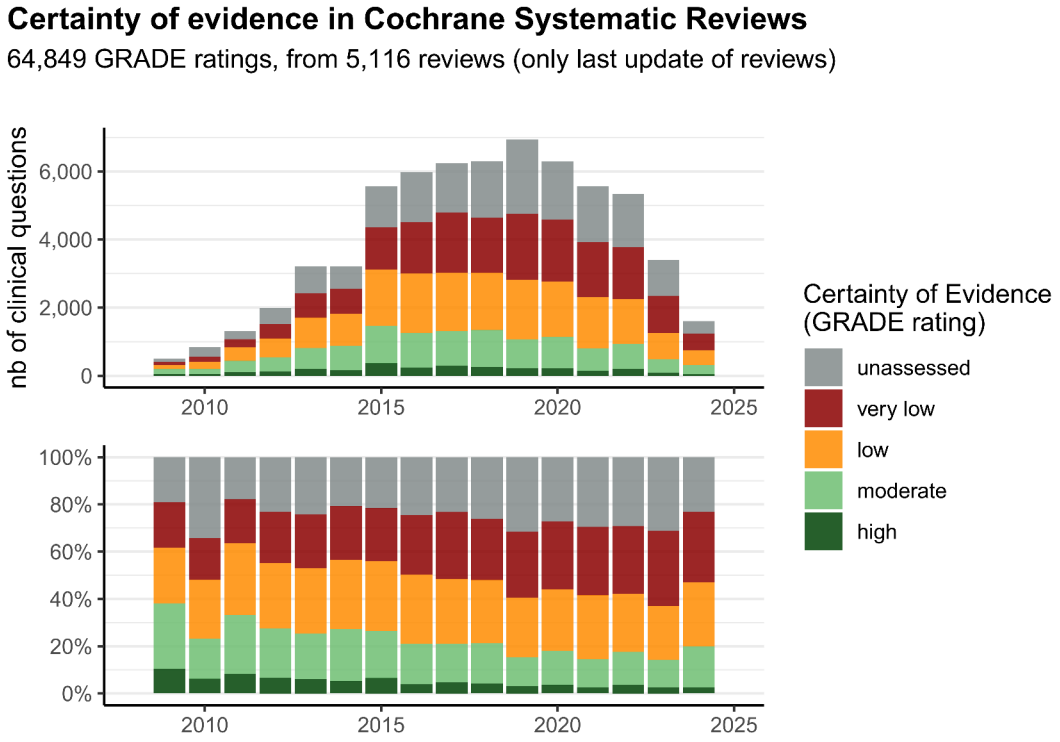

Figure 3: Top: number of PICO questions assessed each year in Cochrane reviews with a "Summary of findings" table, for the different certainties of evidence. Bottom: relative share of each certainty of evidence, for each year.

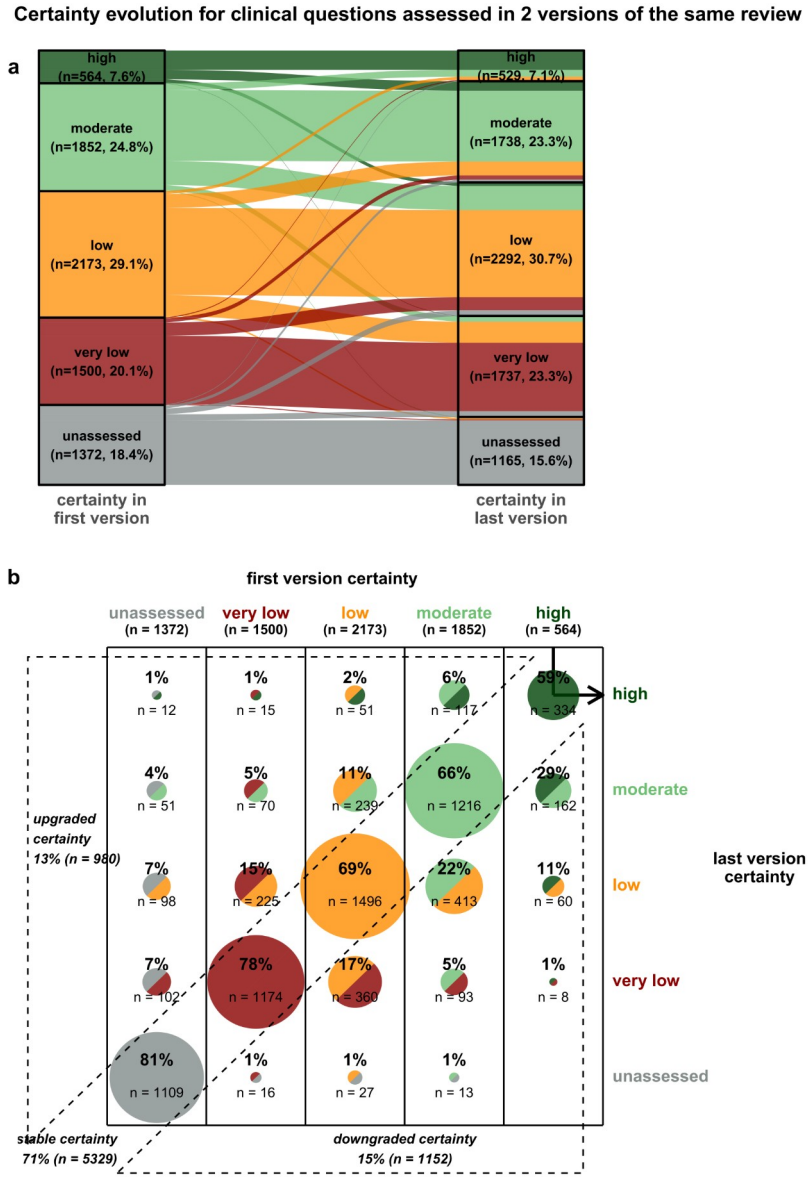

Figure 4: Evolution of the certainty of evidence for the 7,461 pairs of PICO questions (assessed in 986 reviews) for which more than one review version was available. **Top**: Sankey diagram displaying the evolution from the first to the most recent version. **Bottom**: full table of transitions between the first and the most recent version. As an example of how to read this table, the information indicated by the arrow in the top right corner would be interpreted as follows: *out of the 564 certainties graded high in the first review version, 334 (59%) remained high in the*

most recent version of the same review. On the diagonal are unchanged certainties; in the bottom right, downgraded certainties; and in the top left, upgraded certainties.

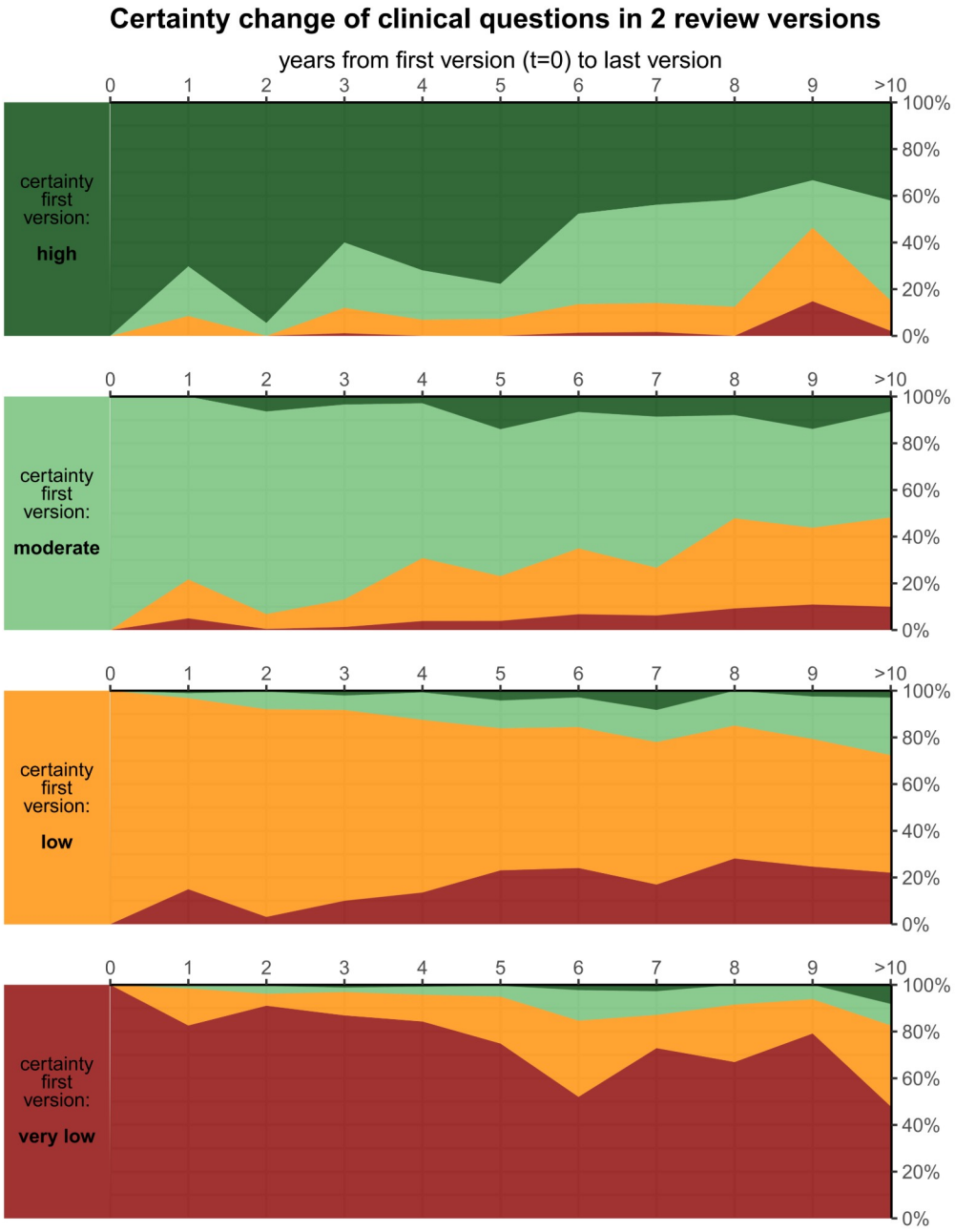

Figure 5: Evolution of the certainty of evidence, with increasing interval between the first version (t=0) and the most recent version (x axis). Results are displayed separately for each level of certainty in the first version (panels from top to bottom). The y axis displays the proportion of PICO questions for which evidence was categorized as each of the four certainties. The absolute number of PICO questions and systematic reviews included for

each year are detailed in Supplementary Figure S9.