Research Square

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

R-SYS for protocol versioning and independent reviewer workflows in systematic reviews

Yehya Tlaiss

yehyatlaiss@hotmail.com

R-Chief Labs Ltd.

Mohamad Ismail

R-Chief Labs Ltd.

Method Article

Keywords: Systematic reviews, Review methodology, Workflow, Artificial intelligence, Protocol amendments, Reproducibility, Evidence synthesis

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11584864/v1

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

Additional Declarations: Competing interest reported. YT and MI are affiliated with R-Chief Labs Ltd, which develops and commercialises R-SYS, and hold a financial interest in it. The core modules described in this paper are released under an open-source licence; the surrounding platform is commercial.

Abstract

Background

Systematic review software must coordinate independent human decisions, protocol amendments and heterogeneous information sources. Synchronizing reviewers at every stage can create waiting time unrelated to the scientific assessment of a record. We describe R-SYS, a workflow that links decisions to versioned protocols and allows reviewers to progress independently before a final reconciliation step.

Methods

The reference implementation combines an eleven-phase guarded lifecycle, immutable locked protocol versions, deterministic rendering of a canonical search strategy, per-reviewer working sets and decision provenance. Each primary reviewer advances through title and abstract screening, full-text assessment, extraction and risk-of-bias appraisal using their own preceding decisions, subject to human adjudication. A final convergence checkpoint resolves outstanding work before synthesis. Artificial intelligence (AI) supplies advisory recommendations that are excluded from human reconciliation. We analyse completion times under explicit scheduling assumptions and describe implementation limitations.

Results

With identical task sets, continuous reviewer availability and no external waiting, independent lanes achieve a completion time equal to the largest reviewer workload. This is a lower bound for schedules that impose additional synchronization constraints. Disagreement can introduce superseded work and delays; a sufficient condition for retaining an advantage over stage barriers is stated. The implementation provides mechanisms for protocol change control, reviewer coordination and traceable reporting. No empirical measurements of accuracy, efficiency, usability or cost are reported.

Conclusions

R-SYS provides a reference design for separating reviewer independence from workflow synchronization. Its theoretical scheduling properties do not establish practical time savings, search completeness or AI accuracy. A separate, prospectively specified validation study is planned after completed reviews using the platform become available.

Background

Systematic reviews require coordinated work across searching, selection, data collection and synthesis. An analysis of PROSPERO-registered reviews estimated a mean interval of approximately 67 weeks from registration to publication [1]. Software can assist with this work, but a coherent audit trail must

distinguish what researchers decided, which methodology they applied and what automation contributed.

Automated citation classification, collaborative screening and active learning have established precedents [2–5]. Evaluations of large language models (LLMs) show that screening performance depends on the task, model and prompting strategy [6–9]. Search translation is also an established area of automation: the Polyglot Search Translator translates strategies across database interfaces and has been evaluated in a randomized trial [10]. These developments mean that a new platform should be assessed against specific methodological contributions rather than a general claim to automate systematic reviews.

The contribution examined here is the integration of protocol versioning with independent reviewer progression and a final reconciliation checkpoint. In a synchronized workflow, a reviewer may be unable to begin the next stage until a peer has completed the same record, or until the whole team has completed the preceding stage. Such dependencies may be useful organizational choices, but they are distinct from the requirement for independent eligibility assessment. Guidance emphasizes independent final inclusion decisions and a predefined disagreement-resolution process; it does not prescribe a universal software scheduling discipline [11].

R-SYS assigns each reviewer a working set derived from their own preceding decisions, with human adjudication taking precedence where available. Reviewers may therefore progress at different rates. The cost is that later adjudication can require catch-up work or invalidate downstream work already performed. A final convergence checkpoint addresses these discrepancies before shared synthesis. Locked protocol versions and decision-level provenance make methodological changes traceable throughout this process.

This article describes the method and the reference implementation, identifies the assumptions behind its scheduling properties and defines the boundaries of the present evidence. It does not report a systematic review of platform effectiveness, a validated AI model or an empirical comparison with other products. Technical details and priorities for the separate follow-up evaluation are provided in Additional file 1.

Methods

Scope and architecture

R-SYS is a web application with a backend service, relational persistence, a single-page interface and an OpenID Connect identity provider. Long-running search, screening and extraction operations execute asynchronously. The implementation separates workflow coordination, review services, external integrations and persistent review state (Fig. 1). AI calls use the Anthropic API through a provider abstraction.

The platform supports single- and dual-reviewer configurations. The independent-lane method described here concerns two assigned primary reviewers. Single-reviewer reviews use the ordinary phase workflow and cannot provide inter-reviewer agreement estimates. Although several review types can be selected in the interface, this article addresses systematic reviews and pairwise meta-analysis; broad support for other evidence-synthesis methodologies is not established by the presence of a review-type option.

Lifecycle and guarded transitions

A review progresses through eleven phases: project creation, protocol, search, title and abstract screening, full-text assessment, extraction, risk of bias, convergence, synthesis, reporting and completion. A guard evaluates the requirements for each of the ten transitions. Preparation requires a locked protocol, accepted search-strategy peer review, and registration recorded as confirmed or explicitly waived. Screening requires an executed search or recorded import and an empty deduplication resolution worklist. Subsequent guards check the required work and unresolved discrepancies. Transition requirements are detailed in Additional file 1.

Each reviewer advances through the four per-record stages independently. The review-level phase summarizes the least-advanced primary reviewer and does not itself define a faster reviewer's working set. Final convergence is a mandatory checkpoint before synthesis. Reconciliation can also occur earlier, including during rolling adjudication; “final convergence” therefore denotes one lifecycle checkpoint, not a claim that statuses are written only once. Living updates of completed reviews are outside the present release.

Protocol versions and amendments

The protocol stores the research question, eligibility criteria, language policy, canonical search strategy, declared information sources, AI configuration and, where applicable, the pooling model and statistical options. Language restrictions are specified as eligibility decisions rather than silently imposed through search syntax [12]. A locked version is immutable. An amendment creates a new version with a reason; previous versions remain available, and decisions retain the version under which they were made (Fig. 2).

The described implementation offers three amendment scopes: all records, unscreened records only, or records previously included or marked uncertain. Locking an amendment marks affected decisions stale and returns affected records to screening. Selective policies can leave a review containing decisions made under different versions and are not interchangeable methodological choices. In particular, broadened eligibility criteria can affect previously excluded records; re-screening only included or uncertain records cannot recover those records. The review team must document the effect of the amendment and reassess all potentially affected records, including exclusions. The software's policy selector does not establish that the chosen scope is adequate.

Changes to search concepts or information sources may additionally require searches to be rerun and newly retrieved records to be screened. Re-screening an existing corpus alone is insufficient when an amendment changes the evidence that should have been retrieved. No automated inference of the complete consequences of an amendment is claimed.

Registration and search development

Registration status is recorded through an OSF pathway, a manually entered PROSPERO identifier, or a waiver with a justification. The OSF pathway prepares a registration from review configuration for researcher review and approval. The platform records confirmation rather than treating draft submission as completed registration. Registration reporting follows the transparency objectives of PRISMA 2020 [13]; the generalized systematic review registration form provides an option when a more specific form is unsuitable [14]. A registry record does not itself validate the conduct of the review.

The canonical query contains concept blocks. Synonyms and controlled-vocabulary terms are combined with OR within a concept and concepts are combined with AND. An AI assistant may propose a draft, the user edits it, and candidate MeSH terms are checked against NLM. A deterministic renderer then produces target-specific expressions. Only PubMed and Europe PMC have programmatic search clients in the described release; the other eleven catalogue entries provide expressions for manual execution or further adaptation (Table 1).

For fixed query and renderer versions, expressions are repeatable. This does not establish equivalent retrieval, valid vocabulary mapping or search sensitivity. NLM checks do not validate Emtree, CINAHL or PsycINFO terminology. Each translation requires specialist review of controlled terms, fields, operators and limits. Removing an OR-connected term can narrow retrieval. The executed expression and manual adaptations should be retained with the canonical query.

The workflow records search-strategy peer review using PRESS items [15]. An independent project member reviews the strategy, but completion of a checklist is not evidence of specialist competence or search completeness. We recommend information-specialist review of the strategy and its translations.

Table 1
Search targets and interpretation of rendered expressions

Targets	Execution in R-SYS	Required interpretation or review
PubMed; Europe PMC	Programmatic clients	Check MeSH and free-text terms, fields, limits and the actual executed query.
Cochrane CENTRAL; Cochrane Database of Systematic Reviews	Manual execution	MeSH-oriented expressions require checking in the selected Cochrane interface.
Embase	Manual execution	MeSH labels are not mapped to Emtree; subject terms and syntax require adaptation.
Scopus; Web of Science	Manual execution	Heading labels become free text; verify field scoping, phrases and interface-specific syntax.
CINAHL; PsycINFO	Manual execution	Free-text expressions do not supply validated native thesaurus mappings.
ClinicalTrials.gov; WHO ICTRP; medRxiv; Google Scholar	Manual execution or adaptation	Generic expressions require target-specific review; executable semantic equivalence is not established.

Acquisition and source provenance

Records enter through programmatic searches, RIS or PubMed-format imports, DOI resolution, or a companion browser extension. The extension can capture bibliographic metadata and accessible full text using the researcher's browser session. DOI resolution provides a metadata check; it does not establish eligibility or study identity.

Acquisition mechanism and discovery source must be distinguished for PRISMA reporting. A database search imported through a file remains a database search; website searching, citation searching and contacts are different discovery methods. The described capture route assigns an OTHER_METHODS label. That label requires review when capture or DOI resolution is used to acquire a record originally identified through a database search. Consequently, the current acquisition route alone cannot guarantee correct PRISMA arm assignment. The team must reconcile original source provenance before exporting a publication flow diagram.

Deduplication

Deduplication follows sequential matching from strict identifiers to less specific metadata, with human resolution of uncertain pairs, as a general principle described by Bramer et al. [16]. The thresholds below are implementation choices and have not been validated against a labelled duplicate set (Fig. 3).

The first pass groups exact normalized DOIs. The second applies exact PubMed-identifier matching to records without a usable DOI.The third considers records with neither identifier, blocks candidates on the first three normalized title tokens and calculates title similarity. A similarity of at least 0.92 with agreement on first-author surname and publication year triggers a merge. Similarity from 0.85 to below 0.92, or at least 0.92 with a failed author/year condition, requires human resolution. Lower-scoring candidate pairs remain distinct. Merged records are flagged rather than deleted and the operation is audited.

This restricted fallback design may miss duplicates with asymmetric identifier availability, conflicting identifiers, changed initial title tokens or incomplete author/year metadata. Exact identifier matches also require safeguards against erroneous source metadata. Comprehensive duplicate detection is therefore not claimed. Record deduplication is distinct from linking separate publications that describe the same underlying study; that distinction must be preserved for synthesis and reporting.

Independent reviewer working sets

For reviewer r and a record at a given stage, the effective preceding decision is the adjudicated decision if one exists and otherwise that reviewer's own human decision. AI recommendations are excluded. Title and abstract screening includes all non-duplicate records. Full-text assessment includes records with an effective title/abstract decision of include or uncertain. Extraction and risk-of-bias appraisal include records with an effective full-text decision of include (Table 2).

Table 2
Working sets for an assigned primary reviewer

Stage	Records requiring that reviewer's assessment
Title and abstract	All non-duplicate records in the screening corpus.
Full text	Effective preceding decision is include or uncertain.
Extraction	Effective full-text decision is include.
Risk of bias	Effective full-text decision is include.

Each reviewer completes their own working set before advancing their lane. When adjudication includes a record that a reviewer had excluded, the required downstream assessment becomes catch-up work. If that reviewer has already advanced, the outstanding task must still be completed before convergence. When adjudication excludes a record, downstream work already performed is retained as superseded and does not enter synthesis. Figure 4 contrasts this scheduling discipline with stage and record barriers.

Independence, visibility and disagreement resolution

The implementation hides a peer's and the AI's record-level decisions from a primary reviewer until that reviewer has submitted their own decision. The review-wide AI suggestion list becomes available after that reviewer's title/abstract lane is complete, and openings are audited. Principal investigators, adjudicators and observers have broader visibility. These rules protect specified initial submissions; they do not establish complete blinding throughout a review. Post-submission exposure can affect amendments, adjudication and later-stage decisions, and overlapping roles require explicit management.

Disagreements can be resolved by an adjudicator or principal investigator with a recorded decision and rationale. The release also contains an optional inclusive bypass that carries unresolved conflicts forward and logs the action separately. Such a bypass must not be interpreted as an independent third-person eligibility assessment. Inclusive progression can protect sensitivity at title/abstract screening, but final inclusion requires documented assessment against full-text eligibility criteria. An unresolved full-text disagreement should receive substantive resolution or remain awaiting assessment, rather than qualify for synthesis solely through a bypass [11]. The presence of the bypass is an implementation limitation, not a recommended substitute for adjudication.

Advisory AI screening

The configured model receives a record's title and abstract and the locked eligibility criteria. It returns per-criterion assessments, an include/exclude/uncertain recommendation, a brief reason and verbalized confidence. The model identifier, prompt-template version, sampling parameters and threshold belong to the protocol. A date-specific model identifier should be used where available; the exact identifier must be reported for each empirical application.

The default routing rule stores a non-uncertain single-call recommendation when verbalized confidence is at least 0.80. Otherwise the system samples additional responses at the configured sampling temperature, with a default sample count of five, and applies a 60% agreement rule. A modal uncertain recommendation remains uncertain; failure to reach the agreement threshold is also represented as uncertainty. Self-consistency motivates repeated sampling [17], but neither agreement among calls nor verbalized confidence establishes calibrated probability or accuracy [18]. Both are routing signals requiring separate validation.

All resulting outputs remain AI recommendations (Fig. 5). They are excluded from the assigned human reviewers' reconciliation, completion checks and inter-reviewer agreement statistic. The platform does not use an accepted AI recommendation as an autonomous eligibility decision. AI screening is limited to titles and abstracts in this release. An input hash supports reuse of stored recommendations for unchanged inputs; such reuse provides traceability, not guaranteed reproducibility of a new provider response.

Extraction, appraisal and synthesis

Extraction uses a review-specific typed template with required fields, permitted values and optional bounds. AI can propose template fields or draft values, but human reviewer records remain distinct from AI drafts. Per-reviewer extractions are compared and a consensus extraction supplies synthesis. Template changes are audited; changing a type for future submissions does not retrospectively harmonize previously collected values, which may require human reconciliation.

The release contains structures for RoB 2, ROBINS-I, QUADAS-2, the Newcastle–Ottawa Scale and AMSTAR 2, and a review-level ROBIS record [19–24]. Instrument scope and version must be recorded for each application. Simplified software summaries must not replace the instruments' signalling questions, decision rules and assessor judgement. Table 3 distinguishes the intended appraisal from reporting limitations; no validation of R-SYS implementations of these instruments is reported.

Table 3
Appraisal structures and interpretation safeguards

Instrument	Scope	Safeguard for use and reporting
RoB 2	Randomized trial results	Apply the declared version and full domain guidance. Multiple domains with some concerns can affect the overall judgement.
ROBINS-I	Non-randomized intervention results	Preserve no-information judgements and the instrument’s overall rules; do not treat missing information as an ordinary severity rank.
QUADAS-2	Diagnostic accuracy studies	Keep risk-of-bias domains and applicability concerns distinct. An R-SYS “at risk” summary is not a validated replacement for the domain profile.
Newcastle–Ottawa Scale	Eligible observational designs	Report domain and total stars. Any categorical cut-off requires explicit justification.
AMSTAR 2	Systematic reviews of interventions	Distinguish critical flaws from non-critical weaknesses; do not reduce the assessment to a total score.
ROBIS	Risk of bias in a systematic review	Use the complete staged assessment and overall judgement; four domain entries alone do not demonstrate a complete ROBIS assessment.

Pairwise synthesis operates on effect estimates and variances using pre-specified statistical options. The described functions support inverse-variance fixed-effect pooling, random-effects pooling with restricted maximum likelihood or DerSimonian–Laird estimation, Wald or Hartung–Knapp–Sidik–Jonkman intervals, heterogeneity measures and prediction intervals [25–29]. Egger and Begg tests are available only when at least ten studies are pooled; study count alone does not establish that either test is appropriate for a particular effect measure or evidence set [30–32]. GRADE judgements are recorded per outcome with reasons [33]. These are implementation capabilities, not reported meta-analytic results; numerical agreement with an established statistical implementation remains to be evaluated.

Final convergence and reporting

Convergence compares the assigned primaries' current, non-superseded records. It identifies title/abstract conflicts, full-text conflicts, extraction mismatches, risk-of-bias mismatches and incomplete catch-up tasks. Each category must be resolved before synthesis (Fig. 6). Human consensus resolves extraction and appraisal differences. Reconciliation and adjudication use a defined write path for shared record status, and final convergence reruns reconciliation rather than trusting stale cached status. This architecture supports consistency but does not by itself prove correct implementation.

Audit events record protocol changes, human decisions, adjudications, AI interactions and workflow movements. AI provenance includes model and prompt identifiers, input hash, protocol version, applied parameters, raw and parsed responses, latency and token counts. Re-issuing a request may still yield a different response. Human decisions are attributable and version-linked; they should not be described as intrinsically deterministic.

The full lifecycle is summarized in Fig. 7. The reporting module assembles flow counts from reconciled records and provides decision-level exports. PRISMA 2020 distinguishes records, reports and studies [13]. The reference description does not establish a validated report-to-study linkage procedure; included-record totals must therefore not be assumed to equal included-study totals. Original discovery source, reports not retrieved, full-text exclusion reasons and multiple reports per study require reconciliation before publication. Figure 8 is an illustrative PRISMA 2020 output template, not a flow diagram of an evaluated R-SYS cohort. Its empty count fields are intentional.

An internal agreement report describes the two primary reviewers' decisions using Cohen's κ and descriptive counts [34]. Proportion estimates may use Wilson confidence intervals [35]; Wilson intervals do not apply to κ. AI–human comparisons are agreement with the review's reference decisions, not external validation, and their interpretation depends on how uncertainty, missing decisions and AI exposure are handled.

Analytical results

Identical task sets

Consider R reviewers and stages s consisting of title/abstract screening, full-text assessment, extraction and appraisal. Let t(r,s) be reviewer r's active work time at stage s and W(r) = Σ_s t(r,s). Assume identical assigned tasks across the compared disciplines, fixed task durations, continuous reviewer availability, no external retrieval delays, no amendments and no adjudication or other shared-resource waiting. Reviewers work whenever an eligible task is available.

$$T_{lanes} = max_r W(r) \leq T_{record}$$

$$T_{lanes} \leq T_{stage} = \Sigma_s max_r t(r,s)$$

Under independent lanes, each reviewer's prerequisites are their own preceding work. Every reviewer therefore finishes after W(r), and the last finishes at max_r W(r). This largest workload is a lower bound for any schedule processing the same work with the same reviewer resources. Consequently, neither record barriers nor stage barriers can improve on that bound under these assumptions. Stage barriers give T_{stage} = Σ_s max_r t(r,s). The maximum of sums cannot exceed the sum of maxima. Equality holds when at least one reviewer attains a maximum workload at every stage, allowing ties.

No universal ordering between particular record-barrier and stage-barrier implementations is asserted without defining their task-ordering and scheduling policies. The independent-lane lower bound also does not imply that a real deployment will achieve it when researchers have unequal availability, retrieval delays or competing responsibilities.

Disagreement and additional delay

When reviewers disagree, let W_c(r) denote the active workload required from reviewer r for the final consensus task set, and w(r) the additional active work later superseded. Let d(r) denote that reviewer's cumulative non-working delay caused by task-release dependencies or adjudication within the comparison horizon. Catch-up work that belongs to the final consensus task set is included in W_c(r), not counted again as additional effort. Its delayed availability can instead contribute to d(r). Under the same fixed-duration and continuous-availability assumptions, and with all required work completed:

$$T_{lanes} = max_r \{W_c(r) + w(r) + d(r)\}$$

$$\leq max_r W_c(r) + max_r w(r) + max_r d(r)$$

Define Δ = $T_{stage,c}$ − max_r W_c(r), where $T_{stage,c}$ is the stage-barrier schedule for the consensus work with no additional waiting. A sufficient, conservative condition for independent lanes to be no slower than this idealized comparator is:

$$max_r w(r) + max_r d(r) \leq \Delta$$

This follows by bounding the maximum of the workload, superseded-work and delay sums by the sum of their maxima. The condition allows delays at multiple stages; it does not assume that all catch-up tasks arrive in a single final batch. It is not a measured benefit, and the bound may be loose. Whether synchronization savings exceed superseded work and dependency delays requires empirical measurement. Both active reviewer effort and elapsed completion time should be reported, because less waiting can coexist with more work.

What the implementation establishes

The reference design specifies where protocol versions, human decisions, AI advice and shared review state are represented and how they interact. These mechanisms make particular checks possible; their

presence is not evidence that searches are complete, decisions are accurate, or completion is faster. Table 4 separates each design mechanism from the evidence still needed. No accuracy, workload, elapsed-time, cost, usability or deployment estimates are presented as results in this article.

Table 4
Design mechanisms and present evidence boundaries

Mechanism	What it makes possible	Evidence not supplied in this paper
Versioned protocols	Attribution of decisions to criteria and settings.	Adequacy of amendment scope or prevention of selection bias.
Canonical query rendering	Repeatable production of a target expression.	Cross-database semantic equivalence or retrieval sensitivity.
Independent reviewer lanes	Progress without specified peer-stage barriers.	Measured time savings, total effort or user benefit.
AI separated from human reconciliation	A distinct automated recommendation record.	AI accuracy, calibration or freedom from indirect influence.
Convergence and controlled state writes	Comparison of outstanding human work before analysis.	An independently verified absence of implementation defects.
Provenance and reporting exports	Inspection of decision history and reported counts.	Validated source classification, study linkage or PRISMA compliance.

Discussion

The central methodological proposal is to separate independent judgement from synchronization. A reviewer may proceed using their own decisions while retaining a requirement to complete the consensus workload before synthesis. Versioned protocols identify which criteria governed a decision; rolling adjudication changes working sets; convergence reconciles differences before analysis. The scheduling result explains when removal of waiting can help, while explicitly allowing additional effort and delay when decisions differ.

Existing systems address related problems through collaborative screening, active learning, search translation, registries and reference management [3–5, 10]. The contribution of R-SYS is the specific integration and representation described here. This article does not establish that these features are absent from competitors or that a single integrated platform is preferable to a well-coordinated collection of tools. No feature superiority matrix is presented because public documentation alone cannot reliably establish absence of a capability, and no licensed head-to-head assessment was conducted.

The advisory AI design preserves a distinction between automated recommendations and human decisions of record. It does not automatically provide an AI-independent reference standard. Reviewers can encounter AI outputs after submitting an initial decision, and adjudicators may have broader access.

A validation study must therefore prevent relevant AI exposure until the reference standard is locked or use an independent reference panel. Audit logs can document exposure but cannot undo its influence. Publication of a review using R-SYS demonstrates an application; it does not, by itself, validate R-SYS or the AI component.

Several design choices need careful interpretation. Protocol locking does not establish that the original protocol was adequate, and a selective amendment policy can preserve inappropriate exclusions. Deterministic search rendering does not equate vocabularies or establish retrieval sensitivity. Conservative progression during initial screening is not equivalent to justified final inclusion. Likewise, a single write path supports consistency of recorded statuses but cannot correct a mistaken source label or an omitted link between reports of one study. These are reasons to retain human methodological review alongside technical controls.

Limitations

This is a descriptive methodology and reference-implementation paper with analytical scheduling results. It contains no empirical validation, usability study, deployment series, independent code audit or comparative trial. Implementation statements refer to the authors' described September 2026 release. The authors have a financial interest in the platform, and only the specified core modules are open source; the complete application is proprietary.

Two search clients are executable. Other search expressions require manual execution and potentially substantial adaptation, particularly for controlled vocabularies and platform-specific syntax. Deduplication thresholds are unvalidated and the restricted fallback strategy can miss duplicates. Report-to-study linkage and source reconciliation must be established before publication counts can be trusted. AI confidence is uncalibrated, the sampling rule is unvalidated, and the model screens titles and abstracts only. Broader visibility privileges and later AI exposure limit claims of blinding. The inclusive conflict bypass and selective amendment policies require the methodological restrictions described above. Appraisal and synthesis implementations have not been independently benchmarked.

The timing analysis omits ordinary interruptions, variable reviewer availability, retrieval delays, shared-resource constraints and behavioural responses to the interface. A replay based on observed task durations would also depend on the assumption that those durations do not change with the scheduling discipline. Audit timestamps alone do not distinguish active work from time away from the platform. These limitations prevent extrapolation from the analytical lower bound to real-world efficiency.

Planned follow-up validation

A separate study is planned after completed reviews using R-SYS become available. Its evaluation protocol will be specified prospectively before validation analyses, with retrospective and prospective components identified explicitly. Priorities are AI screening against an independently locked human reference, duplicate detection, source/report/study accounting, numerical verification of synthesis

functions, reviewer effort, elapsed time, usability and cost. Multiple clinical topics and external reviews are needed to assess transferability. Additional file 1 outlines these evaluation domains without presenting them as completed work or as a registered protocol.

Conclusions

R-SYS describes a method for linking review decisions to versioned protocols while allowing primary reviewers to advance independently before final convergence. Under explicit assumptions, independent lanes attain the lower bound imposed by reviewer workload; disagreement and delayed adjudication can reduce or reverse that advantage. Advisory AI, decision provenance and controlled reconciliation provide an auditable workflow architecture. Their practical reliability and benefit remain to be established in the planned follow-up validation study.

Abbreviations

AI
artificial intelligence
AMSTAR 2
A MeaSurement Tool to Assess systematic Reviews 2
DOI
digital object identifier
GRADE
Grading of Recommendations Assessment,Development and Evaluation
LLM
large language model
MeSH
Medical Subject Headings
NLM
National Library of Medicine
OSF
Open Science Framework
PICO
population,intervention,comparator,outcome
PRESS
Peer Review of Electronic Search Strategies
PRISMA
Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2
Quality Assessment of Diagnostic Accuracy Studies 2
RoB 2
Risk of Bias 2

ROBINS-I
Risk Of Bias In Non-randomised Studies of Interventions
ROBIS
Risk of Bias in Systematic Reviews.

Declarations

Ethics approval and consent to participate:

Not applicable. This paper describes a method and software implementation and reports no human-participant research or patient-level data.

Consent for publication:

Not applicable.

Use of generative AI and writing assistance

Grammarly assisted with language editing, manuscript organization and figure preparation. The authors retain responsibility for the accuracy, interpretation and integrity of the manuscript and figures. These tools are not authors.

Competing Interests

YT and MI are affiliated with R-Chief Labs Ltd, which develops and commercialises R-SYS, and hold a financial interest in it. The core modules described in this paper are released under an open-source licence; the surrounding platform is commercial.

Additional file

Additional file 1

Technical specification and planned validation domains. Describes transition requirements, amendment policies, reconciliation, AI provenance, and the proposed scope of a separate evaluation. It contains no empirical validation results.

Funding:

R-Chief Labs Ltd funded development of R-SYS. No external grant funding was received for this work.

Author Contribution

YT conceived the platform and the methodology, led the system architecture, designed the evaluation protocol, and revised the manuscript. MI implemented the pipeline and both authors drafted the manuscript. Both authors read and approved the final manuscript.

Acknowledgement

We thank the clinical research collaborators who took part in early platform testing and whose feedback shaped the methodology described here.

Availability of data and materials:

No empirical datasets were generated or analysed for this paper. Core source-code availability is specified above. Additional file 1 provides implementation details and the scope of the planned evaluation. Figure source files accompany the manuscript package.

Availability and requirements

Project name: R-SYS core, version 1.0.0.

Project home page: https://github.com/rchief-labs/rchief-rsys-core

Archived version: https://doi.org/10.5281/zenodo.22904806

Operating systems: Platform independent.

Programming language: Java 25 for the core library; the full platform additionally uses Spring Boot 4 and TypeScript with Angular 21.

Requirements: The full platform requires PostgreSQL 17, an OpenID Connect provider such as Keycloak, an AI-provider API key for AI features, and access to the relevant bibliographic services. Core-library use does not require the full application.

Licence: MIT for the published core modules covering review phases and transition interfaces, query rendering, reviewer working-set derivation and convergence comparison. The surrounding platform is proprietary. The open core does not reproduce every integration or independently establish every platform claim. No additional non-academic-use restriction is stated for the MIT-licensed modules.

References

1. 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(2):e012545.

2. Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc. 2006;13(2):206–219.
3. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5:210.
4. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3:125–133.
5. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8:163.
6. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;15(4):616–626.
7. Tran VT, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, et al. Sensitivity and specificity of using GPT-3.5 turbo models for title and abstract screening in systematic reviews and meta-analyses. Ann Intern Med. 2024;177(6):791–799.
8. Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y. Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using GPT-3.5 and GPT-4 for systematic reviews. J Med Internet Res. 2024;26:e52758.
9. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev. 2024;13:219.
10. Clark JM, Sanders S, Carter M, Honeyman D, Cleo G, Auld Y, et al. Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. J Med Libr Assoc. 2020;108(2):195–207. doi:10.5195/jmla.2020.834.
11. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5.1. Cochrane; 2025. Available from: https://www.cochrane.org/handbook.
12. Pieper D, Puljak L. Language restrictions in systematic reviews should not be imposed in the search strategy but in the eligibility criteria if necessary. J Clin Epidemiol. 2021;132:146–147.
13. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
14. van den Akker OR, Peters GJY, Bakker CJ, Carlsson R, Coles NA, Corker KS, et al. Increasing the transparency of systematic reviews: presenting a generalized registration form. Syst Rev. 2023;12:170.
15. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. J Clin Epidemiol. 2016;75:40–46.
16. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. 2016;104(3):240–243.

17. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representations (ICLR); 2023.
18. Xiong M, Hu Z, Lu X, Li Y, Fu J, He J, Hooi B. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. International Conference on Learning Representations (ICLR); 2024.
19. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.
20. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.
21. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–536.
22. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle–Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute.
23. 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.
24. Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–234.
25. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Educ Behav Stat. 2005;30(3):261–293.
26. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–188.
27. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Stat Med. 2001;20(12):1771–1782.
28. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:25.
29. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. 2016;6(7):e010247.
30. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–634.
31. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50(4):1088–1101.
32. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials.

BMJ. 2011;343:d4002.
33. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–926.
34. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
35. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209–212.

Figures

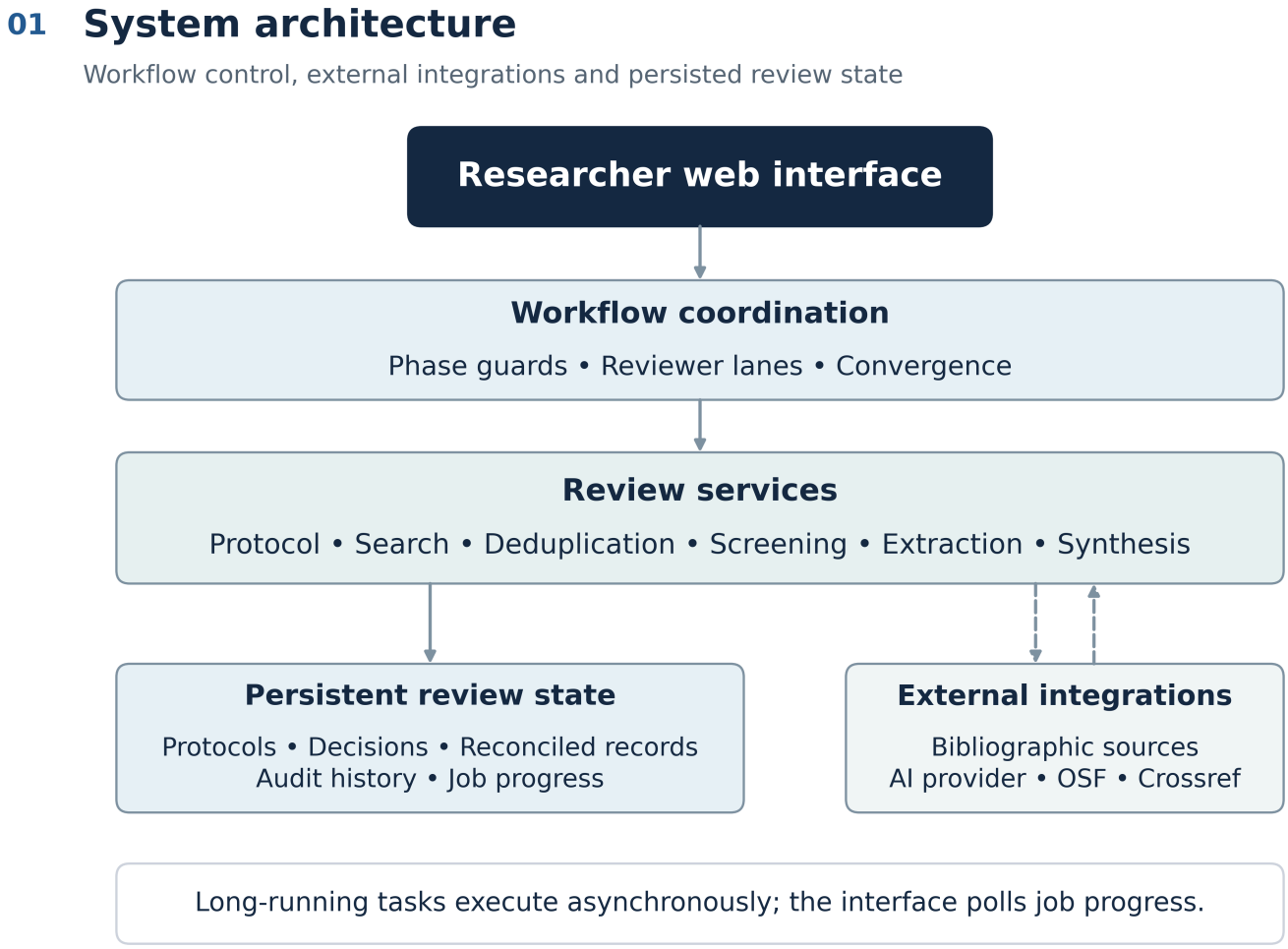


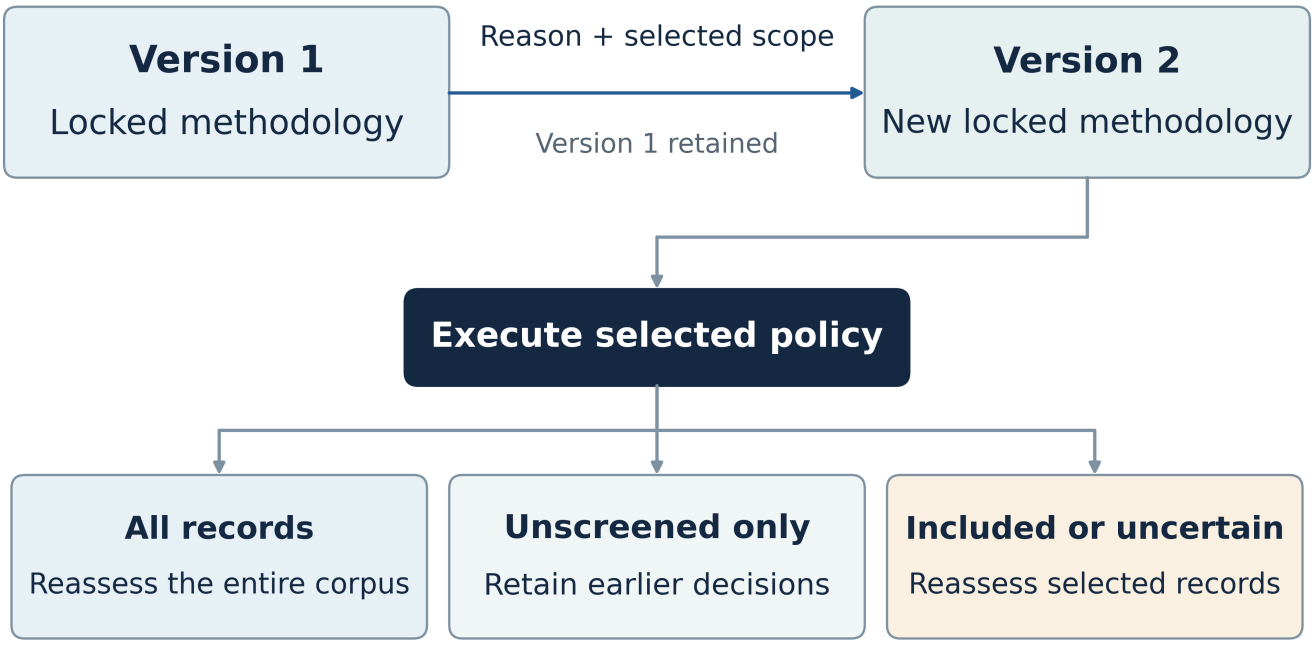
Figure 1

System architecture. The interface invokes workflow coordination and review services. Review state, decisions, audit events and job progress are persisted. Dashed arrows indicate external service interactions; solid arrows indicate internal control or persistence relationships. The diagram is conceptual and does not assert a one-to-one mapping between boxes and deployed services. Long-running tasks execute asynchronously.

02

Protocol versions and amendment scope

Decisions retain the protocol version under which they were made



Affected decisions become stale; affected records return to screening.

Figure 2

Protocol versions and amendment scope. An amendment creates a new locked version while retaining its predecessor. The three branches show selectable scopes in the described release. Affected decisions become stale and affected records return to screening. Policy availability does not establish methodological adequacy: broadened criteria require reassessment of potentially affected exclusions, and changed search concepts may require new retrieval. Details and limitations are given in the text.

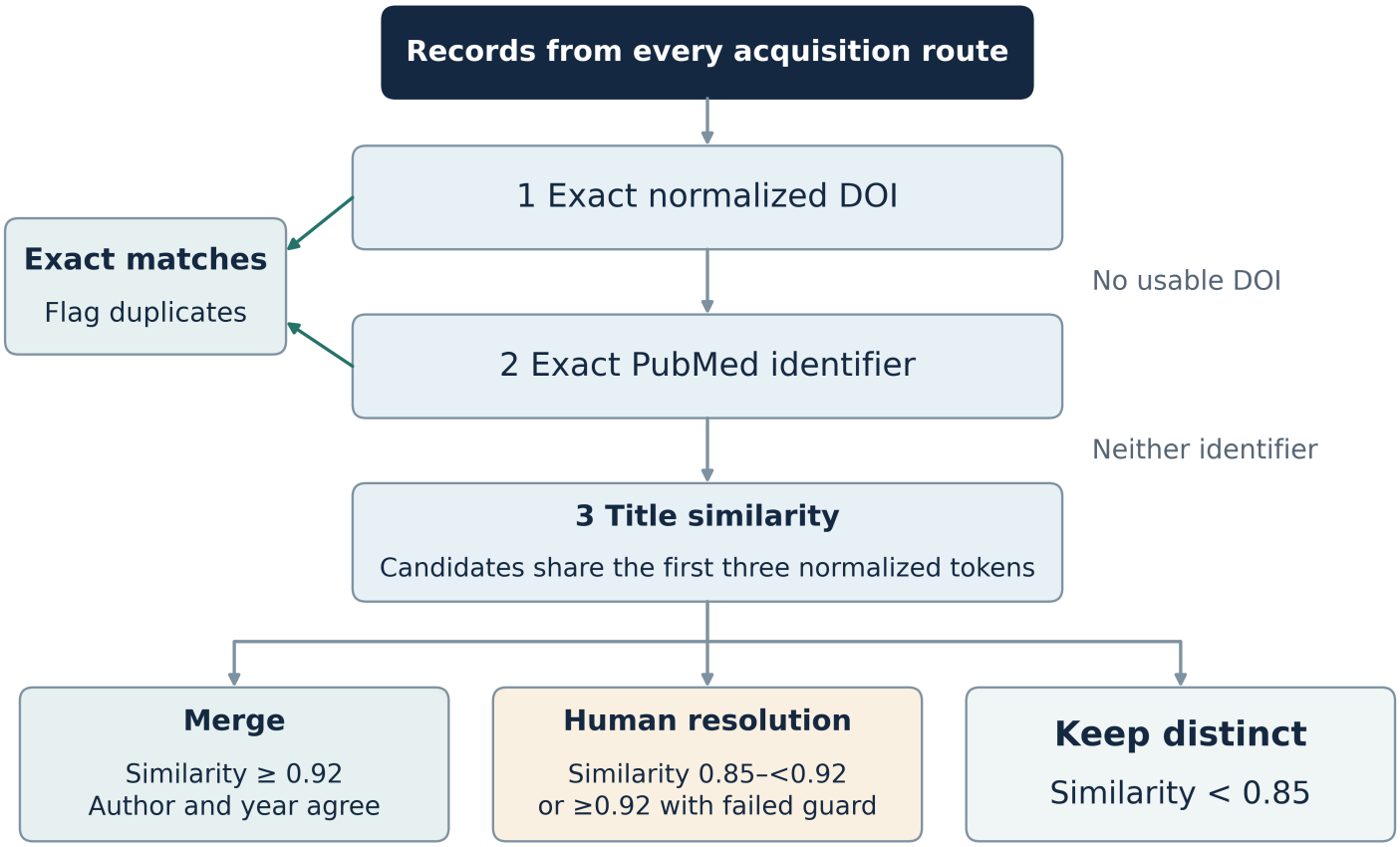


Figure 3

Deduplication and manual resolution. The figure describes the restricted fallback logic in the described implementation: exact DOI matching, then PubMed-identifier matching among records without a usable DOI, then title matching among records with neither identifier. Candidate title pairs are blocked on their first three normalized tokens. Exact identifier matches are resolved at their respective pass. The three final branches describe title-pair handling. Mixed-identifier duplicates and pairs excluded by blocking can be missed; thresholds are unvalidated. Bibliographic deduplication does not link distinct reports of one study.

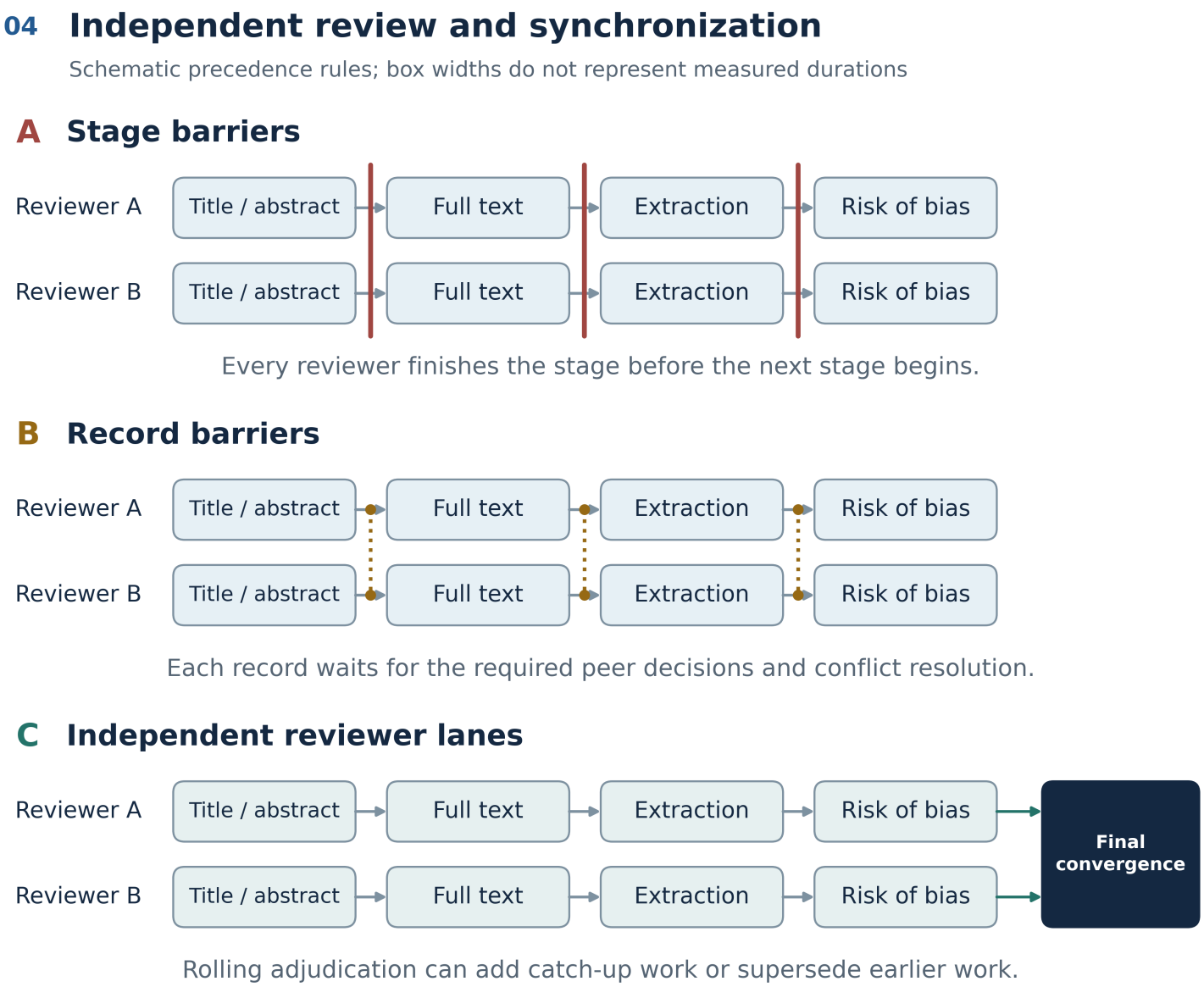


Figure 4

Independent review and synchronization. (A) Stage barriers prevent progression until all primaries finish the preceding stage. (B) Record barriers depend on required decisions for the individual record. (C) Independent lanes permit progression based on the reviewer's own working set, with rolling human adjudication and a final convergence checkpoint. The layout shows dependencies, not measured durations. The largest reviewer workload is attained only under the identical-task, fixed-duration, continuous-availability and no-external-waiting assumptions in the analytical results. No unconditional ranking of record and stage implementations is asserted.

05 AI screening produces advice only

Human eligibility decisions are required regardless of the AI recommendation

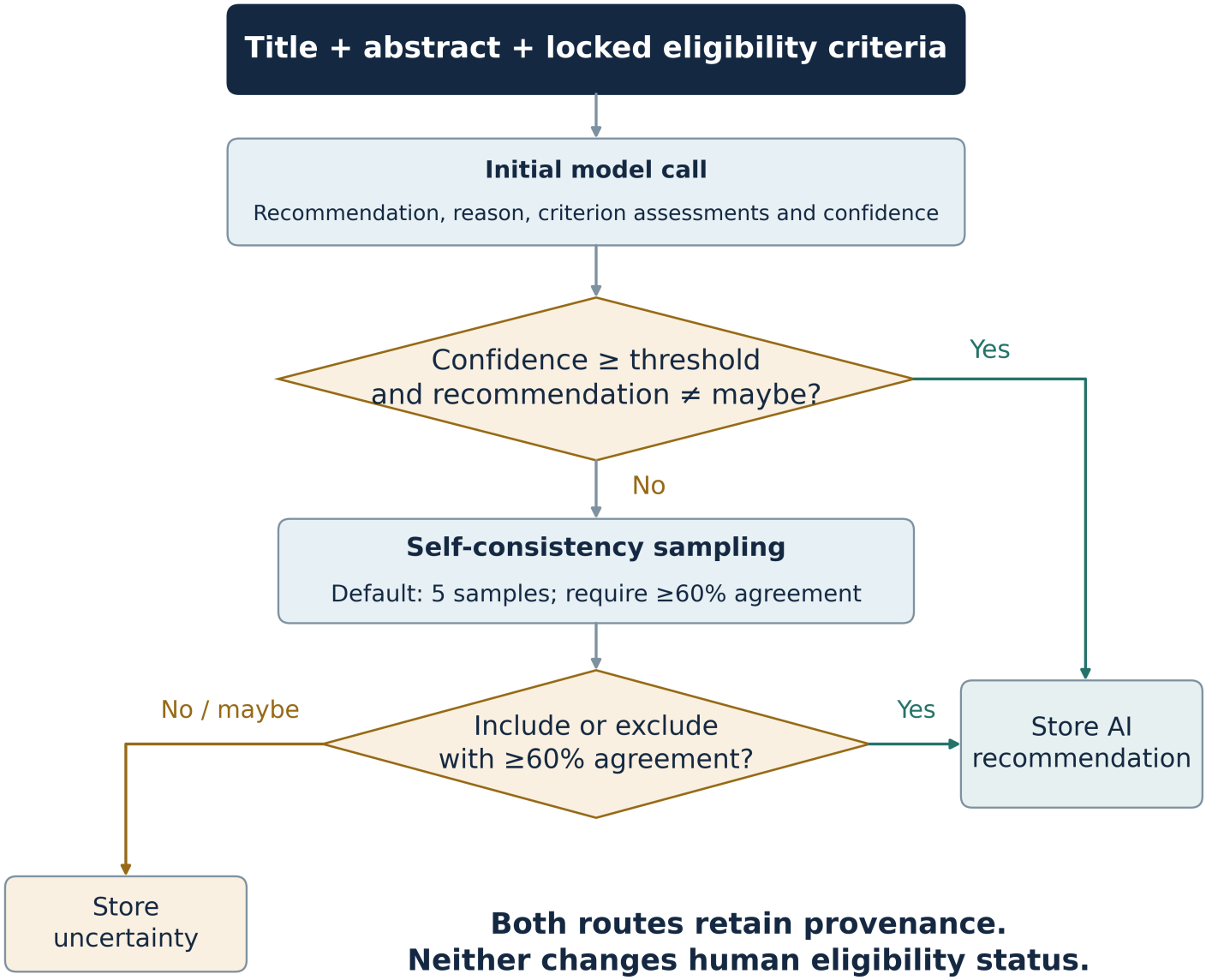


Figure 5

Advisory AI screening. An initial model response provides criterion assessments, a recommendation, a reason and verbalized confidence. The configured threshold routes a non-uncertain response to storage or triggers resampling. The default confidence threshold is 0.80; the default sample count is five and the agreement threshold is 60%. A modal maybe remains uncertain. All outputs are advisory, including outputs that satisfy the routing thresholds. Neither route changes human eligibility status. Confidence and sample agreement are not validated probabilities. Exact model, prompt, sampling and failure-handling details must accompany empirical use.

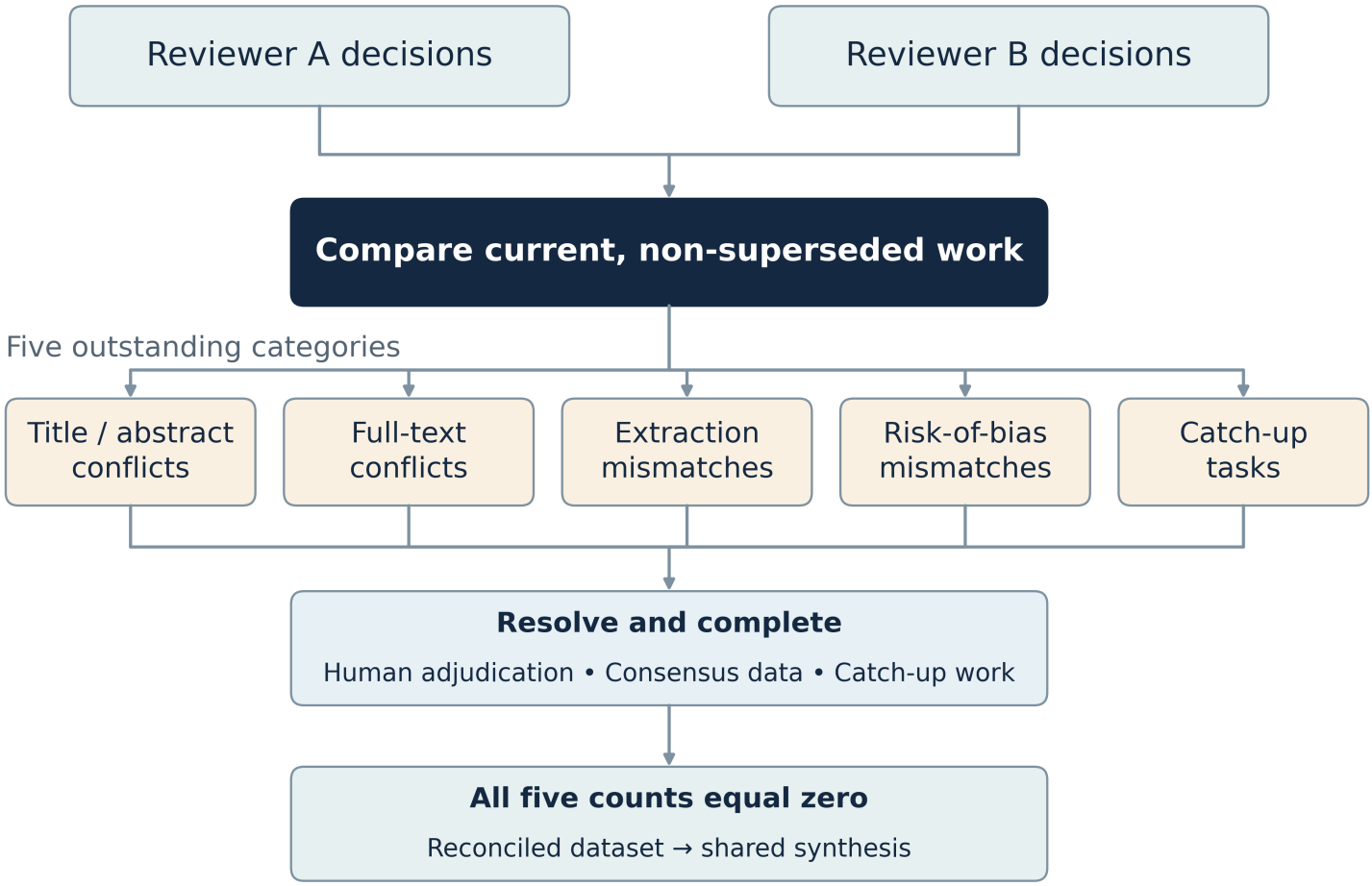
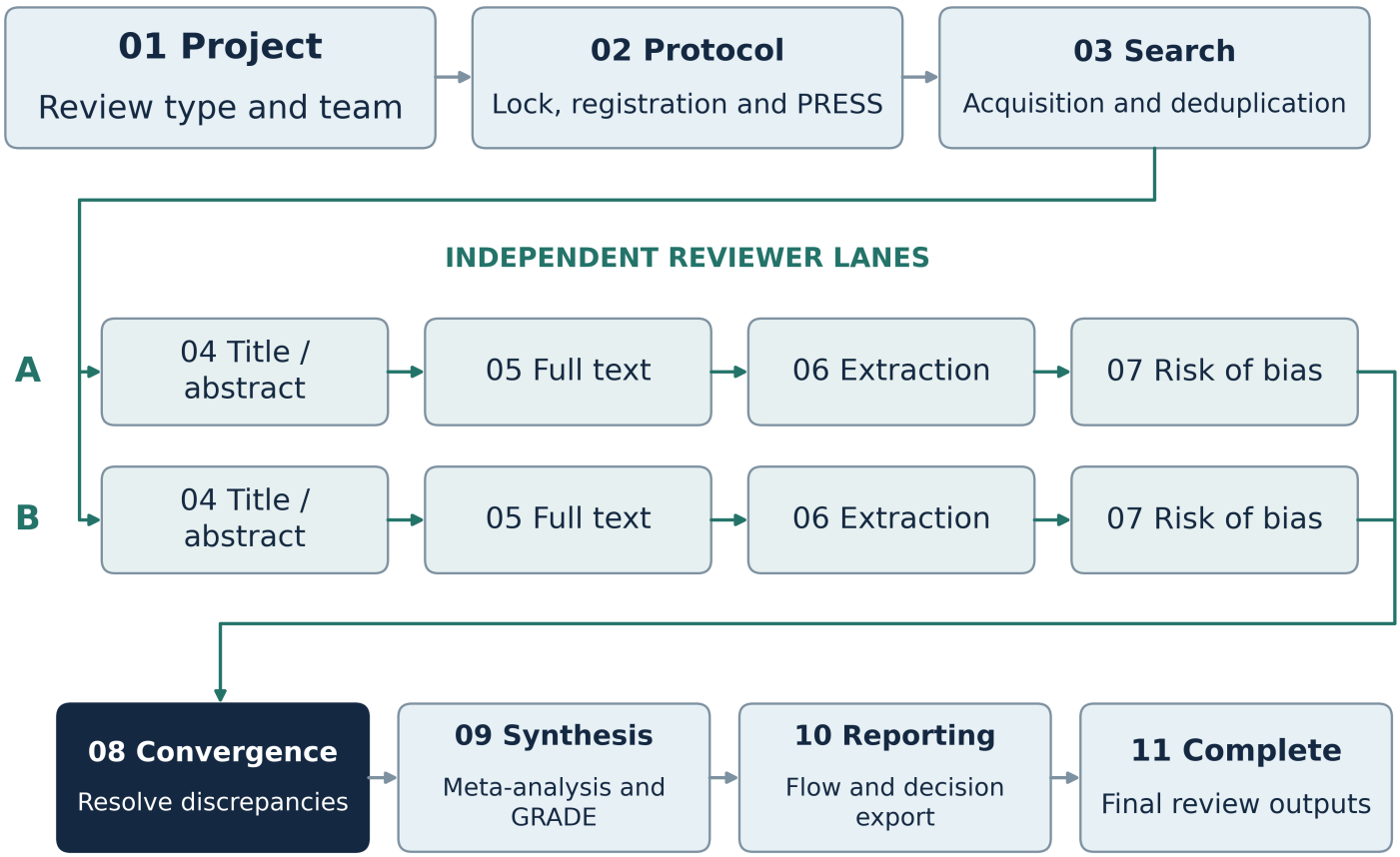


Figure 6

Final convergence. The assigned primaries' current, non-superseded work is compared. Human resolution and completion address five outstanding categories: title/abstract conflicts, full-text conflicts, extraction mismatches, risk-of-bias mismatches and catch-up tasks. Synthesis requires all five counts to reach zero. AI advisory rows are excluded. The diagram summarizes the intended sequence; earlier rolling reconciliation and adjudication may already have resolved some records.

07 The review lifecycle

Shared preparation → independent human review → shared analysis and reporting



Transitions are guarded; amendments can return affected records to screening.

Figure 7

Eleven-phase review lifecycle. Shared preparation comprises project creation, protocol and search. Each primary reviewer independently traverses title/abstract screening, full-text assessment, extraction and appraisal. Final convergence precedes shared synthesis, reporting and completion. The global phase summarizes the least-advanced lane while individual working sets remain reviewer-specific. Registration or waiver and search-strategy peer review are requirements at the protocol-to-search boundary. Amendments and relevant resets can reopen earlier work.

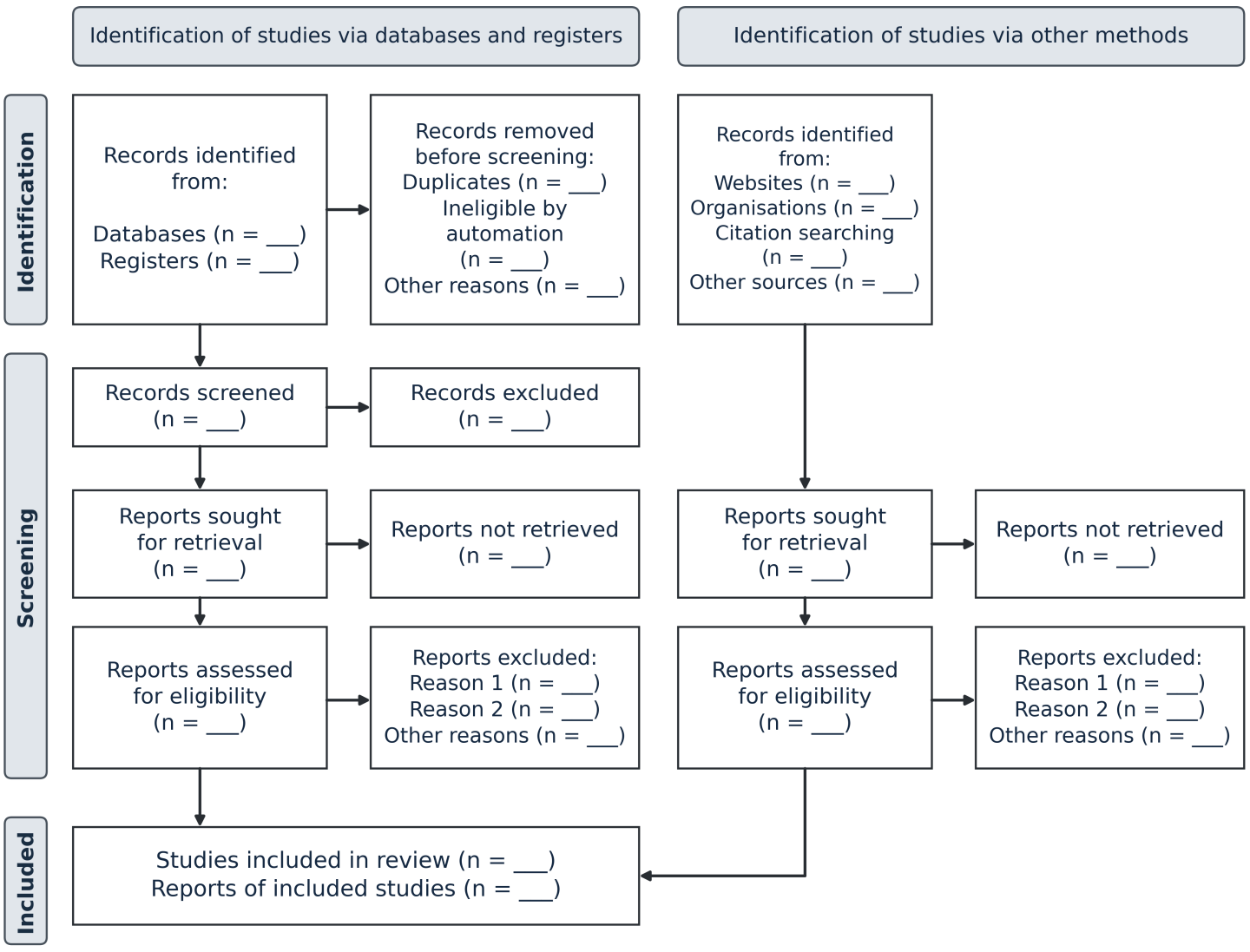


Figure 8

Illustrative PRISMA 2020 reporting template. Adapted from Page et al. [13] and the PRISMA 2020 template for new reviews with databases, registers and other sources (https://www.prisma-statement.org/prisma-2020-flow-diagram), licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Redrawn with revised typography and spacing. Empty count fields are intentional: this methodology paper reports no review cohort. The database/register arm is determined by the discovery source, not by file import or browser capture. Reports and studies are separate units. The figure specifies a reporting structure and does not establish that every required provenance or study-linkage field is automatically produced by the current release.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RSYSTechnicalSupplement.docx